

Statistical Methods and Models for Modern Genetic Analysis

by

Matthew S. Zawistowski

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Assistant Professor Sebastian K. Zöllner, Chair
Professor Michael L. Boehnke
Professor David T. Burke
Associate Professor Thomas M. Braun
Associate Professor Noah A. Rosenberg

© Matthew S. Zawistowski 2011

All Rights Reserved

For my parents,
without whom, I would not have started.

And for my wife, Jamie,
without whom, I would not have finished.

TABLE OF CONTENTS

DEDICATION	ii
LIST OF FIGURES	v
LIST OF TABLES	vii
ABSTRACT	viii
CHAPTER	
I. Introduction	1
II. The Winner’s Curse in Studies of Gene-Environment Interaction	8
2.1 Introduction	8
2.2 Model	11
2.3 Statistical methods	13
2.3.1 Prevalence constraint	13
2.3.2 Naive estimate of θ	14
2.3.3 Corrected estimate of θ	14
2.3.4 Partial likelihood Markov Chain Monte Carlo	14
2.3.5 Proposal density $q(\theta \rightarrow \theta^*)$	18
2.4 Simulation methods	20
2.5 Results	22
2.5.1 Sampling distributions for naive estimates $\hat{\theta}$	22
2.5.2 Corrected estimates $\bar{\theta}$	31
2.6 Discussion	37
2.7 Appendix	40
2.7.1 Stochastic grid search algorithm	40
2.7.2 Initial genome-wide association testing strategies	40
III. Extending Rare Variant Testing Strategies: Analysis of Non-Coding Sequence and Imputed Genotypes	42

3.1	Introduction	42
3.2	Methods	46
3.2.1	Data structure	46
3.2.2	Cumulative minor allele test	47
3.2.3	Alternative rare variant methods	49
3.2.4	Simulations	50
3.2.5	Stratified datasets	53
3.2.6	Simulation settings	54
3.2.7	GWAS application	54
3.3	Results	54
3.3.1	Deep coverage sequencing datasets	54
3.3.2	Covariate correction	58
3.3.3	Imputation datasets	58
3.3.4	Application to GAIN psoriasis data	62
3.4	Discussion	65
3.5	Appendix	69
3.5.1	Empirical distributions for expected minor allele counts	69
3.6	Supplementary material	71
IV. A Coalescent Model for Genotype Imputation		75
4.1	Introduction	75
4.2	Coalescent model for genotype imputation	77
4.3	Methods	81
4.3.1	Derivation of reference panel optimality probabilities	82
4.3.2	Derivation of expected coalescent times	88
4.3.3	Derivations of probabilities and expectations under exponential growth	92
4.4	Results	93
4.4.1	Consistency of computations	93
4.4.2	Constant-sized populations	94
4.4.3	Exponentially growing populations	97
4.5	Discussion	99
4.6	Appendix	103
4.6.1	The quantity $N(i_D, j_D \rightarrow i_C, j_C, k_C)$	103
4.6.2	Table of equations	107
V. Discussion		108
BIBLIOGRAPHY		112

LIST OF FIGURES

Figure

2.1	Naive MLEs conditional on genome-wide significance in the allelic chi-square test ($d > 0, e' > 0$)	24
2.2	Naive MLEs conditional on genome-wide significance in the logistic regression ($d > 0, e' > 0$)	25
2.3	Naive MLEs conditional on genome-wide significance in the allelic chi-square test ($d = 0, e' > 0$)	26
2.4	Naive MLEs conditional on genome-wide significance in the logistic regression ($d = 0, e' > 0$)	27
2.5	Naive MLEs conditional on genome-wide significance in the allelic chi-square test ($d > 0, e' = 0$)	28
2.6	Naive MLEs conditional on genome-wide significance in the logistic regression ($d > 0, e' = 0$)	29
2.7	QQ plots for type I error in follow-up tests of gene-environment interaction	30
2.8	Convergence of $\bar{\theta}$ in the partial likelihood MCMC	34
2.9	Comparison of corrected and naive estimates of θ	35
2.10	Power for a replication analysis based on naive and corrected estimates	36
3.1	Relation between minor allele frequency and relative risk in our disease model	52
3.2	Power to analyze deep sequencing datasets for a range of inclusion probabilities.	56

3.3	Application of the covCMAT to control for population stratification	59
3.4	Comparison of CMAT power for deep sequencing and imputation study designs	60
3.5	CMAT power for imputation datasets	61
3.6	Type I error rates for the rare variant tests	71
3.7	Power for CMAT and WSS using both uniform and maf-based weights	72
3.8	Power to analyze deep sequence datasets for minor allele cutoff $\beta = 1\%$	73
3.9	CMAT power for imputation datasets with minor allele cutoff $\beta = 1\%$	74
4.1	Two population coalescent model for imputation reference panel selection	79
4.2	The two-population coalescent model of divergence, assuming exponential growth in the descendant populations	81
4.3	Coalescence times between the target T and the reference panels . .	89
4.4	Imputation performance for the constant-size two population model.	95
4.5	Imputation performance for the exponential growth two-population model.	98
4.6	An illustration of $N(i_D, j_D \rightarrow i_C, j_C, k_C)$	106

LIST OF TABLES

Table

2.1	Gene-environment penetrance model \mathcal{M}	12
2.2	Additive model of gene-environment interaction \mathcal{M}_+	20
3.1	Summary of the <i>SKIV2L</i> testing unit from CMAT analysis of the GAIN psoriasis dataset	64
3.2	Summary of empirical distributions of minor allele dosage for true heterozygotes	70
4.1	Derivation of the recursion $\tilde{P}(C_1 i, k, j)$	86
4.2	$P(C_1)$ computed analytically using closed form and recursive equations and estimated using coalescent simulations.	94
4.3	Summary of all derived equations and their dependencies	107

ABSTRACT

Statistical Methods and Models for Modern Genetic Analysis

by

Matthew S. Zawistowski

Chair: Sebastian Zöllner

The Genome-Wide Association Study (GWAS) is the predominant tool to search for genetic risk variants that contribute to complex human disease. Despite the large number of GWAS findings, variants implicated by GWAS are themselves unlikely to fully explain the heritability of many diseases. In this dissertation, we propose statistical methods to augment GWAS and further our understanding of the genetic causes of complex disease.

In the first project, we consider the challenges of a gene-environment analysis performed as a follow-up to a significant initial GWAS result. It is known that effect estimates based on the same data that showed the significant GWAS result suffer from an upward bias called the “Winner’s Curse.” We show that the initial GWAS testing strategy can induce bias in both follow-up hypothesis testing and estimation for gene-environment interaction. We propose a novel bias-correction method based on a partial likelihood Markov Chain Monte Carlo algorithm.

In the second project, we shift attention to rare genetic variants that have low power of being detected by GWAS. We propose the Cumulative Minor Allele Test (CMAT) to pool together multiple rare variants from the same gene and test for an excessive burden of rare variants in either cases or controls. We show the CMAT performs favorably across a range of study designs. Notably, the CMAT accommodates probabilistic genotypes, extending applicability to low-coverage and imputed sequence data. We use a simulation analysis to validate study designs that combine sequenced and imputed samples as a means to improve power to detect rare risk variants.

Determining conditions that optimize imputation accuracy is important for successful application. In the final project, we propose a coalescent model of genotype imputation that allows fast, analytical estimates of imputation accuracy across complex population genetic models. We use our model to compare the performance of custom-made reference panels drawn from the same source population as imputation targets to publicly available reference panels (i.e. 1000 Genomes Project) that may differ in ancestry from the targets.

CHAPTER I

Introduction

The identification of genetic variants that contribute to common complex disease is an area of intense interest in the current field of human genetics. As such, substantial research is devoted to the development of statistical methods to powerfully analyze genetic data to detect risk variants. The population-based association study involves testing for a statistically significant correlation between genotype and phenotype in a set of independent samples and has emerged as the preeminent tool in complex disease genetics [65]. The simplicity of the test allows any genetic variant that can be typed in a large sample to be tested for association; no prior knowledge of the potential genotype-phenotype relationship is required. Due to the millions of single nucleotide polymorphisms (SNPs) cataloged across the human genome by the The HapMap [2] and 1000 Genomes Projects [18] and decreasing genotyping costs, the entire genome is now routinely scanned for risk variants in Genome-Wide Association Studies (GWAS) [17]. To date, GWAS have identified more than four thousand genetic variants that predispose for complex diseases ranging from diabetes to bipolar disorder to breast cancer [25, 31, 69, 73].

Although risk variants identified by GWAS account for an appreciable fraction of heritability for some phenotypes [80], it is more often the case that GWAS identified risk variants explain only a small proportion of disease risk [42, 46]. Thus, despite the overwhelming number of GWAS findings, our understanding of many complex diseases remains incomplete. It is therefore necessary to consider conditions under which GWAS are either unlikely to detect genetic risk variants or incorrectly characterize their effect. GWAS experiments are often performed and interpreted under a simple model of marginal genetic effects for individual risk variants. In reality, risk for many complex diseases is likely the cumulative effect of genotypes at multiple risk loci, with the potential for both epistatic effects and interactions with environmental risk factors. The sheer number of GWAS findings with significant marginal effects

supports extensive genetic and allelic heterogeneity. Interpreting GWAS results with the simple marginal effect model ignores these more complicated relationships and misrepresents the true genetic effect. So while GWAS are a powerful tool to initially scan the genome, further analyses on implicated variants are required to fully understand the genotype-phenotype relationship. Moreover, there is increasing theoretical and empirical evidence that low frequency variants play a major role in complex disease [11, 58, 59]. GWAS are underpowered for testing such variants because they often lack detectable marginal effects [34] and are therefore unlikely to detect individual low frequency variants.

It is clear that novel statistical methods are required to complement the GWAS and improve our understanding of complex human disease. In this dissertation, we propose statistical methods that explore genotype-phenotype models beyond the contribution of simple marginal effects. We consider an analysis for gene-environment interaction performed as a follow-up to a significant GWAS result. We propose a statistical test for detecting sets of rare risk variants that affect the same gene. Finally, we address strategies for imputing rare variants into large datasets.

Environment exposures are known to affect risk for many complex diseases. If genetic variation also contributes to the risk of disease, it is possible that the effect of the genetic risk variant can be modified by exposure to the environmental risk factors (ERF) [33]. From a statistical perspective, the genetic risk variant and the ERF are said to interact if the effect of each is dependent upon the other. This dependency may or may not involve an actual biological interaction, but the relationship is important nonetheless [10]. From a medical perspective, knowledge of the interaction can improve risk prediction, treatment and preventative measures contingent upon ERF exposure. Further, knowing the true form of the genotype-phenotype relationship allows a more accurate assessment of the contribution of the variant to overall disease prevalence.

Genetic risk variants that interact with an ERF can be identified by GWAS provided the marginal genetic effect of the locus is large enough. In the presence of a gene-environment interaction, the marginal effect of a risk allele is the average of the genetic effects over all levels of the ERF, weighted according to the frequency distribution of the ERF. However, considering only the marginal effect of a risk variant that interacts with an ERF can cloud the true genotype-phenotype relationship and down play the role of the risk variant in disease etiology. For example, if a risk variant has a large effect in the presence of a rare ERF but little to no effect in the absence of the ERF, the marginal genetic effect of the risk variant will be small. Concluding

that this variant has little effect on disease risk based on small marginal effect ignores the substantial role it plays for a certain portion of the population. In addition, estimates of marginal effect can be misleading if the frequency distribution of the ERF in the dataset from which estimated are derived is not representative of the wider population.

One of the major analysis challenges for gene-environment interactions is that there is no definitive definition or model. In fact, there are likely to be a wide range of genotype-environment-phenotype relations. As such, there is no consensus statistical testing strategy for gene-environment interactions. Both case-control and case-only population-based designs have been proposed. The case-control design directly tests for differences in the effect size for the genetic variant conditional on exposure [10]. The case-only design is more powerful but requires independence between genotype and exposure in the population. Type I error for a case-only analysis can be severely inflated if the assumption is not met [55]. Recently, efforts have been made to combine these strategies into a two-stage design [48]. However, it is often assumed that both an ERF and a risk variant are known in advance for a disease and a “candidate-interaction” analysis of the gene-environment pair can be performed. If a risk variant is not known in advance, putative variants may first be identified through GWAS. SNPs showing significance in the GWAS can then be subjected to a follow-up analysis for interaction with an ERF. Provided exposure data was collected along with genotype data, such a follow-up analysis is trivial to perform but is prone to ascertainment bias due to the requirement for a significant result in the initial GWAS test. Estimates of genetic effect based on the same data that gives a significant GWAS result are well-known to contain an upward bias known as the “Winner’s Curse.” [21, 72] Since GWAS typically test for significant marginal effects, the Winner’s Curse has been studied primarily for estimates of marginal genetic effect, with numerous correction methods proposed [79, 84, 85]. Although a similar bias is certain to exist on estimates for a model of gene-environment interaction, it is not clear how the GWAS requirement for a significant marginal effect will impact parameter estimates when a different genotype-phenotype model is considered in the follow-up.

In chapter 2 of the dissertation, we examine the Winner’s Curse phenomenon in studies of gene-environment interaction that are performed as a follow-up analysis to a GWAS. We consider two distinct testing strategies that require a genome-wide significant result in an initial GWAS-type test followed by formal testing and estimation of gene-environment parameters. We compare the bias on parameter estimation induced by each strategy and show that the choice of initial test can affect infer-

ence for the follow-up analysis. To reduce bias, we propose an extension of the ascertainment-corrected likelihood method introduced by Zöllner and Pritchard [85]. However, when applied to complicated models of gene-environment interaction, the corrected likelihood cannot be written in analytical form making standard maximization methods difficult to apply. Therefore we introduce a novel partial likelihood Markov Chain Monte Carlo algorithm capable of computing bias-reduced estimates from the intractable corrected likelihood function.

Another area we explore in this dissertation is the role of rare variation in complex disease. Due in part to previous limitations in both the knowledge of existing variation and genotyping technology, the search for the genetic causes of complex disease has previously focused primarily on common variation. The ambitious HapMap Project aimed to catalog common variation in the human genome and to date has identified 10 million common variants, mostly SNPs with minor allele frequency (maf) $\geq 5\%$ [2]. Armed with the position and alleles for these SNPs, genotyping platforms were built to allow large study samples to be assayed for hundreds of thousands to millions of HapMap SNPs across the genome, making the GWAS possible. Despite the number of SNPs that could be tested in a GWAS, the testable portion of the genome was primarily limited to common SNPs available in the HapMap. With the frequency spectrum for SNPs heavily skewed toward smaller frequencies, this leaves rare variation as a vastly understudied portion of the human genome [18].

The Common Disease-Rare Variant model proposes that some of the missing heritability for complex diseases can be explained by rare variants with large effect sizes [58, 59]. Under this model, the contribution of individual variants to population prevalence is small but the combined effect of numerous rare variants can account for an appreciable fraction of the prevalence. This model is feasible if risk variants are subject to weak purifying selection and is supported by the fact that allele frequencies for protein-altering mutations are more heavily skewed toward rare variants compared to neutral variants [22]. Despite the obvious potential for rare variants in complex disease genetics, technological limitations have hampered the ability to affordably assay rare variants in large population-based samples. However, the advent of next generation sequencing technology now provides the potential to detect all variation in a genomic region, particularly novel rare variants [47]. The reality of observing all genetic variation in large population-based samples will revolutionize our ability to study the contribution of rare genetic variation to complex disease; however, it will also bring numerous analytical challenges.

Individually testing each variant identified via sequencing in a large population-

based dataset is not a powerful strategy since power diminishes with decreasing allele frequencies and the necessary multiple testing correction may be prohibitively stringent [32]. Instead, statistical tests for excess accumulation of rare variation within a genomic region in either cases or controls, so-called burden tests, have been shown to be a more powerful alternative to single marker tests [34, 41]. Among the numerous proposed burden tests, the common feature is to accumulate marginal evidence for multiple rare variants within the same functional unit (e.g. a gene) into a single statistic. The idea is that, individually, each marginal p-value may be insufficient to declare an association but the p-value of their combined effect in a burden test may reach statistical significance.

In chapter 3 of the dissertation, we introduce an intuitive and computationally efficient burden statistic, the Cumulative Minor Allele Test (CMAT), to identify genomic regions containing rare variation that contribute to disease risk. We assess the performance of the CMAT and other pooling methods on datasets simulated using population genetic models to contain realistic levels of neutral variation. We consider study designs ranging from exon-only to whole-gene analyses that contain non-coding variants. For all study designs considered, the CMAT achieves comparable power to previously proposed methods. A unique advantage of the CMAT is that it easily extends to probabilistic genotypes, allowing application to low-coverage sequencing and imputation data. We illustrate that augmenting sequence data with imputed samples is a practical method to increase power for rare variant studies. We also provide a method to control for confounding variables such as population stratification. Finally, we demonstrate that our method is capable of analyzing rare variants imputed into existing GWAS datasets using external imputation templates. As proof of principle, we performed a CMAT analysis of over 8 million SNPs imputed into the GAIN psoriasis dataset[49] using haplotypes from the 1000 Genomes Project [18]. In our analysis, one gene, *SKIV2L*, maintained a significant test statistic after correcting for multiple testing. The gene is located near the *HLA* region on chromosome 6 thought to harbor multiple psoriasis susceptibility genes[49]. The CMAT statistic for *SKIV2L* contained multiple rare variants ($\text{maf} < 1\%$), none of which achieved genome-wide significance in a single marker test, indicating that the CMAT can identify potentially interesting genes that may otherwise be missed by single marker GWAS.

The previous CMAT analysis on the psoriasis data relied upon genotype imputation, the estimation of genotypes at untyped markers using known patterns of haplotype structure [37]. Imputation has already proven to be a powerful tool in modern genetic studies by increasing the genomic coverage of GWAS and allowing for large-

scale meta-analyses [74, 82]. As we show in the second chapter, imputation promises to be as important in future genetic studies that involve sequencing technology to target rare variation. In particular, power for rare variant burden tests of association can be increased by augmenting sequencing datasets with imputed samples [81]. However, statistical power in genetic association studies is known to be affected by imputation accuracy [28]. Thus, it is of interest to design imputation strategies that optimize accuracy.

The imputation procedure involves a set of target samples in which genotypes are to be imputed and a reference panel of phased haplotypes from which genotypes are copied. Imputation accuracy is known to depend on several factors including reference panel size, diversity, and genetic similarity to the imputation targets [2, 6, 27, 29]. Imputation is likely to be most accurate when the reference panel is drawn from the same population as the target samples, or one that is closely related [2, 26, 39]. Haplotypes from the publicly available HapMap [2] and 1000 Genomes Projects [18] currently serve as the imputation reference panel in most genetic studies despite the fact that these haplotypes are usually not derived from the same population as the imputation targets. Still, imputation accuracy is quite high using these panels, particularly for common variants [2]. In the near future, however, next generation sequencing will allow the creation of custom-made reference panels by sequencing a subset of a large dataset to use as templates for imputing variation into the remaining samples. Creation of these custom panels may be costly, so it is of interest to determine the expected improvement in imputation accuracy for using custom-made panels compared to the large, publicly available panels.

In chapter 4 of the dissertation, we propose a theoretical model of imputation based on coalescent theory [30]. We use the coalescent to model the genealogical history of an imputation target haplotype and sets of reference haplotypes from potential reference panels. We define a rule based on coalescence time to determine the reference haplotype that is expected to provide the best template for imputing untyped alleles on the target haplotype. Based on this rule, we derive analytic equations to quantify imputation accuracy for a given reference panel. The coalescent framework facilitates complex population demographic models, allowing a range of imputation study designs to be evaluated. Here, we use the model to compare imputation accuracy between custom-made reference panels versus large, publicly available panels. We find that custom panels from the same underlying population as the imputation target, even when considerably smaller than competing publicly available panels from closely related populations, are nearly always the optimal choice for imputation. The

relative improvement in imputation accuracy gained by using a custom panel varies according to both the divergence time between and the growth rates for the populations from which the panels are drawn. Improvement in imputation accuracy for the custom panels increases as the divergence time between the two populations increases. However, exponential growth in the populations from which the panels are drawn attenuates the effect of the divergence, slightly reducing the improvement in accuracy for custom panels. Thus, for populations experiencing exponential growth and separated by small divergence times, imputation accuracy based on large, publicly available panels can be comparable to moderately sized custom panels. Our results suggest that future imputation-based genetic studies will benefit from custom reference panels originating from the same population as the samples to be imputed. However, a large publicly available reference panel (≥ 500 haplotypes) can provide similar accuracy provided the reference haplotypes are not too distantly related to the haplotypes to be imputed.

The projects in this dissertation are a next step in the search for the underlying genetic causes of complex human disease. Our analysis method for gene-environment interactions is designed to clarify genotype-phenotype associations previously detected by GWAS. The CMAT is an alternative method to identify rare risk variants unlikely to be detected by GWAS. Finally, our study of imputation strategies provides practical guidelines for designing powerful imputation-based genetic studies in the age of next-generation sequencing.

CHAPTER II

The Winner's Curse in Studies of Gene-Environment Interaction

2.1 Introduction

The Genome-wide association study (GWAS) is a powerful tool for identifying genetic risk variants for complex human disease [65]. The basic design of the GWAS is to individually test hundreds of thousands to millions of Single Nucleotide Polymorphisms (SNPs) across the genome in a set of unrelated samples for association with a phenotype of interest. The primary goal is to identify putative risk variants for the disease but a secondary goal is to estimate the effect size of any locus that achieves statistical significance. Due to the large number of tests performed in a GWAS, the same statistical test is generally applied to all SNPs in an automated fashion. Here the emphasis is on the discovery of putative risk variants rather than accurate modeling of the genotype-phenotype relation for any particular SNP. Thus, the common choice for a GWAS is a simple test for significant marginal genetic effect since it is powerful to declare associations across a range of potential underlying genetic models. Given that a SNP shows significance in the initial GWAS test, follow-up analyses may attempt to replicate the initial finding as well as characterize and quantify the true form of the association [8]. Although the GWAS may indicate a significant marginal genetic effect, the follow-up analysis could test for a more complicated genotype-phenotype relationship. For diseases with a known environmental risk factor, an interesting follow-up analysis is to determine if any of the implicated genetic risk factors interact with the environmental risk factor. Although the statistical interaction does not necessarily reflect an actual biological interaction, the relationship is important nonetheless and can be useful from a medical perspective to prescribe treatment and preventative measures contingent upon environmental exposure [10].

A major challenge to this type of follow-up analysis is to avoid ascertainment bias induced by requiring a significant result in the initial GWAS test. Estimates of genetic effect based on the same data that gave a significant GWAS result are well-known to contain an upward bias known as the “Winner’s Curse” [21, 72]. The reason for the bias is that effect size and significance level for a dataset are highly correlated. Requiring a statistically significant test result restricts estimation to datasets with larger effect estimates. The extent of the bias is directly related to the statistical power of the initial test, with bias increasing as power for the initial test decreases. The stringent significance threshold required for GWAS ($p < 5 \times 10^{-8}$) therefore makes effect estimates from GWAS data especially prone to overestimation. The Winner’s Curse phenomenon and its impact on GWAS parameter estimates is well-appreciated in the field of statistical genetics [21, 72, 84]. It has been cited as a potential cause of replication failure for true risk variants due to underestimation of replication sample size requirements [46]. Characterizing and correcting for the Winner’s Curse effect in GWAS is an active area of research [83, 79, 85]. The numerous statistical correction methods that have been proposed typically assume that the effect estimate of interest is the same effect that is tested for in the GWAS, for example, an odds ratio for carriers of the risk allele in a logistic regression [83]. From the literature, it is clear that conditioning on a statistically significant effect parameter will lead to upward bias in estimates of that particular effect parameter. It is less clear how the Winner’s Curse will affect parameter estimates in a follow-up analysis for a model that is different from the one tested in the initial GWAS. In this dissertation chapter, we examine the effect of the Winner’s Curse on parameter estimation and hypothesis testing for models of gene-environment interaction performed as a follow-up analysis to a significant GWAS result. We propose a general correction method that accounts for the initial GWAS test and can estimate penetrance parameters for models that differ from the initial test.

In the first part of the chapter, we focus on “naive” parameter estimation that ignores the requirement for a significant GWAS result. We use simulation to derive the sampling distributions for naive maximum likelihood estimates of a gene-environment interaction model. We consider two testing strategies for the initial GWAS: a chi-square test for marginal allelic association that ignores an environmental risk factor, and a logistic regression that tests for either a significant genetic main effect or a gene-environment interaction. We purposefully simulate data from a gene-environment model that differs from the modeling assumptions of the two GWAS tests. We show that requiring a significant GWAS result leads to a Win-

ner’s Curse effect on parameters for the gene-environment model. However, the two tests bias the follow-up analysis in different ways. Notably, the chi-square test for marginal genetic effect inflates the false positive rate for a follow-up hypothesis test of interaction when no interaction between the genotype and environmental exposure exists. We conclude that the gene-environment follow-up study design requires an ascertainment-correction method to allow valid analysis.

In the second part of the chapter, we propose a method to correct for the Winner’s Curse effect observed in a follow-up analysis of gene-environment interaction. To do so, we extend the likelihood-based correction method described by Zöllner and Pritchard that conditions on a significant initial GWAS result [85]. It is shown that the conditional likelihood function can be simplified to the ratio of the unconditional likelihood and the power of initial GWAS test. For their analysis, Zöllner and Pritchard assumed that the power function in the denominator could be written in a closed form equation. However, for many underlying disease models this assumption is not realistic. For example, if the initial test is a logistic regression but the underlying disease model does not follow the form of the regression, closed form power functions do not exist. Though the conditional likelihood function is still valid, without an analytic form for the denominator most standard maximization methods are not applicable. Power can be estimated by simulation but this is computationally prohibitive if the quantity needs to be repeatedly computed.

To obtain corrected estimates from the conditional likelihood function, we introduce a novel Markov Chain Monte Carlo (MCMC) algorithm, the partial likelihood MCMC. Our proposed algorithm is a modification of the MCMC without likelihood algorithm described by Marjoram et al. for inference on likelihood functions that are either impossible or computationally prohibitive to compute [45]. Movement in the MCMC without likelihoods algorithm is determined in part by a rejection sampling step that eliminates the need to evaluate the likelihood function. Instead, given a proposed step for the chain, data is simulated from the likelihood function based on the proposed parameter and the step is rejected if the simulated data does not match the observed data. If the simulated data is comparable to the observed data, a Metropolis-Hastings step involving only the proposal density (and the prior, if one exists) is performed. Thus the MCMC without likelihoods can be used to derive the sampling distribution (or posterior distribution) of parameters without the need to ever evaluate a complicated likelihood function.

In the case of the conditional likelihood function of Zöllner and Pritchard, only the denominator is computationally prohibitive and it is preferable to incorporate

the computable numerator in parameter estimation. Thus, the partial likelihood MCMC is specifically designed to account for the intractable denominator while also including information from the numerator in the algorithm. Similar to the MCMC without likelihood, we include simulation-rejection step in our algorithm. However, we evaluate the tractable portion of the likelihood in a Metropolis-Hastings-like step. The simulation-rejection step involves simulating data according to the unconditional likelihood function, testing the data with the initial GWAS test and allowing a move only if the dataset is statistically significant. Given the effect parameters for the current state of the chain, the probability that a dataset drawn according to those parameters is statistically significant in the GWAS test is precisely the statistical power for those parameters. Thus, our simulation-rejection step follows directly from the intractable portion of the conditional likelihood. The main idea behind the partial likelihood MCMC is that the rejection sampling forces the chain to explore portions of the parameter space containing low power for the GWAS test. Effect sizes for these low-power parameters will generally be smaller and therefore reduce estimates of effect.

We formally prove that the partial likelihood MCMC has as its stationary distribution the corrected likelihood function. In applying the algorithm, we produce a random sample of quasi-independent realizations from the corrected likelihood function and use a summary statistic of the sample as corrected parameter estimates. We show that for effect parameter values that have low or moderate power in the initial GWAS, the corrected estimates reduce Winner’s Curse bias and provide more accurate power estimates for a replication analysis. Parameter values that have high power in the initial GWAS can actually be over-corrected resulting in an underestimate of the effect.

2.2 Model

Let \mathcal{M} denote a disease penetrance model for a biallelic locus and dichotomous environmental risk factor (Table 2.1). Let $g \in \{0, 1, 2\}$ be the number of risk alleles at the locus and $e \in \{0, 1\}$ an indicator for exposure to the environmental risk factor. Let A and U indicate affected cases and unaffected controls for disease status, respectively. To simplify notation, we define an ordering on the six genotype-exposure combinations as follows. Let $ge_i, 0 \leq i \leq 5$, be the i^{th} genotype-exposure category where $i = 2^g + e$. Then the model \mathcal{M} is specified by a vector of penetrance values $\boldsymbol{\theta} = \{\theta_0, \theta_1, \dots, \theta_5\}$, where $\theta_i = P(A|ge_i) \in [0, 1]$, and a frequency vector $\boldsymbol{\phi} = \{\phi_0, \phi_1, \dots, \phi_5\}$, where

$\phi_i = P(ge_i) \in [0, 1]$ and $\sum_{i=0}^5 \phi_i = 1$. Given $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, the population disease prevalence for \mathcal{M} is defined to be $F = \boldsymbol{\phi}'\boldsymbol{\theta} = \sum_{i=0}^5 P(A|ge_i) P(ge_i)$.

i	Genotype	Exposure	Frequency $\boldsymbol{\phi}$	Penetrance $\boldsymbol{\theta}$
$2^g + e$	g	e	$P(ge_i)$	$P(A ge_i)$
0	0	0	ϕ_0	θ_0
1	0	1	ϕ_1	θ_1
2	1	0	ϕ_2	θ_2
3	1	1	ϕ_3	θ_3
4	2	0	ϕ_4	θ_4
5	2	1	ϕ_5	θ_5

Table 2.1: Gene-environment penetrance model \mathcal{M}

Given the penetrance model \mathcal{M} , let $D = \{a_0, a_1, \dots, a_5, u_0, u_1, \dots, u_5\}$ denote the data configuration for a case-control dataset where a_i and u_i are the number of cases and controls, respectively, in the i^{th} genotype-exposure category. Then $N_a = \sum_{i=0}^5 a_i$ is the total number of cases and $N_u = \sum_{i=0}^5 u_i$ is the total number of controls in the dataset. Assuming a retrospective case-control sampling design, the counts of cases within the genotype-exposure categories $\{a_0, a_1, \dots, a_5\}$ follows a multinomial distribution with the probabilities for each genotype-exposure category defined by

$$P(ge_i | A, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{P(A|ge_i) \times P(ge_i|\boldsymbol{\phi})}{F} = \frac{\theta_i \times \phi_i}{F}, \quad (2.1)$$

Likewise, the genotype-exposure counts for controls $\{u_0, u_1, \dots, u_5\}$ follow a multinomial distribution with sampling probabilities given by

$$P(ge_i | U, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{P(U|ge_i) \times P(ge_i|\boldsymbol{\phi})}{1 - F} = \frac{(1 - \theta_i) \times \phi_i}{1 - F}. \quad (2.2)$$

Assuming independence between cases and controls, the probability of the full data configuration D , conditional on the model \mathcal{M} , is

$$P(D | \boldsymbol{\theta}, \boldsymbol{\phi}) \propto \prod_{i=0}^5 P(ge_i|A, \boldsymbol{\theta}, \boldsymbol{\phi})^{a_i} \times P(ge_i|U, \boldsymbol{\theta}, \boldsymbol{\phi})^{u_i} \quad (2.3)$$

From an inference standpoint, if D is an observed random sample and our interest is in estimating $\boldsymbol{\theta}$, then the likelihood function for $\boldsymbol{\theta}$ conditional on D is

$$L(\boldsymbol{\theta} | D, \boldsymbol{\phi},) = P(D | \boldsymbol{\theta}, \boldsymbol{\phi}). \quad (2.4)$$

Next we assume the data configuration D is collected as part of a GWAS and is subjected to an association test at a genome-wide significance level. We assume that penetrance parameters will be estimated from D only if D is statistically significant in the initial GWAS test. Otherwise, D is not of interest and the penetrance parameters are not estimated. In this case, the set of data configurations for which penetrance parameters are estimated is not representative of the true underlying distribution of all possible data configurations. Simply put, D is not a random sample from Eq. (2.4) because it is required to have a significant GWAS result.

Following Zöllner and Pritchard [85], we derive the likelihood function for penetrance parameters $\boldsymbol{\theta}$ conditional on the data configuration D being statistically significant in the initial GWAS test. Let S_α denote that the dataset D is statistically significant at level α for the initial test of association. Then the conditional likelihood $L(\boldsymbol{\theta}|D, S_\alpha)$ that accounts for data D being significant in the initial test can be written as

$$L(\boldsymbol{\theta}|D, \boldsymbol{\phi}, S_\alpha) = P(D|S_\alpha, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{P(S_\alpha|D, \boldsymbol{\theta}, \boldsymbol{\phi}) \times P(D|\boldsymbol{\theta}, \boldsymbol{\phi})}{P(S_\alpha|\boldsymbol{\theta}, \boldsymbol{\phi})} = \frac{P(D|\boldsymbol{\theta}, \boldsymbol{\phi})}{P(S_\alpha|\boldsymbol{\theta}, \boldsymbol{\phi})} \quad (2.5)$$

where $P(S_\alpha|D, \boldsymbol{\theta}, \boldsymbol{\theta}) = 1$ since we are restricting attention to datasets that are statistically significant. Thus, the conditional likelihood $L(\boldsymbol{\theta} | D, \boldsymbol{\phi}, S_\alpha)$ reduces to the ratio of $P(D|\boldsymbol{\theta}, \boldsymbol{\phi})$, the unconditional likelihood of $\boldsymbol{\theta}$ (eq. 2.4), and $P(S_\alpha|\boldsymbol{\theta}, \boldsymbol{\phi})$, the statistical power of the initial test at level α .

2.3 Statistical methods

2.3.1 Prevalence constraint

For simplicity, we focus on estimating $\boldsymbol{\theta}$ from the data D and assume that $\boldsymbol{\phi}$ and F are fixed constants known from an independent source, for example a different data set. Under the model \mathcal{M} , fixing $\boldsymbol{\phi}$ and F places a constraint on the parameter space of $\boldsymbol{\theta}$ since, by definition, $F = \boldsymbol{\phi}'\boldsymbol{\theta}$. For fixed values of F and $\boldsymbol{\phi}$, the hyper-plane given by

$$\Theta_{F, \boldsymbol{\phi}} = \{\boldsymbol{\theta} \in [0, 1]^6 | F = \boldsymbol{\phi}'\boldsymbol{\theta}\} \quad (2.6)$$

defines the set of $\boldsymbol{\theta}$ values that satisfy the prevalence constraint. Since we assume throughout that F and $\boldsymbol{\phi}$ are known constants, estimation methods for the penetrance parameter $\boldsymbol{\theta}$ must be performed over the space $\Theta_{F, \boldsymbol{\phi}}$. Since $\boldsymbol{\phi}$ is assumed constant, we hereafter drop the $\boldsymbol{\phi}$ term from the likelihood functions (eq. 2.4 and 2.5) to simplify

notation. All methods described can be modified to include estimation of ϕ .

2.3.2 Naive estimate of θ

We assume that penetrance parameters θ are estimated only for data configurations D that are statistically significant in the initial test of association. Given a statistically significant data configuration D , we define a naive estimate of θ to be any estimate that ignores the significance requirement and instead treats D as a true random sample. Thus naive estimates of θ are based on the unconditional likelihood function $L(\theta | D)$ that ignores the GWAS ascertainment (eq. 2.4). Here, we consider naive maximum likelihood estimates (MLEs) of θ obtained by maximizing the unconditional likelihood function subject to the prevalence constraint (eq. 2.6). Thus, naive MLEs of θ are defined as

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta_{F,\phi}} L(\theta | D).$$

We describe a stochastic grid search algorithm to compute the naive MLE's $\hat{\theta}$ in the Appendix (Section 2.7.1).

2.3.3 Corrected estimate of θ

We define corrected estimates of θ to be estimators based on the conditional likelihood function (eq. 2.5). Obtaining estimates from the conditional likelihood function using standard maximization methods is difficult because the term in the denominator, $P(S_\alpha | \theta)$, cannot be written in analytical form for many testing scenarios. For example, if logistic regression is used for the GWAS test, the non-central chi-square distributed power function is valid only if the data follows the form of the regression [70]. Given a penetrance vector θ , $P(S_\alpha | \theta)$ can be estimated using Monte Carlo simulation and a numerical method such as the grid search algorithm (Section 2.7.1) can be used to maximize the conditional likelihood. This approach can be computationally prohibitive because it requires repeatedly estimating power for the initial test.

2.3.4 Partial likelihood Markov Chain Monte Carlo

We introduce the partial likelihood Markov Chain Monte Carlo (MCMC) algorithm to obtain estimates from the conditional likelihood $L(\theta | D, S_\alpha)$. Below we give

the steps for the partial likelihood MCMC algorithm and then prove that it converges to the conditional likelihood $L(\boldsymbol{\theta}|D, S_\alpha)$.

Due to the prevalence constraint, the parameter space for $\boldsymbol{\theta}$ is $\Theta_{F,\phi}$. Let $q : \Theta_{F,\phi} \times \Theta_{F,\phi} \rightarrow [0, 1]$ be a valid probability density for proposing new points on $\Theta_{F,\phi}$, conditional on the current state of the chain. Then, assuming the chain is currently at $\boldsymbol{\theta}$, let $q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*)$ denote the probability of proposing a move to $\boldsymbol{\theta}^*$. The specific proposal density we employed is described in detail below (2.3.5). Assuming a valid proposal density q , the partial likelihood MCMC algorithm proceeds as follows:

Algorithm 1:

1. If currently at $\boldsymbol{\theta}_i \in \Theta_{F,\phi}$, draw D^* according to $P(D | \boldsymbol{\theta}_i)$.
2. If D^* is statistically significant at level α , draw $\boldsymbol{\theta}^* \in \Theta_{F,\phi}$ according to $q(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}^*)$.
Otherwise set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$ and return to step 1.
3. Calculate $h = h(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*) = \min \left\{ 1, \frac{P(D|\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta}_i)}{P(D|\boldsymbol{\theta}_i) q(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}^*)} \right\}$.
4. Set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}^*$ with probability h , otherwise set $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$ and return to step 1.

We next prove that the chain of vector penetrance values $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \dots\}$ produced by the partial likelihood MCMC has as its stationary distribution the desired conditional likelihood function $L(\boldsymbol{\theta} | D, S_\alpha)$.

Theorem The stationary distribution of the partial likelihood MCMC is $P(\boldsymbol{\theta} | D, S_\alpha)$.

Proof: By construction, $r(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*) = P(S_\alpha|\boldsymbol{\theta}) q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*) h(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the transition probability of the chain. It is sufficient to show that $P(\boldsymbol{\theta} | D, S_\alpha)$ and r satisfy the detailed balance equation

$$P(\boldsymbol{\theta} | D, S_\alpha) r(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*) = P(\boldsymbol{\theta}^* | D, S_\alpha) r(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta}).$$

Consider $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$ such that

$$\frac{P(D | \boldsymbol{\theta}^*) q(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta})}{P(D | \boldsymbol{\theta}) q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*)} \leq 1.$$

Then,

$$\begin{aligned}
P(\boldsymbol{\theta} \mid D, S_\alpha) r(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*) &= \frac{P(D \mid \boldsymbol{\theta})}{P(S_\alpha \mid \boldsymbol{\theta})} P(S_\alpha \mid \boldsymbol{\theta}) q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*) h(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \\
&= P(D \mid \boldsymbol{\theta}) q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*) \frac{P(D \mid \boldsymbol{\theta}^*) q(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta})}{P(D \mid \boldsymbol{\theta}) q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*)} \\
&= P(D \mid \boldsymbol{\theta}^*) q(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta}) \\
&= P(D \mid \boldsymbol{\theta}^*) q(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta}) \frac{P(S_\alpha \mid \boldsymbol{\theta}^*)}{P(S_\alpha \mid \boldsymbol{\theta}^*)} h(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\
&= \frac{P(D \mid \boldsymbol{\theta}^*)}{P(S_\alpha \mid \boldsymbol{\theta}^*)} P(S_\alpha \mid \boldsymbol{\theta}^*) q(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta}) h(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \\
&= P(\boldsymbol{\theta}^* \mid D, S_\alpha) r(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta}).
\end{aligned}$$

Now assume $\frac{P(D|\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*\rightarrow\boldsymbol{\theta})}{P(D|\boldsymbol{\theta})q(\boldsymbol{\theta}\rightarrow\boldsymbol{\theta}^*)} > 1$. Then $h(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = 1$ and similar steps show that the equality holds. \square

The partial likelihood MCMC can be used to generate a random sample from a general function that contains an intractable portion. The algorithm makes use of the observed data by evaluating the tractable portion of the function in a standard Metropolis-Hastings step (step 3) yet accounts for the intractable portion using a simulation-rejection sampling step (step 2). As presented, the partial likelihood MCMC generates a chain of penetrance vectors that converges to $L(\boldsymbol{\theta} \mid D, S_\alpha)$ without requiring the power function $P(S_\alpha \mid \boldsymbol{\theta})$ to be repeatedly computed. The computational requirements for the partial likelihood MCMC are: (1) given a penetrance vector $\boldsymbol{\theta}$, we can simulate data configurations according to the unconditional sampling distribution $P(D \mid \boldsymbol{\theta})$ (eq. 2.3) and (2) we can subject the simulated data set to the same test used in the GWAS. Sampling data configurations using equation (2.3) is computationally straightforward since the unconditional likelihood is simply the product of multinomial distributions.

A challenge in implementing the partial likelihood MCMC is that if the current state of the chain is a set of penetrance parameters that have very low power it is possible that a statistically significant dataset is never drawn and the chain remains stuck at this parameter vector. To avoid this situation, we limit the number of times that the simulation-rejection step is performed on the same parameter vector by forcing a new vector to be proposed and performing a standard Metropolis-Hastings step that includes the complete unconditional likelihood. That is, assume that say

100 datasets have been simulated and rejected for the current vector $\boldsymbol{\theta}_i$. We then propose a new vector $\boldsymbol{\theta}^*$ according to $q(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}^*)$ and compute

$$h(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*) = \min \left\{ 1, \frac{P(D|\boldsymbol{\theta}^*) P(S_\alpha|\boldsymbol{\theta}_i) q(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta}_i)}{P(D|\boldsymbol{\theta}_i) P(S_\alpha|\boldsymbol{\theta}^*) q(\boldsymbol{\theta}_i \rightarrow \boldsymbol{\theta}^*)} \right\}$$

where $P(S_\alpha|\boldsymbol{\theta}_i)$ and $P(S_\alpha|\boldsymbol{\theta}^*)$ are estimated by simulation. We complete the Metropolis-Hastings step by accepting the proposal with probability $h(\boldsymbol{\theta}_i, \boldsymbol{\theta}^*)$. We find that this inclusion encourages the chain to move from areas with low power penetrance values.

Given the chain of penetrance values $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \dots\}$ produced by the partial likelihood MCMC, we create a random sample of quasi-independent penetrance vectors from the corrected likelihood $L(\boldsymbol{\theta}|D, S_\alpha)$ by removing the first $B - 1$ vectors to allow the chain to reach its stationary distribution (burn-in) and then only retaining every T^{th} vector to reduce correlation (thinning). Thus, from the full chain $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \dots, \boldsymbol{\theta}_K\}$, we select the sub-chain $\{\boldsymbol{\theta}_B, \boldsymbol{\theta}_{B+T}, \boldsymbol{\theta}_{B+2T}, \dots, \boldsymbol{\theta}_{B+kT}\}$ as the random sample. Given this random sample, we estimate θ_i , $0 \leq i \leq 5$, the penetrance value for i^{th} genotype-exposure category, with

$$\bar{\theta}_i = \frac{1}{k+1} \sum_{j=0}^k \theta_{B+jT,i}$$

where $\theta_{B+jT,i}$ is the penetrance value for the i^{th} genotype-exposure category in the $B + jT^{th}$ penetrance vector produced by the partial likelihood MCMC. Thus, we take the marginal means for the six genotype-exposure categories from our random sample as corrected estimates of $\boldsymbol{\theta}$ and define

$$\bar{\boldsymbol{\theta}} = \{\bar{\theta}_0, \bar{\theta}_1, \dots, \bar{\theta}_5\}. \quad (2.7)$$

Since $\phi' \boldsymbol{\theta}_{B+jT} = F$ for each $\boldsymbol{\theta}_{B+jT}$ in the random sample,

$$\begin{aligned}
\phi' \bar{\boldsymbol{\theta}} &= \sum_{i=0}^5 \phi_i \bar{\theta}_i \\
&= \sum_{i=0}^5 \phi_i \frac{1}{k+1} \sum_{j=0}^k \theta_{B+jT,i} \\
&= \frac{1}{k+1} \sum_{j=0}^k \sum_{i=0}^5 \phi_i \theta_{B+jT,i} \\
&= \frac{1}{k+1} \sum_{j=0}^k F \\
&= F.
\end{aligned}$$

Thus $\bar{\boldsymbol{\theta}} \in \Theta_{F,\phi}$ and our corrected point estimate satisfies the prevalence constraint.

2.3.5 Proposal density $q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*)$

Here we derive a proposal density for the space $\Theta_{F,\phi}$. Assume that the current state of the chain is the vector $\boldsymbol{\theta}$. Let $\epsilon > 0$ be a fixed constant and let

$$\partial(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

denote Euclidean distance. We define $PR(\boldsymbol{\theta}) = \{\boldsymbol{\theta}^* \in \Theta_{F,\phi} \mid \partial(\boldsymbol{\theta}, \boldsymbol{\theta}^*) < \epsilon\}$, the set of valid penetrance vectors that satisfy the prevalence constraint and are within ϵ of $\boldsymbol{\theta}$, to be the proposal region about $\boldsymbol{\theta}$. In the following, we describe a computational algorithm to sample uniformly from $PR(\boldsymbol{\theta})$ by first drawing a direction and then a distance to define the proposal.

First, draw a random point $\mathbf{p} \in \mathbb{R}^6$ such that $F = \phi' \mathbf{p}$. Note that the point \mathbf{p} lies on the extended plane defined by the prevalence constraint $\Theta_{F,\phi}$, but need not be a valid penetrance vector (i.e. $0 \leq p_i \leq 1, \forall i$). Next compute the unit vector $\mathbf{d} = \frac{\boldsymbol{\theta} - \mathbf{p}}{\partial(\boldsymbol{\theta}, \mathbf{p})}$. The vector \mathbf{d} will indicate the proposal direction. To determine the proposal magnitude, first let $m = \min\{\theta_0, \theta_1, \dots, \theta_5, 1 - \theta_0, 1 - \theta_1, \dots, 1 - \theta_5\}$ be the minimum distance between the penetrance vector $\boldsymbol{\theta}$ and the edge of the parameter space $([0, 1]^6)$. Then draw the step size δ from $\text{Uniform}[0, \min\{m, \epsilon\}]$. Note that δ is allowed to be as large as ϵ only if $\boldsymbol{\theta}$ is more than ϵ away from the boundaries. Otherwise, δ is at most the minimum distance between $\boldsymbol{\theta}$ and the boundary.

Let $\boldsymbol{\theta}^* = \boldsymbol{\theta} + \delta \mathbf{d}$ and note that $\boldsymbol{\theta}^* \in PR(\boldsymbol{\theta})$ since $\partial(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \delta < \epsilon$ and

$$\begin{aligned}\phi' \boldsymbol{\theta}^* &= \phi' \boldsymbol{\theta} + \delta \phi' \left(\frac{\boldsymbol{\theta} - \mathbf{p}}{\partial(\boldsymbol{\theta}, \mathbf{p})} \right) \\ &= F + \frac{\delta}{\partial(\boldsymbol{\theta}, \mathbf{p})} (\phi' \boldsymbol{\theta} - \phi' \mathbf{p}) \\ &= F.\end{aligned}$$

because $\phi' \boldsymbol{\theta} = \phi' \mathbf{p} = F$. Since both the directional vector \mathbf{d} and the proposal magnitude δ were uniformly chosen, $\boldsymbol{\theta}^*$ is a uniform random draw from the proposal region $PR(\boldsymbol{\theta})$. It follows that the proposal density has the form

$$q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*) = \begin{cases} V(PR(\boldsymbol{\theta}))^{-1}, & \boldsymbol{\theta}^* \in PR(\boldsymbol{\theta}), \\ 0, & \textit{otherwise} \end{cases} \quad (2.8)$$

where $V(PR(\boldsymbol{\theta}))$ is the volume of the proposal region $PR(\boldsymbol{\theta})$, which depends on m , the minimum distance of $\boldsymbol{\theta}$ to the boundary of the parameter space. If $m \geq \epsilon$, $V(PR(\boldsymbol{\theta}))$ is the volume of the intersection between the plane $\Theta_{F,\phi}$ and a six dimensional ϵ -ball centered at $\boldsymbol{\theta}$. If $m < \epsilon$, $V(PR(\boldsymbol{\theta}))$ is the the volume of the intersection of the plane $\Theta_{F,\phi}$ and a six dimensional m -ball centered at $\boldsymbol{\theta}$. In either case, an analytical computation of $V(PR(\boldsymbol{\theta}))$ is difficult.

However, since the proposal density q appears in the Metropolis-Hastings ratio (Algorithm 1, step 3) as

$$\frac{q(\boldsymbol{\theta}^* \rightarrow \boldsymbol{\theta})}{q(\boldsymbol{\theta} \rightarrow \boldsymbol{\theta}^*)} = \frac{V(PR(\boldsymbol{\theta}))}{V(PR(\boldsymbol{\theta}^*))}, \quad (2.9)$$

we need only compute the ratio of the two densities rather than the actual value of each density. If the minimum distance between the boundary and both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ is at least ϵ , then $V(PR(\boldsymbol{\theta})) = V(PR(\boldsymbol{\theta}^*))$ and the ratio is one. If either or both of the minimum distances is less than ϵ , we use rejection sampling to obtain a Monte Carlo estimate of the ratio as follows.

Let $B_{\boldsymbol{\theta},\epsilon} = \{\mathbf{b} \in \mathbb{R}^6 \mid \phi' \mathbf{b} = F, \partial(\mathbf{b}, \boldsymbol{\theta}) \leq \epsilon\}$. Then $B_{\boldsymbol{\theta},\epsilon}$ is a superset of $PR(\boldsymbol{\theta})$ containing vectors less than ϵ from $\boldsymbol{\theta}$ that satisfy the prevalence constraint but are not required to fulfill the $[0, 1]$ boundaries for the individual vector components. Note that if the minimum distance between $\boldsymbol{\theta}$ and the $[0, 1]^6$ boundary is greater than ϵ then $B_{\boldsymbol{\theta},\epsilon} = PR(\boldsymbol{\theta})$. Let $V(B_{\boldsymbol{\theta},\epsilon})$ be the volume of $B_{\boldsymbol{\theta},\epsilon}$. It is important to note that $V(B_{\boldsymbol{\theta}_i,\epsilon}) = V(B_{\boldsymbol{\theta}_j,\epsilon}), \forall \boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \Theta_{F,\phi}$ since the volume of the intersection between the ϵ -ball and the extended $\Theta_{F,\phi}$ plane is the same regardless of where the ϵ -ball is

centered. Then for any $\boldsymbol{\theta}$, we can estimate the ratio $V(PR(\boldsymbol{\theta}))/V(B_{\boldsymbol{\theta},\epsilon})$ by sampling vectors from $B_{\boldsymbol{\theta},\epsilon}$ and tallying the count that are also in $PR(\boldsymbol{\theta})$. Thus for $\boldsymbol{\theta}, \boldsymbol{\theta}^* \in \Theta_{F,\phi}$, we can estimate Eq. (2.9) in the Metropolis-Hasting ratio as follows

$$\frac{V(PR(\boldsymbol{\theta}))}{V(PR(\boldsymbol{\theta}^*))} = \frac{\frac{V(PR(\boldsymbol{\theta}))}{V(B_{\boldsymbol{\theta},\epsilon})}}{\frac{V(PR(\boldsymbol{\theta}^*))}{V(B_{\boldsymbol{\theta}^*,\epsilon})}}$$

2.4 Simulation methods

In this section, we describe our simulation method to derive the sampling distributions for naive estimates of $\boldsymbol{\theta}$ that ignore the GWAS significance requirement and corrected estimates based on the partial likelihood MCMC algorithm. Thus far we have defined the penetrance vector $\boldsymbol{\theta}$ and the frequency vector $\boldsymbol{\phi}$ in as general terms to allow versatility of the methods. However, we simulate datasets according to an additive penetrance model of gene-environment interaction, denoted \mathcal{M}_+ .

For simplicity, we assume a risk allele frequency of p with Hardy-Weinberg proportions at the locus, an environmental exposure frequency of f and independence between genotype and exposure status. This allows the elements of $\boldsymbol{\phi}$ to be computed as functions of p and f . For the penetrance vector $\boldsymbol{\theta}$, we assume a baseline penetrance of m for unexposed individuals homozygous for the non-risk allele with additive genetic (d) and environmental main effects (e) and an interaction (e') within exposed individuals homozygous for the risk allele. Parameters for the additive penetrance model \mathcal{M}_+ are summarized in Table 2.2.

i	Genotype	Exposure	Frequency $\boldsymbol{\phi}$	Penetrance $\boldsymbol{\theta}$
$2^g + e$	g	e	$P(ge_i)$	$P(A ge_i)$
0	0	0	$(1-p)^2 \times (1-f)$	m
1	0	1	$(1-p)^2 \times f$	$m + e$
2	1	0	$2p(1-p) \times (1-f)$	$m + d$
3	1	1	$2p(1-p) \times f$	$m + d + e$
4	2	0	$p^2 \times (1-f)$	$m + 2d$
5	2	1	$p^2 \times f$	$m + 2d + e + e'$

Table 2.2: Additive model of gene-environment interaction \mathcal{M}_+ . For simulations we assume risk allele frequency p , exposure frequency f with independence between genotypes and exposure. The penetrance vector is parameterized with additive genetic (d) and environmental main effects (e) and an interaction term (e').

We construct the sampling distribution for naive MLEs of $\boldsymbol{\theta}$, conditional on a

significant GWAS result, using the following algorithm:

Algorithm 2:

1. For fixed values of θ and ϕ , simulate a data configuration D according to \mathcal{M}_+ .
2. Test D for association at genome-wide significance α .
If D is significant, go to step 3; otherwise discard D and return to step 1.
3. Assuming F and ϕ are known, compute the naive estimator $\hat{\theta}$.
4. Return to step 1.

By design, the algorithm computes MLEs only for datasets from $P(D \mid S_\alpha, \theta)$, the conditional likelihood for θ (Eq. 2.5). However, the naive MLEs are computed based on the unconditional likelihood $L(\theta \mid D)$ (Eq. 2.4). For our simulations, we generate data according to the additive penetrance model \mathcal{M}_+ (Table 2.2) and then computed naive MLEs for that model. That is, $\hat{\theta} = \{\hat{m}, \hat{e}, \hat{d}, \hat{e}'\}$. We use a stochastic grid search algorithm to perform the constrained maximization (Appendix 2.7.1) but other methods may also be used. For simulations with no interaction effect ($e' = 0$), we performed a follow-up hypothesis test for evidence of a gene-environment interaction. Formally, we tested the null hypothesis $e' = 0$ using a likelihood ratio test after re-computing naive MLEs with e' fixed at zero.

We consider two testing strategies for the initial GWAS (Algorithm 2, step 2). The first is a 2x2 Allelic Chi-Square Test that collapses the data D into risk allele counts for cases and controls, ignoring the environmental exposure, and testing strictly for a marginal allelic effect at the locus. The second testing strategy employs a logistic regression that includes main effects for risk allele carriers (β_1^G) and homozygotes (β_2^G), environmental exposure (β^E) and interaction parameters ($\beta_1^{GE}, \beta_2^{GE}$) for each genotype-exposure level. We use a Likelihood Ratio Test in the logistic regression framework to test for either significant main genetic or interaction effects ($H_0 : \beta_1^G = \beta_2^G = \beta_1^{GE} = \beta_2^{GE} = 0$) while controlling for the main effects of the exposure. Whereas the first testing strategy ignores the environmental exposure in the initial GWAS, the second strategy explicitly models and tests for potential interactions in the initial GWAS. Note that neither of these tests assume the true additive form of the penetrance vector is known. The testing strategies are described in more detail in the Appendix (2.7.2).

To compute sampling distributions for the corrected estimates $\bar{\theta}$ we use the same steps in Algorithm 2, except at step 3 we use the partial likelihood MCMC to compute

$\bar{\theta}$. However, to demonstrate the model-free aspect of the algorithm, we compute the corrected estimates $\bar{\theta}$ assuming the general six-parameter penetrance model \mathcal{M} (Table 2.1) .

2.5 Results

2.5.1 Sampling distributions for naive estimates $\hat{\theta}$

We simulated datasets for an additive penetrance model of gene-environment interaction (Table 2.2) and used Algorithm 2 to we generate the sampling distribution for parameter MLEs of the model, conditional on first showing significance in an initial GWAS. We considered three scenarios of gene-environment interaction: (1) true genetic and interaction effects ($d > 0, e' > 0$), (2) true interaction but no genetic effect ($d = 0, e' > 0$), and (3) true genetic but no interaction effect ($d > 0, e' = 0$). The initial association tests were performed at $\alpha = 10^{-8}$. Simulated datasets contained $N_a = 1000$ cases and $N_u = 1000$ controls. For each scenario, we show results for three sets of penetrance parameters that provide low, medium and high power for the two initial GWAS tests considered. We report sampling distributions for maximum likelihood parameter estimates of the additive penetrance model \mathcal{M}_+ and the estimated odds ratios for increase in risk associated with an additional copy of the risk allele.

2.5.1.1 True genetic and interaction effects ($d > 0, e' > 0$)

At high power, the parameter estimates for genetic effect (d) and interaction effect (e') are unbiased for both initial testing strategies. As power decreases to the medium level, the estimates of d and e' begin to show an upward bias for both tests (figs. 2.1 & 2.2). The extent of the bias increases when power is low. The baseline (m) and exposure (e) estimates are underestimated at medium and low power due to the prevalence constraint on the penetrance parameters. Note that for a fixed set of parameter values, the logistic regression provides a more powerful test resulting in less extreme bias. Although bias appears subtle for parameter values, it leads to more substantial bias in the odds ratios.

2.5.1.2 True interaction but no genetic effect ($d = 0, e' > 0$)

Next we consider a model with a true interaction effect but lacking a main genetic effect. Here the two tests differ with respect to bias on MLEs for the genetic and interaction parameters (figs. 2.3 & 2.4). Initial screening with both the chi-square test

and the logistic regression result in similar bias on point estimates of the interaction parameter e' . However, the chi-square test also induces a clear bias on the genetic term while the logistic regression remains unbiased for that parameter. This is the result of the logistic regression explicitly modeling both main genetic and interaction effects whereas the allelic chi-square test merges the signal of the two. As a result, odds ratios based on samples significant in the chi-square test suffer from a larger bias, especially at low power.

2.5.1.3 True genetic but no interaction effect ($d > 0, e' = 0$)

As in the previous scenario, the chi-square test induces a bias on both the genetic and interaction terms (fig. 2.5). The estimate of genetic effect is biased for the logistic regression but the interaction estimate is unbiased, again the result of explicitly separating the two effects in the initial model (fig. 2.6). For datasets showing a significant result in the initial genome scan, we performed a follow-up test for a gene-environment interaction. We formally tested the null hypothesis of no interaction ($H_0 : e' = 0$) using a likelihood ratio statistic and comparing it to the chi-square distribution with one degree of freedom. The QQ plots of observed versus expected p-values for the interaction test show that, at low power, the logistic regression induces only a slight departure from the expected null distribution (fig. 2.7). However, initial screening with the chi-square test results in a more extreme departure from the null. For the parameter values with the lowest power considered, the follow-up test with expected α level of 5% has actual Type I Error of 7% when screened with the logistic regression method and 14% for the chi-square test. Thus, the initial test used to identify the locus in a genome scan affects the α -level of future hypothesis testing on the same data.

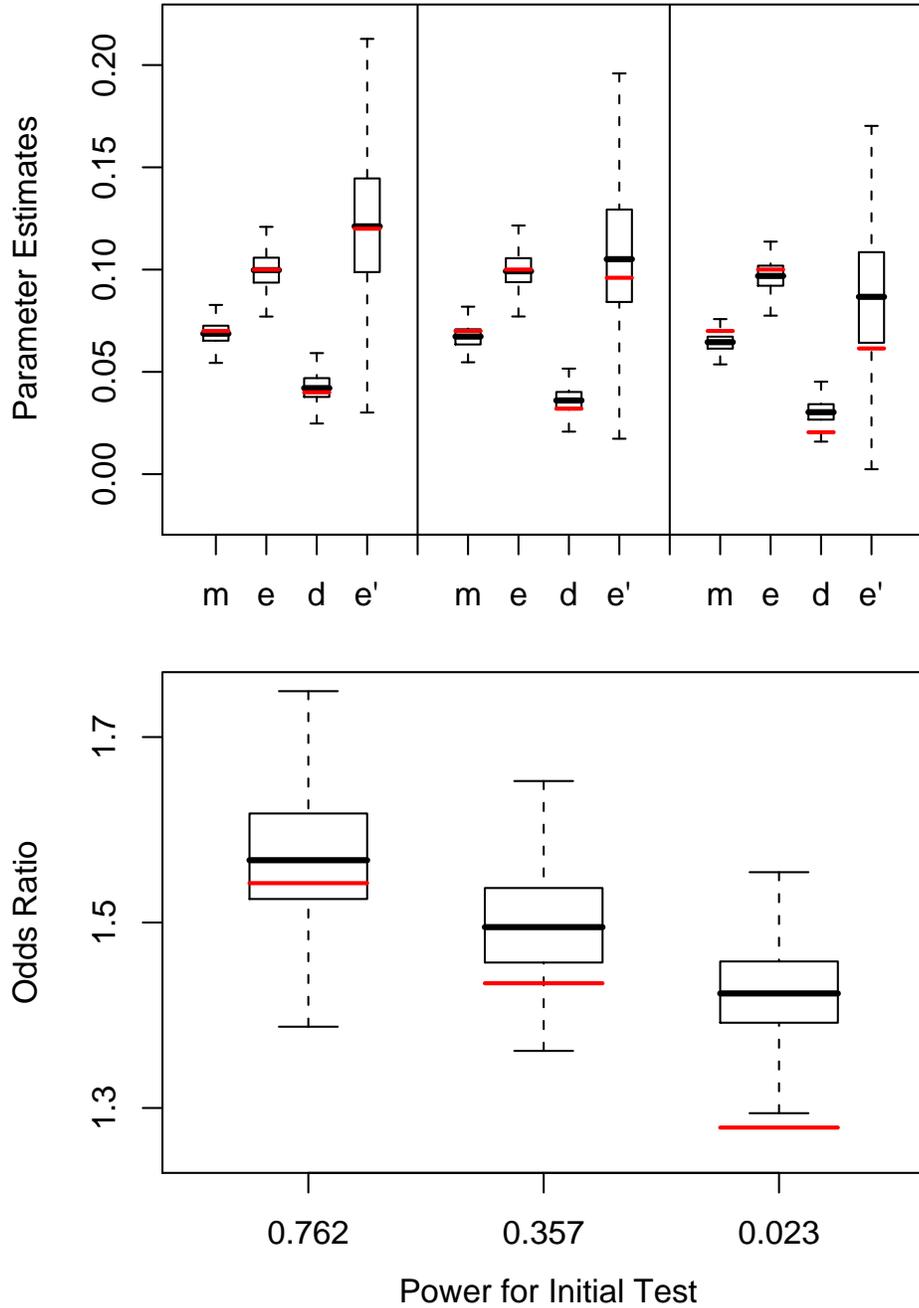


Figure 2.1: Naive MLEs conditional on genome-wide significance in the allelic chi-square test ($d > 0, e' > 0$). The top graph shows the distribution of parameter estimates at three power levels for the initial test. True parameter values are indicated by the red lines. At medium and low power, the estimates of d and e' have an upward bias. The bottom graph shows estimated odds ratios for each additional copy of the risk allele. Subtle biases in the individual parameters leads to a large bias in the odds ratios.

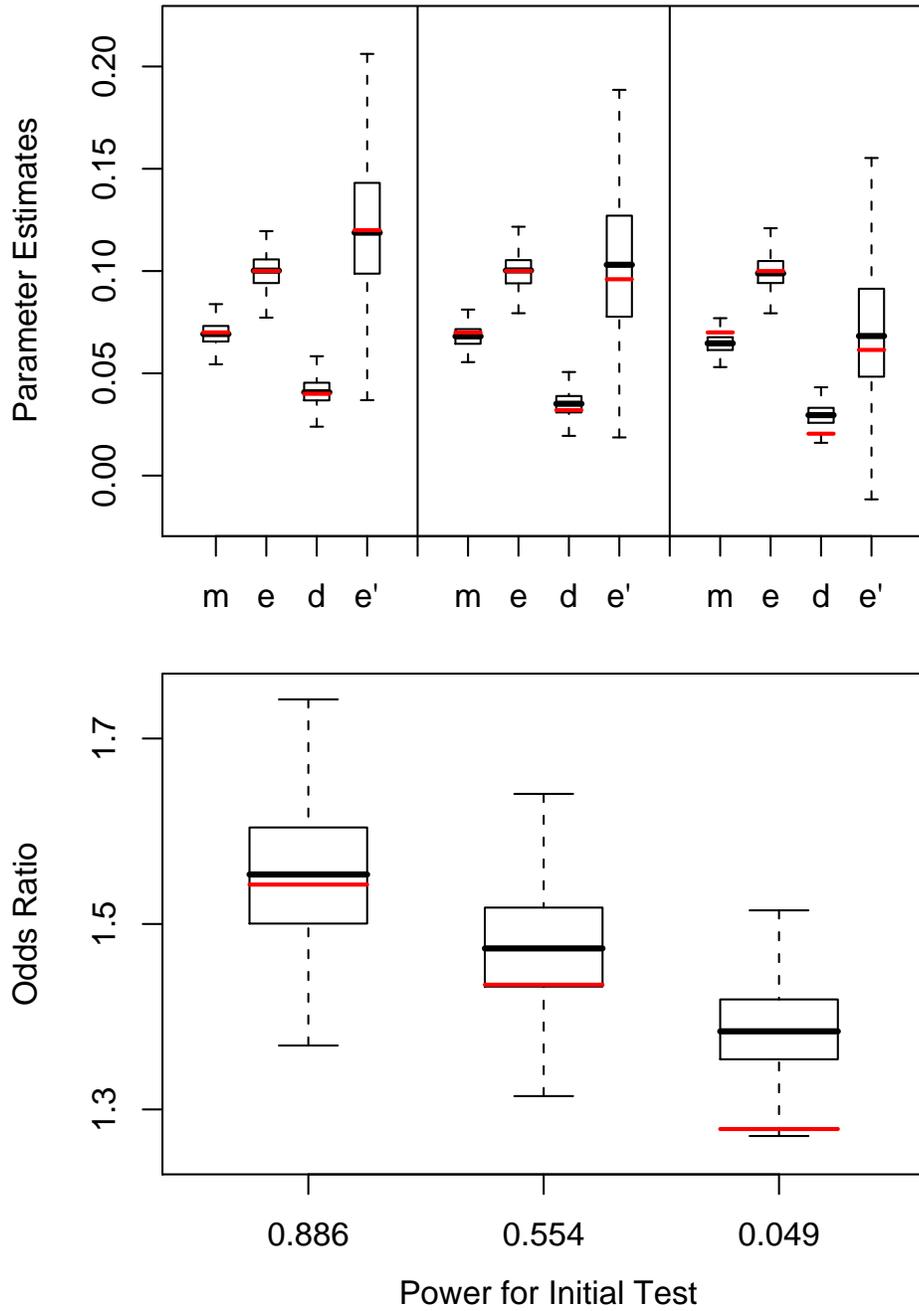


Figure 2.2: Naive MLEs conditional on genome-wide significance in the logistic regression ($d > 0, e' > 0$). As with the Allelic Chi-Square Test, the logistic regression framework results in bias for both d and e' . However, logistic regression is more powerful and leads to less extreme bias.

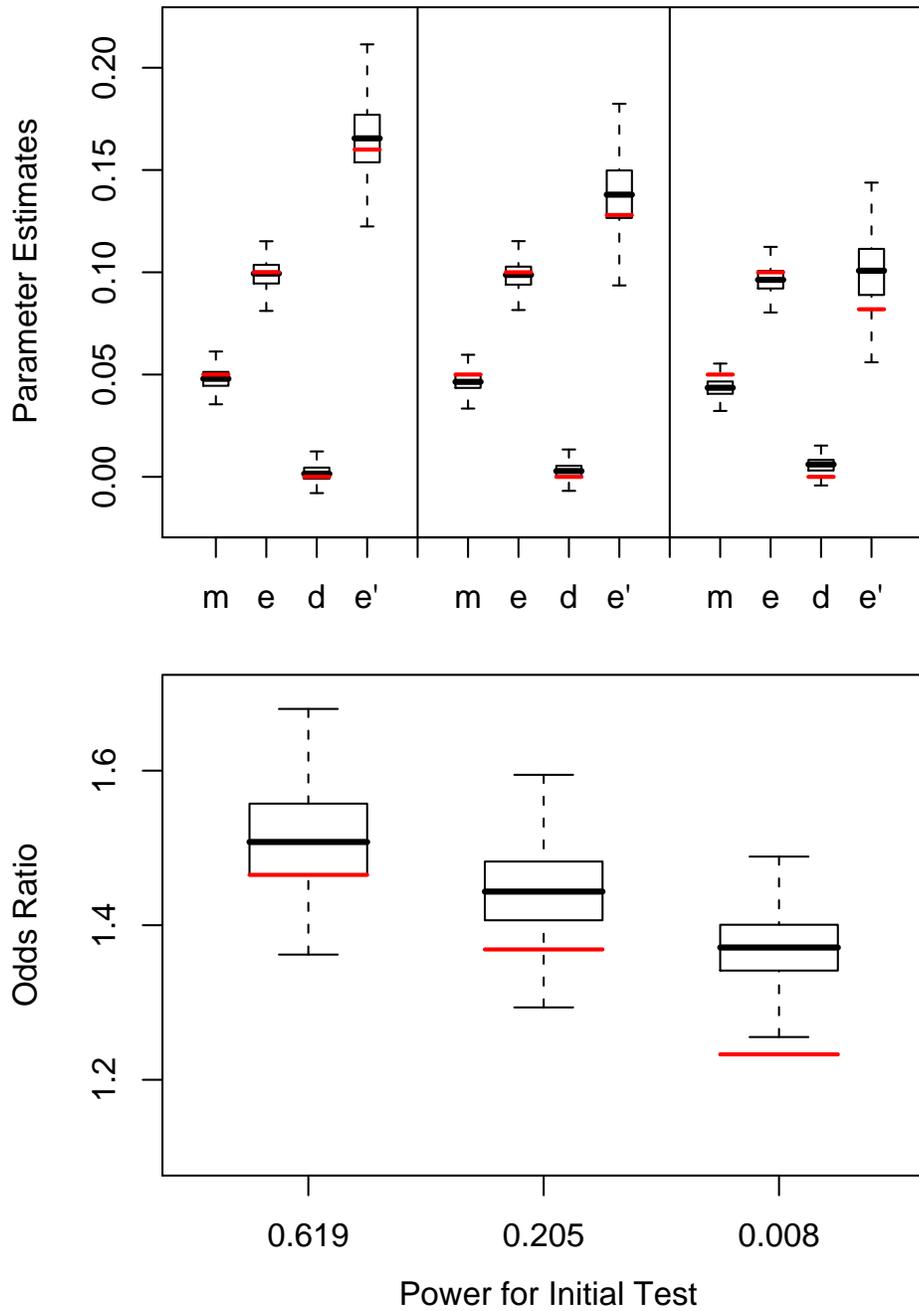


Figure 2.3: Naive MLEs conditional on genome-wide significance in the allelic chi-square test ($d = 0, e' > 0$). Initial screening with the chi-square test results in a bias on both d and e' .

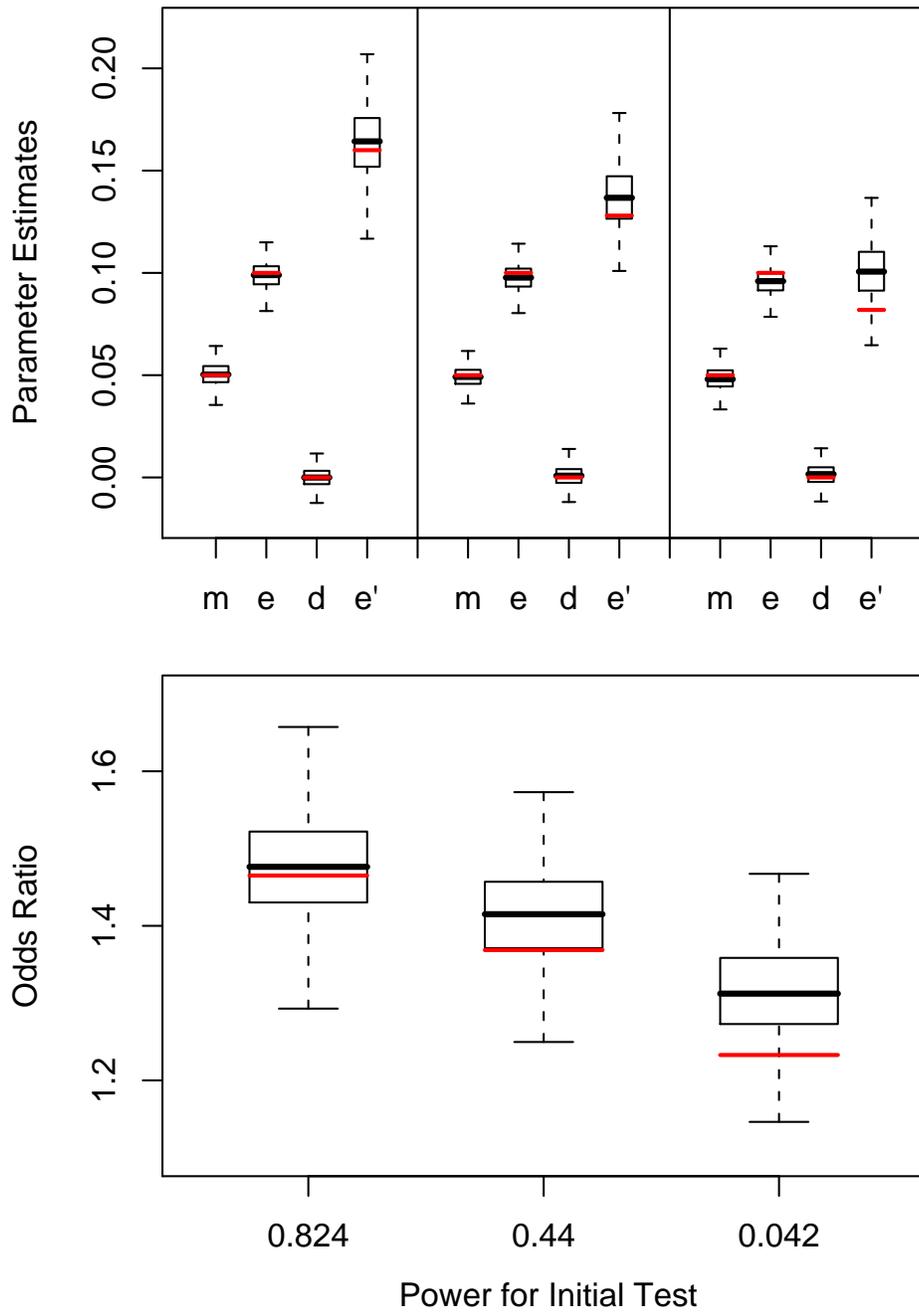


Figure 2.4: Naive MLEs conditional on genome-wide significance in the logistic regression ($d = 0, e' > 0$). Like the Allelic Chi-Square Test, the logistic regression framework leads results in biased estimates of the true signal source e' . Unlike the Allelic Chi-Square Test, estimates for d remain unbiased under the logistic regression.

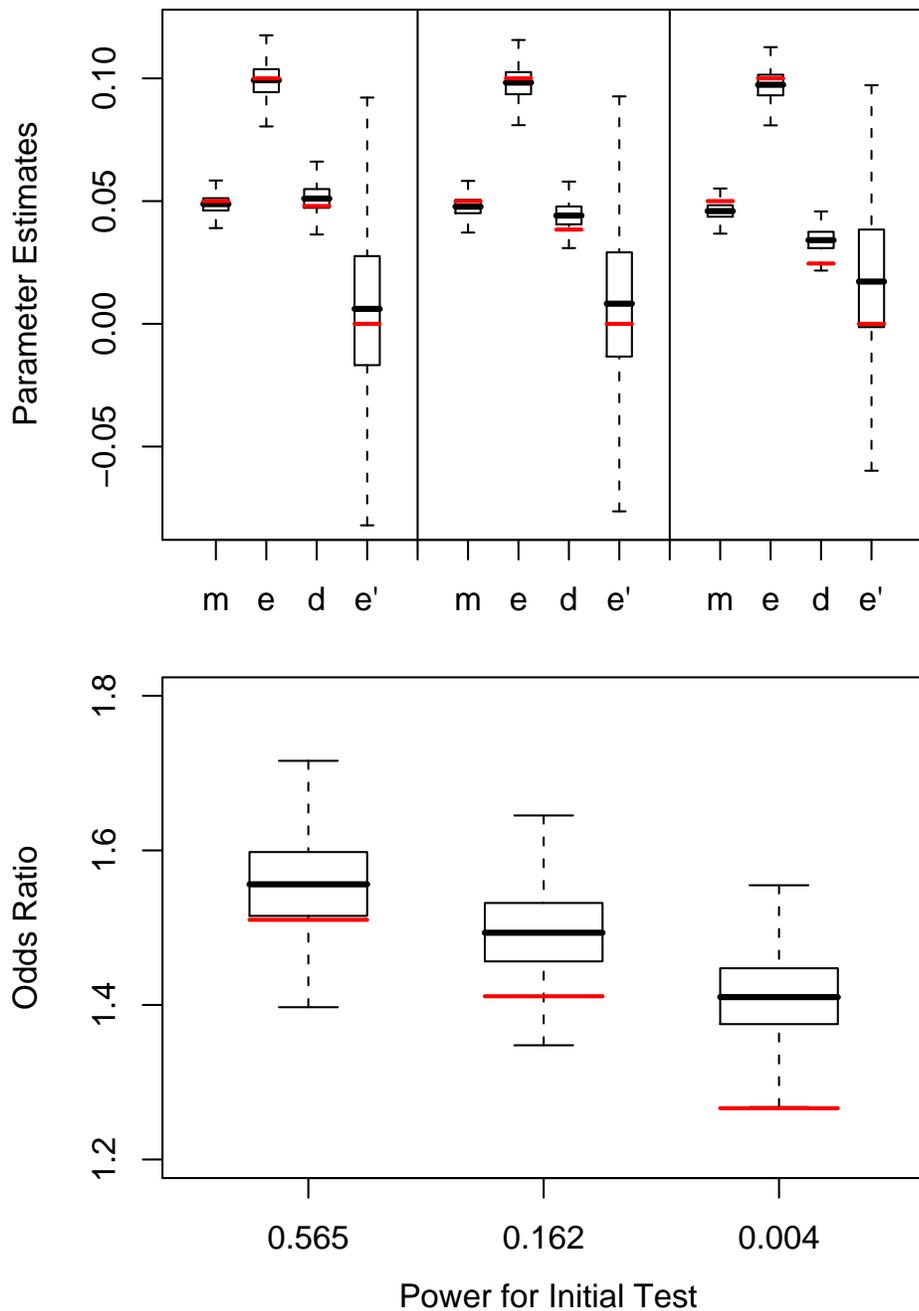


Figure 2.5: Naive MLEs conditional on genome-wide significance in the allelic chi-square test ($d > 0, e' = 0$). Again, the chi-square test induces a bias on parameter estimates for both d and e' . Combined with fig. 2.3, this implies that regardless of the true source of signal, the screening with the allelic chi-square test that ignores environmental exposure results in biased estimates for both genetic main effects and interaction.

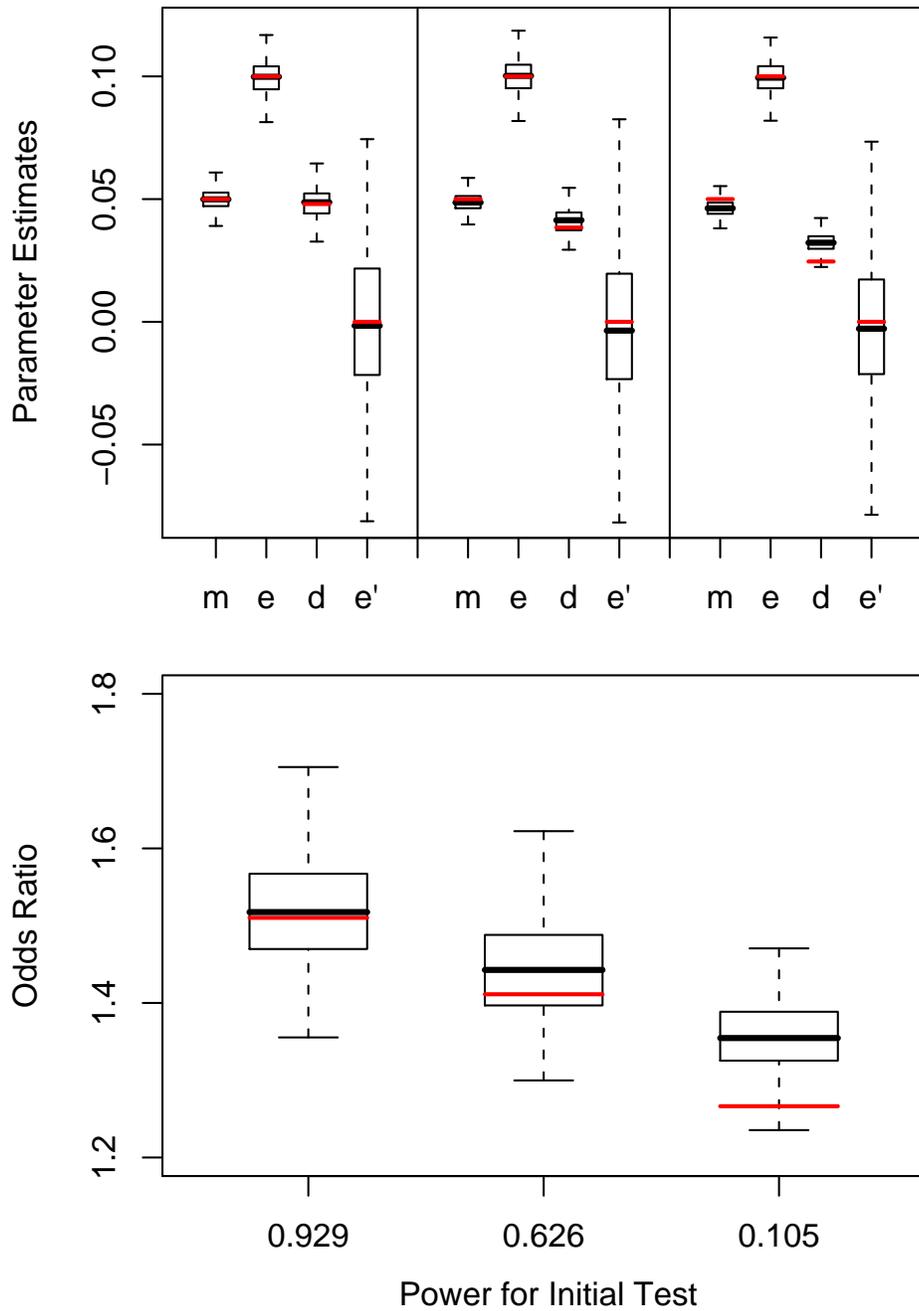


Figure 2.6: Naive MLEs conditional on genome-wide significance in the logistic regression ($d > 0, e' = 0$). Conditioning on the logistic regression result biased estimates of the true signal parameter d and unbiased estimates for e' . Thus, the logistic regression testing framework results in biased estimates of true effect parameters but unbiased estimates for null value parameters.

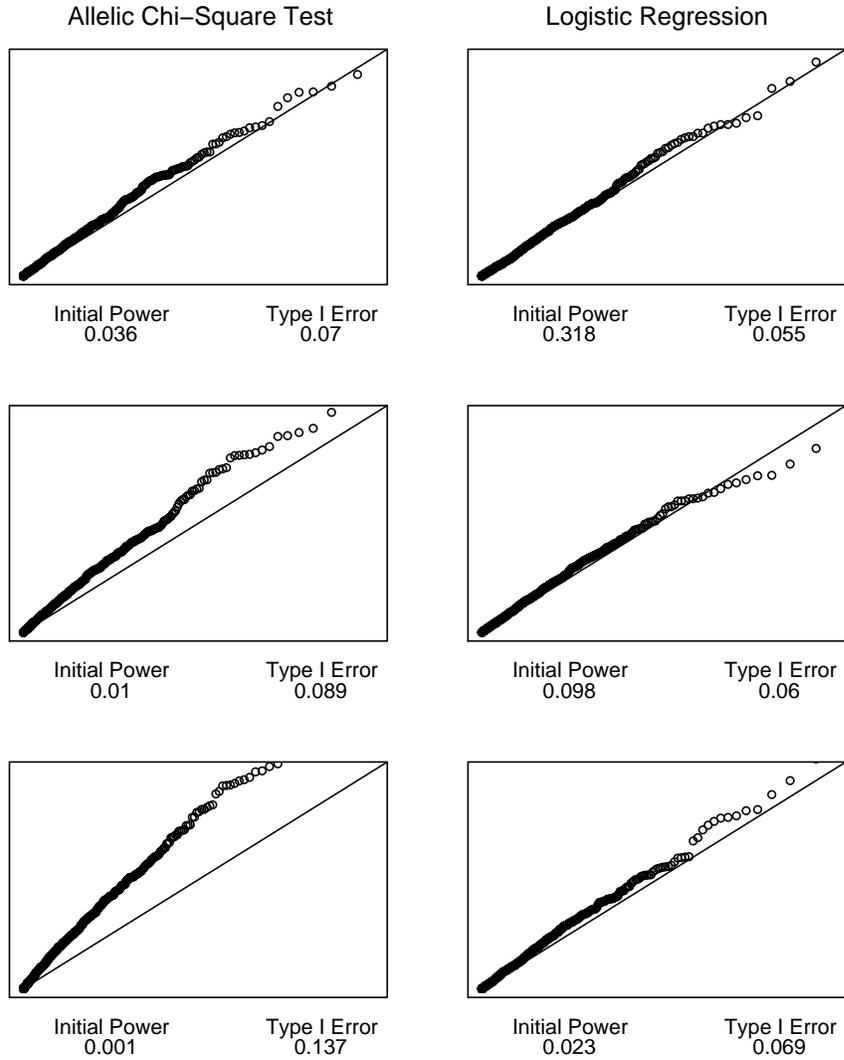


Figure 2.7: QQ plots for type I error in follow-up tests of gene-environment interaction. Initial GWAS testing with the allelic chi-square test (left column) results in an elevated Type I error for a follow-up test of interaction when no interaction exists. Type I error remains controlled when the logistic regression (right column) is used in the initial test.

2.5.2 Corrected estimates $\bar{\theta}$

In this section we present results for the partial likelihood MCMC algorithm and demonstrate that the corrected estimates $\bar{\theta}$ reduce bias on parameters for the gene-environment interaction model. Following Algorithm 2, simulated datasets according to the additive penetrance model of gene-environment interaction \mathcal{M}_+ (Table 2.2) and computed $\bar{\theta}$ for datasets significant in the initial GWAS test. To demonstrate the model-free aspect of the partial likelihood MCMC, we assumed the additive form of the gene-environment model was not known during estimation. That is, we used the partial likelihood MCMC to estimate penetrance parameters for the general six-parameter model \mathcal{M} (Table 2.1). The results presented in this section are based on the additive penetrance parameters $(m, e, d, e') = (0.12, 0.12, 0.02, 0.02)$, corresponding to $\theta = \{0.12, 0.24, 0.14, 0.26, 0.16, 0.30\}$. We used the logistic regression framework at $\alpha = 5 \times 10^{-8}$ for the initial GWAS test. Datasets contained $N_a = 10,000$ cases and $N_u = 10,000$ controls

It is difficult to directly show that the partial likelihood converges to the conditional likelihood function $L(\theta|D, S_\alpha)$ (Eq. 2.5). Instead we show that each component of the six-dimensional estimator $\bar{\theta}$ converges. Our results are based on running the chain for a total of ten million iterations, removing the first one million for burn-in and thinning the chain by keeping only every thousandth state of the chain. This resulted in a total of nine-thousand penetrance vectors in the quasi-independent random sample used to estimate $\bar{\theta}$.

Figure (2.8) shows the results for ten independent runs of the partial likelihood MCMC on the same dataset. Each plot in the figure shows the running estimate for one of the values $\bar{\theta}_i$ computed at each step in the random sample produced by the chain. Each of the colored lines in the plots indicates an independent run of the MCMC and shows that the estimate for $\bar{\theta}$ converges for each of the ten runs. Further, for each $\bar{\theta}_i$, the estimates from the ten independent runs converge to a common value, indicating consistent estimates of the parameters between runs. We note that the penetrance estimates for the first four genotype-exposure combinations converge faster than the estimates for $\bar{\theta}_4$ and $\bar{\theta}_5$ because of large case-control counts in those categories due to higher allele and exposure frequencies ($p = 0.1, f = 0.4$). The estimates for $\bar{\theta}_4$ and $\bar{\theta}_5$ show slight variability when the chain is stopped at 10 million iterations, however, when we allowed the chain to continue to 50 million total iterations the individual estimates continued to converge toward a common value. We found that the slight variability in estimates for $\bar{\theta}_4$ and $\bar{\theta}_5$ obtained at 10 million iterations of the chain did not affect the results in the subsequent analysis.

Next we present results comparing the corrected parameter estimates to naive estimates at a range of power values for the initial GWAS (Figure 2.9). For each power level, we computed naive and corrected estimates for 100 datasets statistically significant in the initial GWAS test and report the mean for each estimator. To obtain different power levels, we kept the penetrance parameters fixed and adjusted the allele and exposure frequencies. The corrected estimates are generated using the partial likelihood MCMC algorithm. Here, the naive estimates are computed using a standard MCMC algorithm to allow comparison.

When power is low (20%), each component of the naive estimator is biased. Penetrance for non-risk allele homozygotes (θ_0 and θ_1) are underestimated and penetrance for carries of the risk allele ($\theta_2 \dots \theta_5$) are overestimated, indicating an upward bias of genetic effect. Corrected estimates based on the partial likelihood MCMC reduce, but do not eliminate, the bias for each penetrance value. At moderate power (50%), the naive estimates show a bias for each parameter though slightly reduced from the low power setting. The corrected estimates again reduce this bias for each of the individual penetrance parameters and for several, the bias is nearly zero.

For the parameter setting with high power (80%), the ascertainment effect on the naive estimates is considerable reduced and they approach the true parameter values. The naive estimates that still show a noticeable bias at high power (θ_1, θ_3 and θ_4) are nearly perfectly estimated by the corrected estimates. However, the corrected estimates actually over-correct the remaining penetrance values.

The multidimensionality of the general six-parameter penetrance vector has the advantage that it removes assumptions on the underlying gene-environment model. However, the generality of the model also makes it difficult to fully appreciate the significance of bias on parameter estimates. We provide an interpretation for the bias by computing the estimated power for a replication study based on both the naive and corrected estimates derived in the previous result. That is, for each set of parameter estimates we computed the power for the logistic regression that was applied in the initial GWAS testing. Figure (2.10) shows the estimates of power for a replication analysis at the low, moderate and high power parameter settings. For comparison we also computed replication power based on true random samples, that is, datasets that were not required to show significance in the logistic regression and do not contain an ascertainment effect. The large spread on these estimates show the inherent variability in estimation for the six-parameter model and subsequent power calculations.

At the low power setting, the replication power for uncorrected estimates is sub-

stantially biased, with mean replication power nearly triple the true power value. Replication power based on the corrected estimates substantially reduces the bias although the mean is still nearly double the true value. At moderate power, uncorrected estimates again show a clear bias in estimates of replication power. Here, the corrected estimates provide a nearly identical result to the estimates derived without ascertainment. This corresponds to the result in Figure (2.9) that shows at moderate power (50%), the correction algorithm has eliminated nearly all ascertainment bias on individual parameter estimates. Finally, at the high power setting, the uncorrected estimates provide replication power estimates similar to the estimates with no ascertainment. The median of the distribution for replication power based on corrected estimates is the true power. However, the distribution for corrected estimates is more heavily skewed toward low estimates than either the no ascertainment or uncorrected estimates. This reflects the slight over-correction that was observed for some of the individual parameter estimates (Fig 2.9). In summary, the subtle bias in uncorrected estimates of penetrance parameters can lead to substantial overestimation of power for a replication study. The corrected estimates reduce this bias and produce estimates of replication power that more closely resemble estimates derived from true random samples.

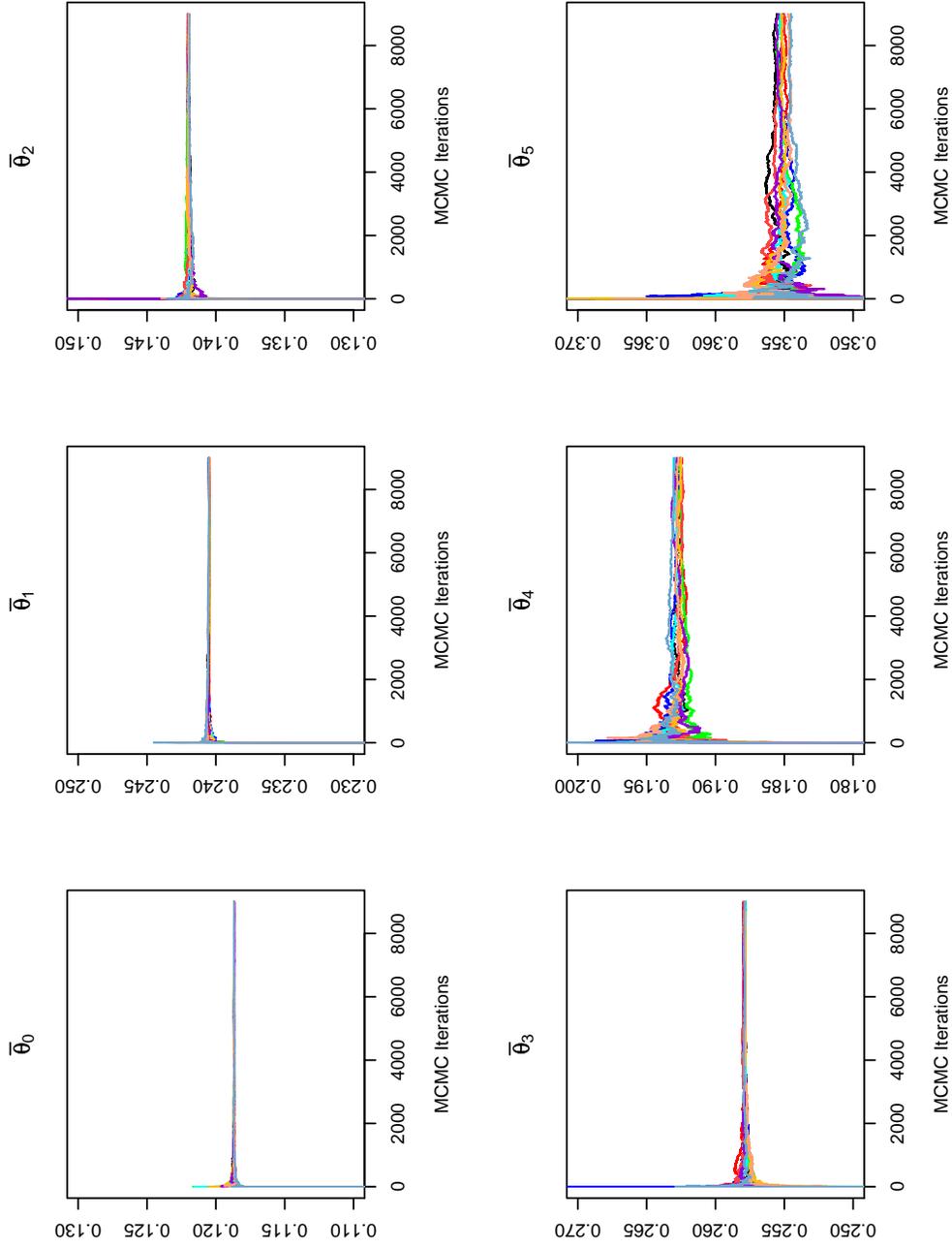


Figure 2.8: Convergence of $\bar{\theta}$ in the partial likelihood MCMC. Each plot shows the running estimate for a component $\bar{\theta}_i$ of the corrected parameter estimate $\bar{\theta}$. Each colored line in the plots indicates an independent estimates of $\bar{\theta}$ for the same dataset, generated by rerunning the partial likelihood MCMC. All six components of the estimate ($\bar{\theta}_i, 0 \leq i \leq 5$) converge in each of the ten runs. Further, the estimate from the independent realizations converge to the same estimate $\bar{\theta}$.

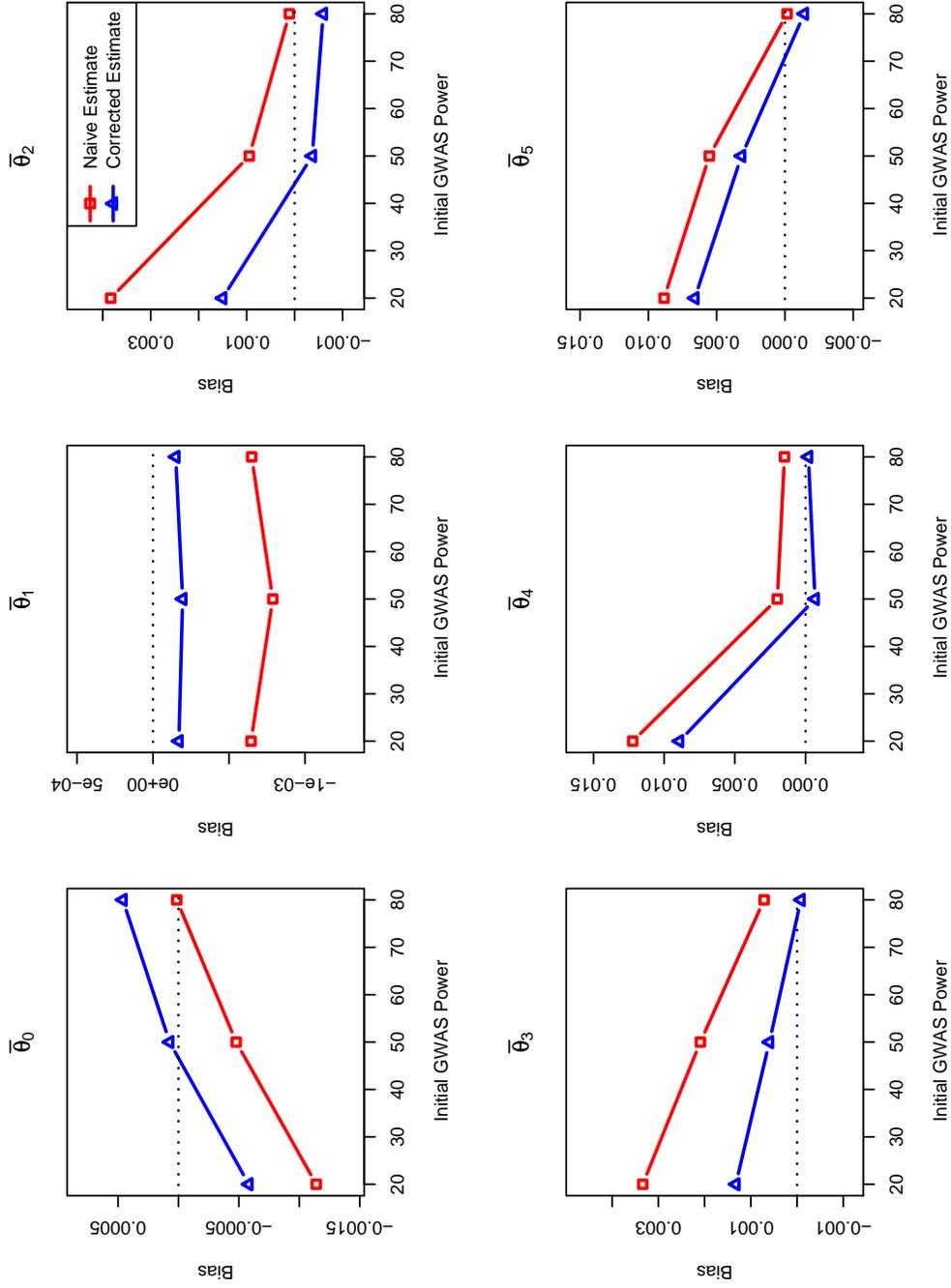


Figure 2.9: Comparison of corrected and naive estimates of θ . The corrected estimates $\hat{\theta}$ based on the partial likelihood MCMC (blue) reduce bias compared to Naive estimates $\hat{\theta}$ (red) at a range of power values for the initial GWAS. The horizontal dotted line in each plot indicates a bias of zero. There is a slight overcorrection on some of the parameters when power is high.

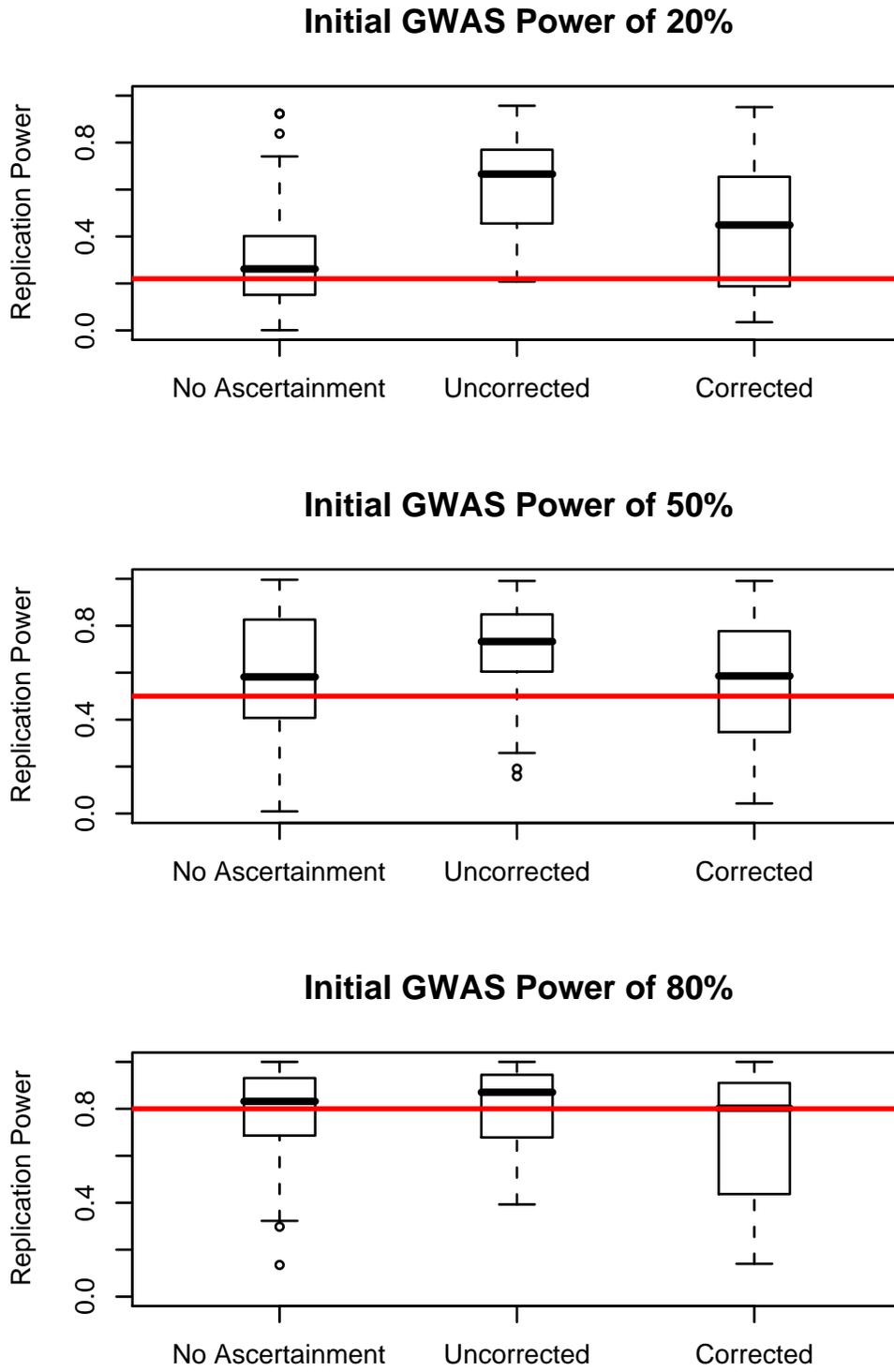


Figure 2.10: Power for a replication analysis based naive and corrected estimates.

2.6 Discussion

The detection of gene-environment interactions will improve our understanding of human complex disease, however, testing strategies for interactions require care to avoid potential bias in the analysis. In this chapter, we considered an analysis of gene-environment interaction as a follow-up to a significant GWAS result. We showed that naive parameter estimates of both the main genetic and interaction effects that ignore the ascertainment effect of requiring the significance result show an upward bias. This result is consistent with the standard Winner’s Curse phenomenon that is known to bias estimates of marginal genetic effects following a significant GWAS hits.

We considered two testing strategies for the initial GWAS, the first was a test for significant marginal genetic effects that ignored the environmental exposure, and the second testing for either significant main genetic effects or an interaction between genotype and exposure. We found that the follow-up analysis for interaction is affected differently by the two tests. Parameter estimates for true genetic or interaction effects have an upward bias for both tests, the classical Winner’s Curse. However, the test for marginal genetic effect also overestimates genetic and interaction effect when there is no actual effect for these parameters. As a result, the marginal effect test that did not initially model potential gene-environment interactions had a substantially inflated type I error in the follow-up hypothesis test for interaction. Alternatively, the test that did initially model an interaction maintained the appropriate type I error.

Corrected estimates can be obtained through a likelihood function that conditions on the significant GWAS result. The denominator of this likelihood is the power function for the GWAS test which, for most disease models and statistical tests, cannot be expressed analytically. Thus we introduced a new type of MCMC algorithm, a partial likelihood MCMC, to compute estimates from the conditional likelihood function. The partial likelihood MCMC accounts for this intractable denominator with a simulation-rejection step and makes use of the observed data through a standard Metropolis-Hastings step that contains the tractable portion of the likelihood. We applied the partial likelihood MCMC algorithm to data simulated under our additive penetrance model of gene-environment interaction. The estimates derived using the partial likelihood MCMC show a decrease in ascertainment bias compared to naive estimates and provide a more accurate estimates of power for follow-up testing.

In this chapter we considered a follow-up test for gene-environment interaction, but our results can be extended to other analyses that first screens for a significant

GWAS result. An obvious example is follow-up tests for gene-gene interactions. As with gene-environment interactions, initially testing all possible SNPs for interaction is not a powerful approach. Instead, an initial GWAS could identify SNPs with significant marginal effect and these SNPs could then further scrutinized for interactions with other SNPs. Our results indicate the potential for increased false positives in such an analysis if the initial GWAS test is not properly accounted for. An advantage of our partial likelihood MCMC algorithm is that it can be applied to any type of follow-up testing, including the gene-gene example describe here. The initial test, whatever the form, can be accounted for in the simulation-rejection step. And because we do not require explicit model assumptions the for penetrance parameters, the partial likelihood MCMC algorithm can be applied to any type of underlying model structure. This is particularly appealing since the true form of gene-environment and gene-gene interactions are often unknown.

In our application of the partial likelihood MCMC, we generated model-free estimates for the penetrance parameters of the six genotype-exposure categories. However, it may be of interest to test the plausibility of a specific interaction model while still controlling for the GWAS ascertainment effect. Due to the versatility of the algorithm, it is possible to incorporate hypothesis testing into the partial likelihood MCMC framework. Corrected parameter estimates for the “saturated” general model would first be computed, as we have done here. Then ascertainment-corrected estimates for parameters of a specific model of interest, for example our additive penetrance model would be computed. The fit of the more specific model can be assessed using a likelihood ratio test if the models are nested and Akaike information criterion (AIC) [1] or similar comparison of fit if the models are not nested. Of course, for both of these methods, the conditional likelihood function must be evaluated at the parameter estimates. Given that parameter estimates for the general and specific models, evaluation of the numerator is straightforward because it is simply the unconditional likelihood function. The denominator, statistical power for the parameter estimates in the initial GWAS test, can be estimated using Monte Carlo simulation. Then, for both the general and specific model, the corrected likelihood value can be computed as the ratio of the uncorrected likelihood and the power estimate. This testing framework allows interaction models that are not easily incorporated into a regression framework, the most common format for initial GWAS tests, to be properly analyzed.

The partial likelihood MCMC produces a random sample of penetrance vectors from the corrected likelihood. Without corresponding likelihood values for the vectors,

we based our corrected point estimates on a summary statistic from the sample. In our gene-environment model, each point in the sample was actually a six-dimensional penetrance vector, meaning that the choice for corrected estimates is not straightforward. We chose to use the marginal mean for each penetrance value as corrected estimates due to computational simplicity and because the six marginal means are guaranteed to satisfy the prevalence constraint. In actuality, the MLE for the corrected likelihood is the six-dimensional mode of the random sample but the multidimensionality and continuous scale of the space make this value difficult to compute. It is likely that the correction provided by the partial likelihood MCMC can be improved by choosing a corrected estimate that is guaranteed to be closer to the true mode of the corrected likelihood.

Gene-environment interactions are a promising model to explain complex human disease. However, when performed as a follow-up analysis to a GWAS experiment, care must be taken to avoid biased results. In this chapter, we have demonstrated the GWAS ascertainment effect on gene-environment analysis. To reduce bias, we introduced a flexible partial likelihood MCMC algorithm to compute parameter estimates from a likelihood function that conditions on the initial significant result.

2.7 Appendix

2.7.1 Stochastic grid search algorithm

We first describe a two-step, space-search algorithm used to compute MLEs of θ under the additive penetrance model (table 2.2) for the uncorrected likelihood function (eq. 2.4). We assume the true values of population prevalence F and frequencies ϕ are known. Fixing both F and ϕ places a constraint on possible values of θ (Eq. 2.6). In the additive penetrance model, given F and θ , $\Theta_{F,\phi} = \{\theta \in \mathbb{R}^4 \mid \sum_{i=0}^5 P(A|ge_i) P(ge_i) = F\}$ defines the subspace of penetrance parameter values that satisfy the prevalence constraint. The first step in the space-search algorithm is a coarse search over the domain $\Theta_{F,\phi}$ to identify the general region of the MLEs. To do this, we randomly chose vectors $\theta^* = (\theta_1^*, \dots, \theta_C^*) \in \Theta_{F,\phi}$ with uniform probability and computed the likelihood of the data D for each vector θ_i^* . Let $\hat{\theta}^* = \text{argmax}_{\theta^*} L(\theta \mid D)$ be the penetrance vector that maximizes the likelihood function in the coarse search.

Next we implemented a local random-walk to fine-search the region of $\hat{\theta}^*$ to determine precise MLEs. At the i^{th} step in the fine search, a new parameter vector θ_i is proposed by perturbing three of the parameters from $\theta_{i-1} = (m_{i-1}, d_{i-1}, e_{i-1}, e'_{i-1})$ by a value less than $\epsilon > 0$ and solving for the fourth value subject to the prevalence constraint. We accepted the proposed step only if it increased the likelihood of the data D. We reduced ϵ at each iteration. The parameter set $\hat{\theta} = (\hat{m}, \hat{d}, \hat{e}, \hat{e}')$ at the conclusion of the fine search was declared the naive MLE for θ given data D.

2.7.2 Initial genome-wide association testing strategies

We considered two testing strategies for the initial genome-level scan. First, we ignore the information on exposure status and test exclusively for a marginal allelic association using a standard allelic χ^2 test. Let $a_g = \sum_{g=1}^2 \sum_{e=0}^1 g \times a_{2g+e}$ be the number of putative risk alleles observed in cases and $u_g = \sum_{g=1}^2 \sum_{e=0}^1 g \times u_{2g+e}$ be the same quantity in controls. We compute the Pearson χ^2 statistic

$$\chi_{Pear}^2 = \frac{(N_a + N_u) \{a_g(2N_u - u_g) - u_g(2N_a - a_g)\}^2}{2(a_g + u_g)(2N_a + 2N_u - a_g - u_g)(N_a)(N_u)}$$

and determine significance by comparing to the theoretical χ^2 distribution with one degree of freedom.

For the second testing strategy we use logistic regression to simultaneously test

for either main genetic or interaction effects while controlling for exposure status. We fit the saturated logistic regression model

$$\text{logit}[P(A|g, e)] = \alpha + e\beta^E + 1_{\{g=1\}}\beta_1^G + 1_{\{g=2\}}\beta_2^G + e1_{\{g=1\}}\beta_1^{GE} + e1_{\{g=2\}}\beta_2^{GE}$$

This is the most general logistic model and contains main effect parameters for risk allele genotypes (β_1^G, β_2^G) and environmental exposure (β^E) plus interaction parameters ($\beta_1^{GE}, \beta_2^{GE}$) for each genotype-exposure level. We test the null hypothesis of no main genetic or interaction effects ($H_0 : \beta_1^G = \beta_2^G = \beta_1^{GE} = \beta_2^{GE} = 0$) by fitting the reduced model

$$\text{logit}(P(A|g, e)) = \alpha + e\beta^E$$

and comparing the fits using a Likelihood Ratio Test with 4 degrees of freedom.

CHAPTER III

Extending Rare Variant Testing Strategies: Analysis of Non-Coding Sequence and Imputed Genotypes

3.1 Introduction

The Genome-Wide Association Study (GWAS) is a powerful tool for analyzing common variation across the human genome[65]. In recent years, GWAS have identified risk alleles for a wide range of complex human diseases[25]. However, most of these alleles provide only small to moderate increases in risk and contribute little to the overall heritability of the disease[42]. Since it is unlikely that the remaining heritability can be completely explained by undetected common variants with even lower effects[20], heritable factors besides common variation must contribute to complex diseases. The Common Disease-Rare Variant Hypothesis proposes that some of the missing heritability can be explained by low frequency variants with larger effect sizes[58, 59]. Under this model, the contribution of individual variants to population prevalence is small but the combined effect of numerous rare variants can account for an appreciable fraction of the prevalence. This model is feasible if risk variants are subject to weak purifying selection and is supported by the fact that allele frequencies for protein-altering mutations are more heavily skewed toward rare variants than those for neutral variants[22].

Previously, technological limitations hampered the ability to affordably assay and test rare variants in large population-based samples. However, recent advances in next generation sequencing technology now provide the potential to detect all polymorphisms in a genomic region[47]. Thus rare variants can be tested directly rather than relying on indirect LD-based methods. Already, candidate region resequencing

has discovered numerous rare variants contributing to phenotypic variation and complex disease in humans. Resequencing of coding regions and consensus splice sites in *NPC1L1* and *PCSK9* has successfully identified multiple rare non-synonymous mutations collectively associated with variation in sterol absorption and plasma levels of LDL-C[11, 12].

Individually testing each variant identified by resequencing is not a powerful strategy since it requires stringent multiple testing correction and power diminishes with decreasing allele frequencies[32]. To avoid these issues, several statistical methods have been proposed that instead pool together multiple rare variants from the same gene and jointly test them for association[11, 34, 41, 56]. The recent literature has addressed two related question in rare variant testing: first, the question of how to effectively combine multiple rare variants in a gene into a single test and second, how to weight variants based on some prior assumption about the likelihood of functionality. Cohen et al. performed a pooled analysis of rare variants in *NPC1L1*, identifying non-synonymous variants observed only in cases or only in controls and comparing the distributions of cases and controls carrying these variants with a Fisher Exact Test[11]. Li and Leal proposed the Combined Multivariate and Collapsing method that pools variants below a specified minor allele frequency (maf) then dichotomizes individuals dependent on whether they carry a variant allele at one of the pooled sites[34]. A multivariate statistic is used to jointly analyze the set of pooled variants together with more common variants in the region.

Madsen and Browning introduced two features in the Weighted Sum Statistic (WSS)[41]. First, the WSS accumulates rare variant counts within the same gene for each individual rather than collapsing on them. Second, it introduced a weighting term to emphasize alleles with low minor allele frequency in controls. The result is a quantitative genetic score for each individual that is more informative than a qualitative score, especially for individuals harboring more than one rare allele in the region. The scores for all samples are ordered and the WSS is computed as the sum of ranks for cases. Significance is determined by permuting affection status and re-ranking. The ranking protects against outliers but becomes computationally expensive for large sample sizes.

Price et al. [56] showed that the power gain of weights based on minor allele frequency is dependent on the relation between risk allele frequency and likely effect size which in turn is dependent on selection strength. The weights used by Madsen and Browning, for example, correspond to strong purifying selection. If this model is correct, the WSS provides a significant power gain over the previous methods.

To generate a test that is powerful under multiple evolutionary models, Price et al proposed a variable maf threshold approach. For a given frequency threshold, a likelihood ratio statistic is computed to compare summed minor allele counts for variants below the maf threshold for cases and controls. The likelihood ratio statistic is maximized across a range of frequency thresholds to adapt the statistic to the underlying model of selection.

All pooling statistics are subject to variant misspecification, that is, potential inclusion of neutral variants or exclusion of risk variants. Study designs to date have opted to minimize inclusion of neutral variants by limiting analysis to non-synonymous coding variants of candidate genes[32]. The power of this strategy depends on the cumulative effect of rare risk variants that are exonic. While coding variants are most likely to be functional, they account for only a tiny fraction of variation in the genome. Numerous pieces of evidence indicate that non-coding variants play an extensive role in disease etiology. 88% of trait associated variants identified by GWAS have occurred outside of known coding regions[25]. Large portions of non-coding regions in the human genome are subject to negative selection indicating a functional purpose to the sequence[3]. In addition, non-coding risk variants have already been verified for numerous diseases[19, 23, 54]. Resequencing non-coding intronic and regulatory regions may enable detection of these more elusive risk variants but also presents new technical and analytical challenges to rare variant analysis. In particular, non-coding sequence contains substantially more neutral variation than coding regions.

Existing pooling methods have not been carefully assessed under a paradigm where many risk variants reside outside exons. Instead, these methods have only been considered for fairly optimal testing conditions assuming few variants per gene, most of which are causative[34, 41]. Moreover, previously published pooling methods assume high-quality rare variant genotypes that are only available through deep-coverage sequencing. Exon-only studies can attain high quality genotype calls because sequencing is limited to relatively small regions. Generating high quality sequence data of larger genomic regions (including whole-genome sequencing) is still expensive, limiting the number of samples that can be sequenced at deep-coverage for a given study. Instead, cost effective strategies such as low-coverage sequencing[18] and genotype imputation[37] will be used to produce sample sizes large enough to powerfully analyze rare variants. Genotype calls from these methods are less precise than deep sequencing, generating probabilistic rather than exact genotypes. Thus tests applied to whole-gene sequence data containing both coding and non-coding regions must

accept probabilistic genotypes and be robust to potentially high inclusion rates for neutral variants.

In this article, we consider a simple pooling statistic, the Cumulative Minor Allele Test (CMAT), and show that it is easily extended to accommodate practical analysis considerations such as qualitative covariates and probabilistic genotypes. The CMAT is closely related to the tests described in Madsen and Browning[41] and Price et al. [56] in that it aggregates allele counts rather than collapsing on them. Like these methods, the CMAT jointly analyzes sets of variants in the same gene that would otherwise be missed by a standard single marker analysis. Since power for single marker tests is dependent on study sample size and risk allele frequency, the CMAT is computed on variants with maf below a preset threshold. In this paper we especially focus on markers with maf $< 5\%$ and hereafter refer to these as rare variants.

The CMAT statistic is computed by summing rare allele counts for sites predicted to be functionally relevant separately for cases and controls. Our test statistic is analogous to the single marker allelic χ^2 statistic typically used to test for allele frequency difference between cases and controls. Significance is determined by permutation to account for correlation between pooled variants.

We compare the power of several pooling methods on case-control sequencing datasets simulated using population genetic models designed to mimic the overall level of diversity seen in European HapMap samples. We create a disease model of allelic heterogeneity by placing multiple rare risk variants in the population. The effect size for each risk variant is determined by allele frequency to ensure low power for a single marker test. Since our datasets contain realistic levels of neutral variation we can consider the effect of variant misspecification, both inclusion of neutral variants and exclusion of causal variants, in study designs ranging from exon-only to whole-gene analysis. We show that, dependent on the proportion of non-coding risk variants, whole gene designs may be more powerful than exon-only designs even if they include a large number of neutral variants.

The form of the CMAT statistic conveniently allows for categorical covariates and probabilistic genotypes. These extensions allow rare variant analysis for datasets containing imputed genotypes or low-coverage sequence data as well as common confounding variables such as population stratification. We demonstrate the importance of these extensions by analyzing two previously unconsidered rare variant study designs. First, we simulate rare variant datasets containing spurious associations created by population stratification. Ignoring the stratification leads to an elevated Type I Error rate while controlling for it with the covariate form of the CMAT maintains the

desired α -level.

Second, we present a study design consisting of both sequenced and imputed samples. We assume the sequenced samples are used to identify novel rare variants in a region of interest and then serve as templates to impute genotypes for these variants into the remaining (non-sequenced) samples. While carefully accounting for the uncertainty involved in imputing rare variants, we simulate datasets for this study design and analyze them with the CMAT. We show that using imputation to increase sample size of a sequencing dataset can substantially improve power. Hence we predict that imputation will provide a powerful cost-saving strategy for future resequencing studies. Moreover, our results suggest that existing resources such as the 1000 Genomes Project can be used to reanalyze existing GWAS datasets by imputing rare variants and performing tests such as the CMAT.

Finally, we illustrate the possibility of reanalyzing GWAS datasets without re-sequencing samples. As a proof of principle, we imputed over 8 million SNPs into the GAIN psoriasis GWAS dataset using CEU haplotypes from the 1000 Genomes Project. This dataset had previously been augmented with genotypes imputed from HapMap haplotypes and analyzed with a single marker association test[49]. That analysis identified numerous common risk loci that were subsequently replicated, including several variants in the *HLA* region on chromosome 6. We reanalyzed 3000 genes with at least 2 rare variants ($\text{maf} \leq 5\%$) using the CMAT. One gene maintained a significant test statistic after correcting for multiple testing, *SKIV2L*, located on chromosome 6 near the *HLA* region.

3.2 Methods

In the following, we develop notation for exact and probabilistic genotype calls then introduce the CMAT along with three alternative rare variant tests. Following that, we describe our algorithm to simulate case-control sequencing data based on population genetic models. Finally, we provide details for our application of the CMAT to the GAIN Psoriasis dataset.

3.2.1 Data structure

We assume a dataset of N_A cases and N_U controls. Let $x_{ij} \in \{0, 1, 2\}$ be the true number of minor alleles at the j^{th} variant site in the i^{th} case. Let y_{ij} be the same value for the i^{th} control. We consider two possible types of genotype calls in the data: exact calls, discrete values from $\{0, 1, 2\}$ giving the observed minor allele

count, and probabilistic calls consisting of a posterior probability mass function $P(\cdot)$ giving the likelihood for each possible minor allele count. Exact genotypes reflect high confidence calls possible in deep coverage sequencing data whereas the probabilistic calls represent the uncertainty in low coverage sequencing and imputation. In the dataset, we define the observed value for the j^{th} variant site in the i^{th} case to be

$$X_{ij} = \begin{cases} x_{ij}, & \text{for exact genotype calls} \\ \sum_{n=0}^2 nP(x_{ij} = n), & \text{for probabilistic genotype calls.} \end{cases}$$

That is, we assume the true minor allele count is observed if an exact call is made, otherwise we observe the expected minor allele count based on the posterior probability distribution. Similarly, we define Y_{ij} for the j^{th} variant site in the i^{th} control, replacing x_{ij} with y_{ij} .

3.2.2 Cumulative minor allele test

We assume the genetic data are partitioned into a collection of discrete testing units, genomic regions to be individually tested for association with disease susceptibility. The most natural choice for a testing unit is a single gene but highly conserved non-genic regions or pathways containing multiple genes are also suitable. Assume $F > 1$ variants in the testing unit, each with a weighting factor $w_j \geq 0$, ($j = 1, \dots, F$). A variant can be filtered out of the analysis by setting the respective weight to zero or its presence emphasized by assigning a large weight. For this paper, w_j is a simple indicator function to identify variants included in the analysis (described in more detail later). Note that a testing unit containing only a single variant with positive weight is equivalent to a single marker test on that variant.

We first describe application of the CMAT to a dataset containing exact genotype calls for all N_A cases and all N_U controls. Let $m_A = \sum_{i=1}^{N_A} \sum_{j=1}^F w_j X_{ij}$ and $m_U = \sum_{i=1}^{N_U} \sum_{j=1}^F w_j Y_{ij}$ be the weighted minor allele counts across all sites in the testing unit for cases and controls respectively. Then $M_A = \sum_{i=1}^{N_A} \sum_{j=1}^F w_j (2 - X_{ij})$ and $M_U = \sum_{i=1}^{N_U} \sum_{j=1}^F w_j (2 - Y_{ij})$ are therefore the weighted major allele counts across all sites for cases and controls respectively. We define the CMAT statistic Σ_{CMAT} to be

$$\Sigma_{\text{CMAT}} = \frac{N_A + N_U}{2N_A N_U \sum_j w_j} \times \frac{(m_A M_U - m_U M_A)^2}{(m_A + m_U)(M_A + M_U)}. \quad (3.1)$$

The statistic Σ_{CMAT} is derived from the standard Pearson χ^2 statistic for testing independence between allele frequency and disease status in a single marker association

test. However Σ_{CMAT} does not have an asymptotic χ^2 distribution as independent counts are required for the asymptotic properties to be valid. Since we sum over multiple sites in a testing unit, some of which may be in LD with each other, the counts are not independent. Instead, we determine statistical significance of Σ_{CMAT} by permuting affection status while holding the genetic data fixed. For each permuted realization, Σ_{CMAT} is recomputed and the p-value is defined as the proportion of permutations with a test statistic greater than or equal to the observed statistic.

In the presence of qualitative covariate data on potential confounders, the weighted allele counts are computed separately within each covariate level and the form of Σ_{CMAT} is changed to a Cochran-Mantel-Haenszel-like statistic. Assume a qualitative covariate $c = 1, \dots, C$. Using similar notation, we define the observed value for the j^{th} variant site in the i^{th} case of the c^{th} covariate class to be

$$X_{ijc} = \begin{cases} x_{ijc}, & \text{for exact genotype calls} \\ \sum_{n=0}^2 nP(x_{ijc} = n), & \text{for probabilistic genotype calls.} \end{cases}$$

Similarly, we define Y_{ijc} for the j^{th} variant site in the i^{th} control of the c^{th} covariate class, replacing x_{ijc} with y_{ijc} . Assume $N_{A,c}$ cases and $N_{U,c}$ controls within the c^{th} covariate class and $N_c = N_{A,c} + N_{U,c}$. Weighted allele counts are then computed within each covariate class separately. Let $m_{A,c} = \sum_{i=1}^{N_{A,c}} \sum_{j=1}^F w_j X_{ijc}$ and $m_{U,c} = \sum_{i=1}^{N_{U,c}} \sum_{j=1}^F w_j Y_{ijc}$ be the weighted minor allele counts across all sites in the testing unit for cases and controls in the c^{th} covariate class, respectively. Then $M_{A,c} = \sum_{i=1}^{N_{A,c}} \sum_{j=1}^F w_j (2 - X_{ijc})$ and $M_{U,c} = \sum_{i=1}^{N_{U,c}} \sum_{j=1}^F w_j (2 - Y_{ijc})$ are therefore the weighted major allele counts across sites for cases and controls of the c^{th} covariate class respectively. We define the covCMAT statistic Σ_{covCMAT} to be

$$\Sigma_{\text{covCMAT}} = \frac{\left[\sum_c m_{A,c} - \frac{N_{A,c}(m_{A,c} + m_{U,c})}{N_c} \right]^2}{\sum_c \frac{N_{A,c}N_{U,c}(m_{A,c} + m_{U,c})(M_{A,c} + M_{U,c})}{2N_c^3 \sum_j w_j}} \quad (3.2)$$

Statistical significance is determined by permuting case-control status while keeping the genetic and covariate data fixed. Eq. (3.2) resembles the Cochran-Mantel-Haenszel χ^2 statistic and simplifies to Eq. (3.1) when $C = 1$.

We now consider a dataset containing N_{seq}^A cases and N_{seq}^U controls with exact genotype calls and $N_A - N_{\text{seq}}^A$ cases and $N_U - N_{\text{seq}}^U$ controls with probabilistic calls. Computation of Σ_{CMAT} (Eq. 3.1) remains the same except expected minor allele counts replace exact counts for imputed samples. Significance is again determined

by permuting affection status. However, to account for the difference in quality between the two data types, affection status must be shuffled separately for exact and probabilistic calls. That is, for all permutations, the number of cases and controls with exact genotype counts must remain constant. Failure to modify the permutation method in this manner can affect Type I error, especially for unbalanced designs ($N_{seq}^A \neq N_{seq}^U$).

3.2.3 Alternative rare variant methods

We compared the performance of the CMAT to three alternative rare variant methods. First, we implemented the collapsing method described in Li and Leal[34] comparing the distribution of cases carrying at least one rare variant to that of controls carrying at least one rare variant. Let the indicator variable X_i denote whether the i^{th} case carries at least one rare variant at a site of interest as follows

$$X_i = \begin{cases} 1, & w_j X_{ij} > 0 \text{ for any } 1 \leq j \leq F \\ 0, & \text{otherwise.} \end{cases}$$

Y_i is analogously defined to indicate controls carrying at least one rare variant. Then $X = \sum_{i=1}^{N_A} X_i$ and $Y = \sum_{i=1}^{N_U} Y_i$ are, respectively, the number of cases and controls carrying at least one rare variant. The Pearson χ^2 statistic,

$$\chi^2_{COLL} = \frac{(N_A + N_U) \times (X N_U - Y N_A)^2}{N_A N_U (X + Y) (N_A + N_U - X - Y)}$$

tests the null hypothesis that cases and controls are equally likely to be carriers of a rare variant. χ^2_{COLL} has an asymptotic χ^2 distribution with one degree of freedom.

Next, we considered a private allele test similar to the method used by Cohen et al[11], comparing the number of rare variants unique to either cases or controls. For this test we require an equal number of cases and controls ($N_A = N_U$). A site is defined to be private if it is polymorphic in either cases or controls but monomorphic in the other group. The minor allele at a private site is called a private allele. For example, the minor allele at the j^{th} site is private to cases if $\sum_{i=1}^{N_A} X_{ij} > 0$ but $\sum_{i=1}^{N_U} Y_{ij} = 0$. Under the null hypothesis, rare variants are not associated with disease risk and private alleles are therefore equally likely to occur in cases and controls. This is tested formally using a χ^2 test in the following manner: Let n_{priv} be the total number of private alleles in the dataset and n_A and n_U the number of private alleles unique

to cases and controls respectively ($n_{priv} = n_A + n_U$). Define

$$\chi^2_{PRIV} = \frac{(n_A - \frac{n_{priv}}{2})^2 + (n_U - \frac{n_{priv}}{2})^2}{\frac{n_{priv}}{2}}$$

Under the null distribution of no association, χ^2_{PRIV} is asymptotically χ^2 distributed with one degree of freedom. As with the CMAT and collapsing test, the private allele test considers only variants with positive weighting terms.

Finally, we implemented the Weighted Sum Statistic as described by Madsen and Browning[41]. For the i^{th} individual in the dataset, we compute a genetic score defined as $\gamma_i = \sum_{j=1}^F w_j X_{ij}$. The genetic scores for all samples in the dataset (cases and controls combined) are sorted and the sum of ranks of genetic scores for cases, $x = \sum_{i \in cases} \text{rank}(\gamma_i)$, is computed. Statistical significance of x is determined by permutation. Madsen and Browning recommend upweighting rare variants by defining weighting terms according to minor allele frequency in controls. We do not directly consider the question of how to weight rare variants in this paper. Therefore, we apply a simple uniform weighting scheme to all tests. However, for comparative purposes, we include application of the WSS and CMAT using the weights defined in Madsen and Browning (Supplementary Figure 3.7). The three alternative methods have been formally defined only for exact genotypes, thus we limit power comparisons to datasets containing only exact genotype calls.

3.2.4 Simulations

3.2.4.1 Deep sequence datasets

We simulated deep sequence datasets containing exact genotype calls for an equal number N of cases and controls. We first created a population of ten-thousand $\sim 100\text{kb}$ haplotypes using the coalescent simulator *cosi* with parameters chosen to reflect characteristics seen in the European HapMap samples[67]. Let n_{tot} be the total number of polymorphic sites among the ten-thousand population haplotypes. Denote the allele at the j^{th} site on the i^{th} haplotype as A_{ij} where $A_{ij} = 0$ if the major allele is present and $A_{ij} = 1$ if the minor allele is present ($i = 1, \dots, 10,000$ and $j = 1, \dots, n_{tot}$). We fixed a maximum allele frequency p_{max} for risk alleles and randomly chose k sites with $\text{maf} < p_{max}$ to be causative. Let $c_j = 1$ if the j^{th} variant site is selected to be causative and $c_j = 0$ if it is neutral.

For each risk variant, we assigned an effect size that ensured a single marker association test would have low probability of being statistically significant. Specifically,

we computed the relative risk γ_p necessary for a risk variant with maf= p to have 10% power in a one degree of freedom χ^2 test of 1000 cases and 1000 controls performed at $\alpha = 10^{-5}$ (Figure 1). As a result, rarer variants are assigned larger relative risks, although we capped relative risks for variants with maf < 10^{-3} at 6. Assuming the maf for the j^{th} variant is p , we set the relative risk at that site $RR_j = \gamma_p^{c_j}$.

Assuming a multiplicative effect between causative variants, the penetrance ϕ_i for haplotype i is,

$$\phi_i = \sqrt{b} \times \prod_{j|A_{ij}=1} RR_j$$

where b is the risk for an individual with wildtype (non-risk) alleles at all causative sites and is set to ensure the population prevalence remains fixed at a desired level.

Diploid cases and controls were then sampled by randomly drawing two haplotypes conditional on disease status using Bayes' Theorem. The probability that the i^{th} and j^{th} haplotypes are chosen for a case, for example, is

$$\Pr(h_i, h_j | case) = \frac{\Pr(case | h_i, h_j) \times \Pr(h_i) \times \Pr(h_j)}{\Pr(case)} = \frac{\phi_i \times \phi_j}{10,000^2 \times \Pr(case)}.$$

assuming that unconditionally each of the ten-thousand population haplotypes are equally likely to be selected. We treat the unconditional probability of being a case (population prevalence) as a fixed parameter in our simulations.

Following the construction of a dataset, we mimicked a bioinformatic annotation process to determine the set of variants predicted to be functional and therefore included in the analysis. Each observed variant was randomly labeled as either 'included' or 'excluded' from the analysis conditional on whether it was causative or neutral. Define p_c to be the probability that a causative variant is predicted to be functional and therefore included in the analysis. Likewise, p_n is the same probability for neutral variants. Then, letting I_j be an indicator for inclusion in the analysis, the j^{th} variant is included with probability

$$\Pr(I_j = 1) = \begin{cases} p_c, & c_j = 1 \\ p_n, & c_j = 0. \end{cases}$$

The values p_c and p_n were treated as parameters to simulate study designs with alternative inclusion thresholds. Using the functional annotations, we defined the

weighting terms used in our simulations

$$w_j = \begin{cases} I_j, & maf_j \leq \beta \\ 0, & maf_j > \beta. \end{cases} \quad (3.3)$$

This weighting scheme therefore acted as a filter to retain variants with both $maf \leq \beta$ and predicted to be functional in the annotation step.

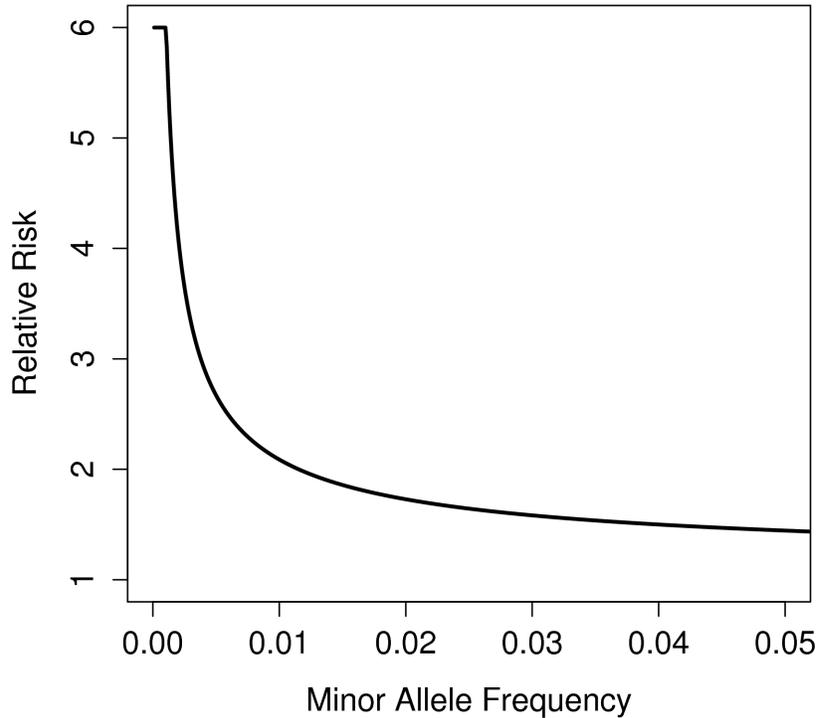


Figure 3.1: Relation between minor allele frequency and relative risk in our disease model. The relative risk is chosen to ensure that a single marker test of 1,000 cases and 1,000 controls performed at $\alpha = 10^{-5}$ on a risk variant with the specified maf has a maximum power of 10%. Relative Risks for variants with $maf < 10^{-3}$ were truncated to 6.

3.2.4.2 Imputation datasets

Next, we created datasets containing exact genotypes for N_{seq} cases and controls assumed sequenced at deep coverage and probabilistic genotypes for an additional $N-$

N_{seq} imputed cases and controls. Thus, in contrast to our deep sequence simulations where we assumed deep sequence data for all samples, here we assumed deep sequence data for only a fraction of the total sample size. It was computationally infeasible to phase and impute genotypes for each simulated dataset, therefore we drew haplotypes for N cases and controls using the previously described method and replaced the true minor allele counts with expected minor allele counts for the imputed portion of the sample. Expected minor allele counts were drawn from empirical sampling distributions created using independent imputation runs (see Appendix). Individual draws were made conditional on the true minor allele count at the locus to be imputed and the number of times the minor allele at that site was observed in the sequenced samples. We created separate empirical distributions for $N_{seq} = 100, 200$ and 400 . Only sites polymorphic among the sequenced samples were eligible for inclusion in the analysis. Singletons in the sequenced samples cannot be accurately phased and are therefore imputed. Hence, the minor allele must be observed at least twice in the sequenced samples to be imputed.

3.2.5 Stratified datasets

To demonstrate the covariate form of the CMAT statistic we simulated datasets containing population stratification. To do so, we used *cosi* to simulate sets of haplotypes that reflect variation observed in European and African populations[67]. We drew datasets containing N cases and N controls under the null hypothesis of no risk variants ($k = 0$), however we preferentially chose haplotypes from the African population to be cases. For each sample in a dataset, we first chose a population of origin for the sample. We let p be the probability that a control is derived from the African population and $p + \delta$, ($\delta > 0$), to be the probability that a case is derived from the African population. Controls and cases are therefore drawn from the European population with probability $1 - p$ and $1 - p - \delta$, respectively. Once population origin is determined, we randomly selected two haplotypes from the appropriate population to create a diploid sample. We analyzed each simulated dataset with both the CMAT and the covCMAT, controlling for population of origin in the latter. When applying the covCMAT, we assumed that the true population of origin was known for each sample.

3.2.6 Simulation settings

We fixed the population disease prevalence at 1% throughout the simulations. Under our disease model, increasing the number of causative sites k while holding prevalence constant increases the proportion of disease prevalence explained by variation at the locus. We focused our analysis on risk alleles with $\text{maf} \leq 5\%$ by setting parameters p_{max} and β to 0.05. However we repeated our analysis restricting risk variants to $\text{maf} \leq 1\%$ and report those results as well. However, since causative sites were chosen at random and the allele frequency spectrum was heavily shifted toward extremely low frequencies, approximately 95% of risk alleles in our simulations have frequency $< 1\%$ even for simulations with $p_{max} = 0.05$. We estimated power for each test at different parameter settings as the proportion of simulated datasets with statistically significant p-values (based on a minimum of at least 1000 simulated datasets). We report power at a critical level of $\alpha = 0.01$ assuming the sequenced region contains several genes to be tested.

3.2.7 GWAS application

We imputed 8.2 million autosomal SNPs into the GAIN Psoriasis dataset using 112 CEU haplotypes from the August 2009 release of the 1000 Genomes Project as reference. We filtered the imputed SNPs, removing all variants with very low estimated imputation accuracy ($\hat{R}^2 < 0.3$). We annotated SNPs discovered in the 1000 Genomes Pilot One Project using a custom Perl script. The tool reports for each SNP the gene locus (if available) and the predicted protein effect, based on a set of curated transcripts from Refseq and Genbank. We included SNPs annotated as missense, nonsense, splice-site or UTR in our analysis. We filtered out variants with $\text{maf} > 5\%$ and pooled the remaining variants together by genes. That is, we used the following weighting strategy

$$w_j = \begin{cases} 1, & \text{maf}_j \leq 0.05 \text{ and annotated as missense, nonsense, splice-site or UTR} \\ 0, & \text{otherwise.} \end{cases}$$

3.3 Results

3.3.1 Deep coverage sequencing datasets

To evaluate the performance of the CMAT, we used coalescent simulations to generate realistic case-control sequence data for a 100kb region of interest, representing

the exons, introns and surrounding regulatory regions for a large gene. A dataset of $N = 1000$ cases and controls drawn from a population with $k = 15$ rare ($\text{maf} \leq 5\%$) causative sites contained, on average, $S = 1565$ segregating variable sites with a mean pairwise sequence difference $\pi = 0.00114$. Of the observed sites, 1272 had $\text{maf} \leq 5\%$ and we observed 12.4 of the 15 risk alleles. A larger dataset with $N = 2000$ cases and controls contained an average of 1556 polymorphic sites with frequency $< 5\%$ and 14.1 of the 15 risk alleles. Larger sample sizes therefore increase both the number of risk alleles observed in the sample as well as the number of neutral variants.

We mimicked functional filtering by analyzing only a subset of the variants observed in a dataset. Conditional on being observed, causative variants were ‘predicted’ to be functional and therefore included in the analysis with probability p_c ; neutral variants were included with probability p_n . Since few of the observed variants are actually causative, p_n is approximately the overall proportion of rare variants included in the analysis while p_c can be thought of as the success rate for including causal variants.

We determined practical values for p_c and p_n by investigating the distribution of functional annotations for genic SNPs in the dbSNP database[71]. Of genic SNPs with at least one annotation, approximately 1.6% were denoted as non-synonymous coding or splicing variants (nonsense, missense, frameshift or altered splice-site), 1% were synonymous coding variants, 2.7% occurred in the UTR and 5.3% outside the transcribed region of the gene. Intronic SNPs accounted for the remaining class of variants. Thus, an overall inclusion rate (p_n) of 1 – 2% roughly corresponds to analyzing only non-synonymous variants whereas extending the analysis to include variants in the UTR and outside the transcribed region has an inclusion rate of approximately 10%.

We computed power for the rare variant methods on a misspecification grid with values of p_n between 0 and 0.1 and p_c between 0.2 and 1.0. First we computed Type I Error for each test by setting $k = 0$. The CMAT, collapsing method and WSS each maintained the desired false positive rate for all values of p_n . Type I error for the private allele test was initially conservative for smaller values of p_n , then increased with the number of variants included, becoming anti-conservative for larger values of p_n (Supplementary Figure 3.6). Increased false positives for the private allele test were likely due to including neutral variants in high pairwise r^2 , violating the independence assumption required for the asymptotic distribution.

In the presence of causative variants ($k > 0$), the power to identify a gene depended on the inclusion parameters p_c and p_n (Figure 2). We discuss results generated with a

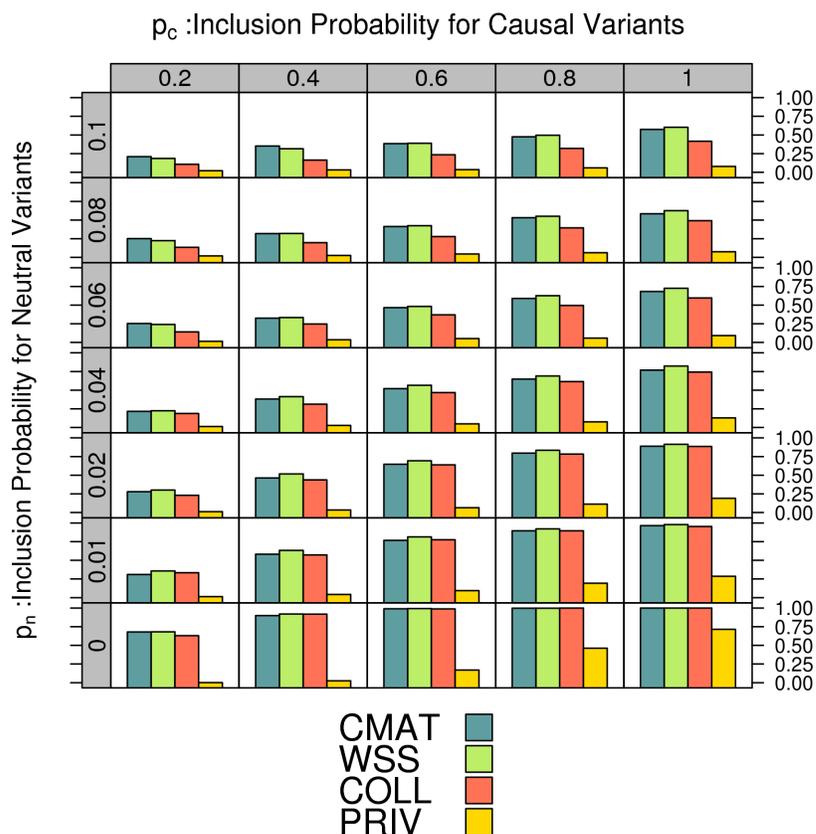


Figure 3.2: Power to analyze deep sequencing datasets for a range of inclusion probabilities. Each dataset contains exact genotypes for $N = 1000$ cases and controls based on $k = 15$ causative variants in the population. Along the vertical axis we vary the probability of (incorrectly) including a neutral variant (p_n) in the analysis and along the horizontal axis we vary the probability of (correctly) including a causative variant (p_c). The height of the bars in each cell indicates the power for the four tests at $\alpha = 0.01$.

sample size of $N = 1000$ and $k = 15$ causative variants; results for $k = 7$ and $k = 30$ were similar (not shown). When all variants were correctly specified ($p_c = 1, p_n = 0$), the CMAT, WSS and collapsing test attained power near 100% and the private test a power of 72%, indicating that each test is quite powerful under perfect filtering. However, power for each test dropped when we allowed for misspecification. Increasing the probability of including neutral variants ($p_n \uparrow$) reduced power. Decreasing the probability of including causative variants ($p_c \downarrow$) also lowered power.

A comparison of power between tests illustrates that the CMAT and WSS had nearly identical performance and were the most powerful tests at all levels of misspecification considered. The private allele test had power $< 20\%$ for most parameter settings. Power for the CMAT, WSS and collapsing test was nearly identical when only a small number of neutral variants were included in the test statistic ($p_n \leq 0.02$). Here, the absolute power for the three tests was heavily dependent on the inclusion rate for causal variants, increasing from 30% up to 95% as the number of causal variants included increased.

The CMAT and WSS showed a clear power gain over the collapsing method for larger neutral variant inclusion probabilities. In fact, the power gain was greatest when filtering accuracy was poorest. The CMAT had power of 24% compared to 11% for the collapsing test when $p_n = 0.1, p_c = 0.2$. This trend continued for values of $p_n > 0.1$ (data not shown). This difference is caused by the way the different tests account for individuals with more than one rare variant of interest. For larger values of p_n , individual samples are increasingly likely to contain multiple rare variants. By directly testing the number of rare variants rather than the number of rare variant carriers, the CMAT and WSS have additional power over the collapsing test.

Appropriately weighting variants in the test statistic may further improve power. However, it is presently unclear which is the most powerful weighting strategy and it is likely it will differ from case to case. Although we do not directly address the issue of most powerful weighting scheme in this paper, we computed power for both the CMAT and WSS using the weighting scheme described by Madsen and Browning[41]. Under this scheme, allele counts for the j^{th} variant are weighted by the inverse of the estimated standard deviation of minor allele frequency in controls. To facilitate comparison, we included only variants with maf below our pre-determined threshold ($\beta = 0.05$) in the analysis. The maf-based weights correspond more closely to our disease model (Figure 1) than do the simple uniform weights and therefore provided a more powerful analysis for both methods except when misspecification rates are highest (Supplementary Figure 3.7). Conditional on weighting scheme, the CMAT

and WSS had similar power across the grid.

To assess the influence of variants with maf of 1 – 5% on the presented results, we repeated all simulations restricting attention to variants with maf $\leq 1\%$ (ie $\beta = 1\%$, $p_{max} = 1\%$). The misspecification grid for these settings (Supplementary Figure 3.8) showed that overall power for each test was slightly lower than the presented results. The noticeable change was that for the largest values of p_n and p_c , the WSS showed a power advantage over the other tests with the CMAT and the collapsing test having similar power.

For the remainder, simulation results are based on inclusion parameters of $p_n = 0.1$, $p_c = 0.8$ to reflect a whole-gene analysis strategy that includes non-synonymous coding and splice-site mutations plus variants in the UTR and potential regulatory regions flanking the gene.

3.3.2 Covariate correction

Next, we created datasets in which samples were drawn from two distinct populations meant to resemble European and African haplotypes. We simulated the datasets under the null hypothesis of no association ($k = 0$) but preferentially drew cases from the African population. Since the African haplotypes contain more rare variation than do the European haplotypes, the datasets contain a spurious association between disease status and excess of rare variants. Datasets contained $N = 1000$ cases and controls drawn from the African population with probability $p + \delta$, $\delta > 0$ and p , respectively. We analyzed each dataset at $\alpha = 0.01$ with the CMAT and the covCMAT, controlling for population of origin in the latter.

We present results for $p = 0.5$ and $0 \leq \delta \leq 0.25$ (Figure 3). Ignoring the population stratification resulted in an elevated CMAT Type I Error, increasing sharply for $\delta > 0.025$. The magnitude of this increase is affected by the inclusion probability for the summary statistics. For strategies that attempt to capture all variants near a gene (shown here) the false positive rate is substantially larger than for strategies focusing on exonic variation. Controlling for ancestry by including it as a covariate into the covCMAT maintained the desired Type I error across all values of δ considered.

3.3.3 Imputation datasets

The CMAT is easily applied to imputation datasets containing probabilistic genotype calls. To consider the potential of a study design combining sequenced and imputed samples, we simulated exact genotype calls for the sequenced samples and

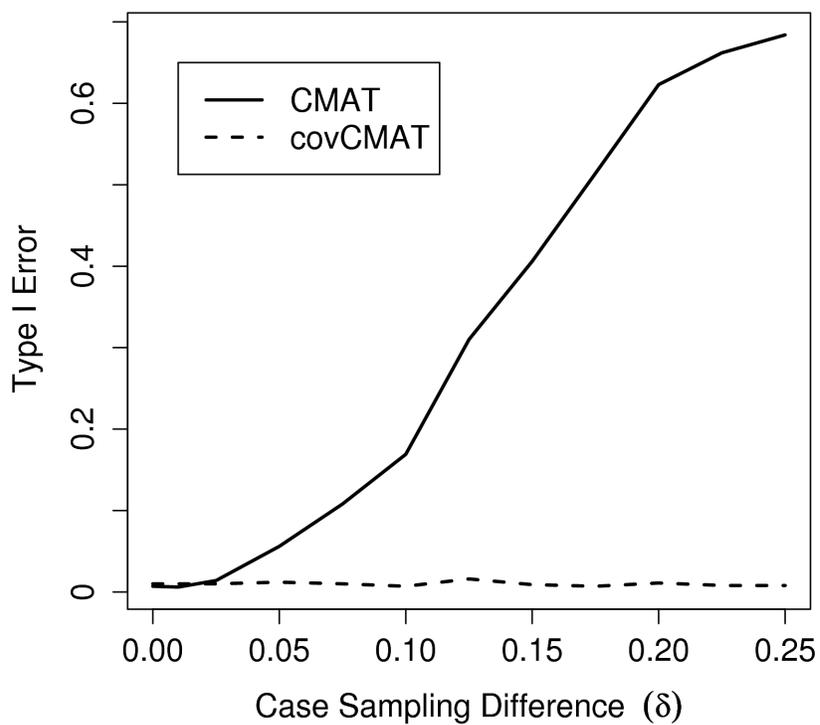


Figure 3.3: Application of the covCMAT to control for population stratification. Cases were preferentially sampled from a population containing a larger number of rare variants. Failure to account for population stratification leads to inflated false positive rates for the CMAT. When applied with the covariate correction, the covCMAT maintained the appropriate Type I Error.

probabilistic genotypes for the remaining samples. We considered a design with an equal number of cases and controls sequenced in a 100kb region of interest and genotyped for tagSNPs in a 1Mb encompassing region. Imputed samples are assumed genotyped for the same set of tagSNPs. In this design, variants observed at least twice in the sequenced samples are imputed in the non-sequenced samples.

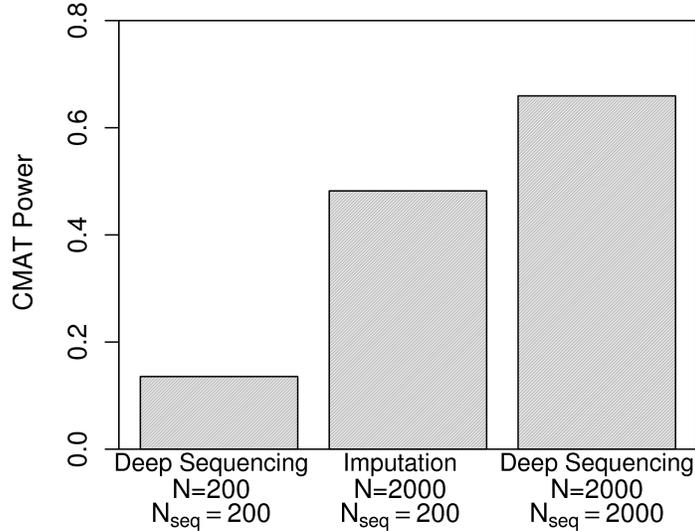


Figure 3.4: Comparison of CMAT power for deep sequencing and imputation study designs. From left to right, the bars show power at $\alpha = 0.01$ for a deep sequencing dataset with $N = 200$, an imputation dataset with $N_{seq} = 200, N = 2000$ and a deep sequencing dataset with $N = 2000$. For each, we used the whole-gene inclusion threshold ($p_n = 0.1, p_c = .8$).

We found that the addition of imputed samples to a fixed number of sequenced samples can provide a considerable power gain over analyzing only the sequenced samples (Figure 4). A whole-gene CMAT analysis of datasets drawn from a population containing $k = 15$ causative variants and constrained to $N = N_{seq} = 200$ sequenced cases and controls has power of 14%. By augmenting these sequenced samples with an additional 1800 imputed cases and controls (total sample size $N = 2000$), power improves to 48%. This compares favorably with the optimal $N = 2000$ design that sequences all samples and has power of 66%. Thus, the additional information from imputed samples recovered much but not all of the power of a fully sequenced dataset.

We extended our analysis to a wide range of sample sizes with N from 200 to

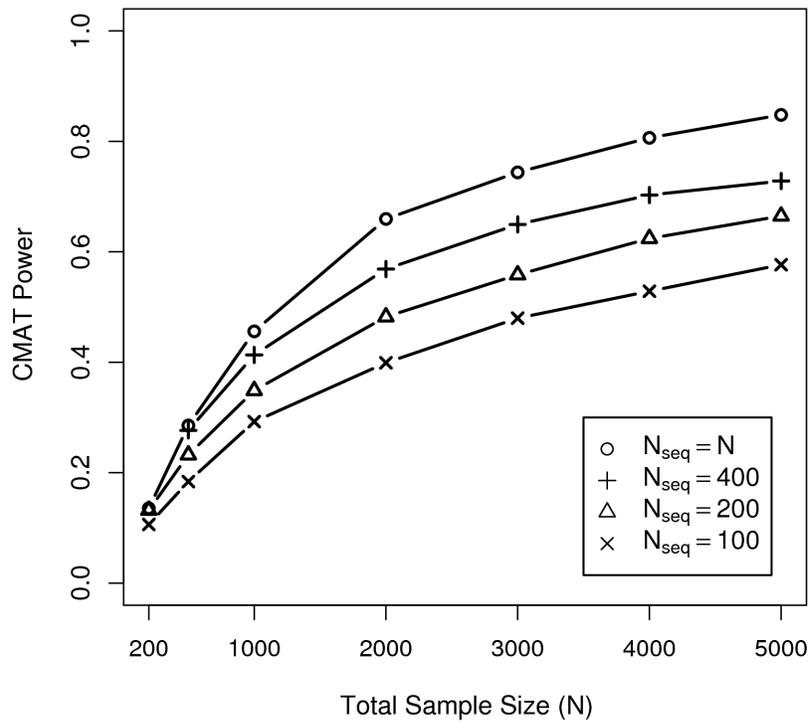


Figure 3.5: CMAT power for imputation datasets. Datasets contain exact genotypes for N_{seq} sequenced cases and controls and probabilistic genotypes based on imputation for the remaining samples. The top line shows CMAT power when all samples are sequenced ($N_{seq} = N$) and serves as an upper bound for power at a fixed total sample size N . We report power at $\alpha = 0.01$ using the whole-gene inclusion threshold ($p_n = 0.1, p_c = 0.8$).

5000 and considered the effect of sequencing $N_{seq} = 100, 200$ or 400 samples for each N . The CMAT had a well controlled type I error when applied to datasets simulated with $k = 0$ causative variants (data not shown). We present results based on $k = 15$ causative variants in the population analyzed using whole-gene inclusion parameters (Figure 5). Power curves for inclusion thresholds that reflect an exon-only analysis ($p_n = 0.02, p_c = 0.4$) were slightly lower across all values of N considered (data not shown). For comparison, we also computed CMAT power for a dataset containing exact genotypes for all samples (i.e. $N_{seq} = N$). We found that for a given total sample size N , CMAT power increased with the number of sequenced samples. At $N = 3000$, datasets containing 100, 200 and 400 sequenced samples had powers of 48%, 56%, and 65%, respectively. To attain similar power in a set of fully sequenced samples requires $N = N_{seq} = 1000, 1500$ and 2000 samples, respectively.

The dependence of power on the number of sequenced individuals is likely driven by three factors. First, replacing an exact genotype with a probabilistic genotype results in a loss of information. Thus, for a fixed sample size, datasets containing fewer sequenced samples suffer a larger information loss. Second, increasing the number of sequenced samples increases the chance that a risk allele is observed at least twice and can therefore be imputed. Of $k = 15$ risk alleles in the $\text{maf} < 5\%$ simulations, an average of 3.2 were observed at least twice among 100 sequenced cases and controls. This number increased to 5.0 and 7.5 for datasets with 200 and 400 sequenced cases and controls, respectively. Third, imputation accuracy for an individual allele improves as that allele is observed more often in the sequenced samples. Sequencing a larger number of samples increases the number of times a risk allele is observed, thus improving imputation accuracy for that allele.

We repeated the imputation simulations using 1% maf parameter settings (Supplementary Figure 3.9). We observed only a small reduction in power compared to the $\text{maf} \leq 5\%$ analysis. Only datasets with 100 sequenced cases and 100 sequenced controls showed a notable reduction in power. For $N_{seq} = 100, 200$ and 400 sequenced cases and controls, a study with total sample size of $N = 3000$ had powers of 38%, 52%, and 63%, respectively. Hence, provided a sufficiently large set of sequenced templates, imputation of rare variants is a useful strategy, even if variants with $\text{maf} < 1\%$ are of particular interest.

3.3.4 Application to GAIN psoriasis data

Our simulation study assumed that imputation templates were sequenced individuals from the study sample. It is feasible to instead use haplotypes from a public

dataset as the imputation templates. This has the advantage that it allows rare variant analysis in any existing GWAS dataset without requiring additional sequencing by the investigator.

As proof of principle for this approach, we applied the CMAT to the GAIN Psoriasis dataset consisting of 1,359 psoriasis cases and 1,400 unaffected controls of white European ancestry. We imputed 8.2 million autosomal SNPs into the dataset using 112 CEU haplotypes from the August 2009 release of 1000 Genomes Project as our reference panel. Previously, this dataset had been imputed for 2.5 million SNPs using the CEU HapMap samples and analyzed with a standard single marker test for association[49]. The strongest signal for association (*rs12191877*, single marker $p = 4 \times 10^{-53}$) was located 13kb upstream of the *HLA-C* gene, a previously known psoriasis locus on chromosome 6. Ten of the top 18 loci identified in the initial analysis were subsequently replicated in a larger, independent sample.

To apply the CMAT, we assigned the imputed SNPs to genes and retrieved functional annotations for genic variants (see Methods). We retained only SNPs with $\text{maf} < 0.05$ and annotated as either non-synonymous, splice-site or UTR. In total, 2889 genes containing two or more SNPs following filtering were analyzed with the CMAT. Of the genes tested, 55% contained two SNPs, 23% three SNPs, 11% four SNPs and the remaining 11% contained 5 or more SNPs. None of the ten replicated SNPs from the original analysis remained following filtering and only three genes near a replicated signal (*IL12B*, *TSC1* and *TNFAIP3*) were included in the CMAT analysis.

SKIV2L (CMAT $p < 10^{-6}$)

Variant	Imputation		Single Marker p-value	Correlation Between Imputed Genotypes	
	maf	function		\hat{R}^2	Imputed Genotypes
rs17201466	0.0496	UTR	0.98	0.0018	1.000
rs36038685	0.0109	R324W	0.99	0.1210	-0.016 1.000
rs3911893	0.0427	D887N	0.94	0.0029	-0.055 -0.006 1.000
rs106287	0.0359	V917M	0.91	0.0888	-0.059 -0.027 -0.034 1.000

Table 3.1: Summary of the *SKIV2L* testing unit from CMAT analysis of the GAIN psoriasis dataset. We imputed 8.2 million SNPs into the GAIN psoriasis dataset using 112 CEU haplotypes from the 1000 Genomes Project. *SKIV2L* was statistically significant following Bonferroni correction and is located on 6p21.33, 700kb away from *HLA - C*, a known psoriasis susceptibility locus. The table lists the maf, functional annotation, imputation accuracy \hat{R}^2 , and single marker p-value of individual variants included in the pooled statistic for each gene. The last columns contain the pairwise correlations between imputed minor allele counts.

After Bonferroni correction for the number of genes tested, one gene achieved statistical significance: *SKIV2L*, ($p < 10^{-6}$; $p < 3 \times 10^{-3}$ after Bonferroni correction) [MIM 600478]. *SKIV2L* is located on 6p21.33, 700kb away from *HLA-C*, the previously implicated psoriasis susceptibility locus. The *SKIV2L* testing unit contained four imputed variants with $\text{maf} < 0.05$ (Table 1). Although each variant trended toward significance in the single marker test, no individual p-value is sufficient to explain the level of significance observed in the CMAT. Genotypes for these variants were uncorrelated, indicating they are likely on different haplotype backgrounds and therefore independently contribute to the CMAT statistic. Thus, the significance of the *SKIV2L* CMAT statistic is driven by the cumulative effect of the four variants. Since imputation accuracy, indicated by \hat{R}^2 , is high for each variant, it is unlikely that the observed signal is the result of low imputation quality.

Analysis of common variation in the HLA region indicated the potential for additional functional variants in the same or different genes after conditioning on *rs12191877*. Our result for *SKIV2L* may indicate such an additional psoriasis locus in this region and is an interesting candidate for further investigation.

3.4 Discussion

We described the CMAT, a simple method for jointly testing multiple rare variants in case-control sequence data that can be easily extended to deal with typical challenges of modern genomic studies. Notably, our statistic accepts expected minor allele counts from probabilistic genotypes, making it applicable to both low-coverage sequencing and imputed data. The statistic can incorporate qualitative covariates allowing correction for confounders such as population stratification. Moreover, the CMAT is both computationally fast and straightforward to implement.

We assessed the CMAT by applying it to simulated case-control sequencing datasets specifically designed to contain realistic levels of neutral variation. We also considered three alternative testing strategies, a private allele test similar to the one used by Cohen et al[11], the collapsing test described by Li and Leal[34] and the Weighted Sum Statistic (WSS) of Madsen and Browning[41]. We considered levels of variant misspecification that are representative of exon-only sequencing to entire genic regions. Our results indicated that the strategy of focusing on exonic variants is appropriate if most risk rare variants are located in exons. However, if we assume that the majority of rare risk variants are located in regulatory regions, including variants from outside the exons may be the more powerful strategy than analyzing only exonic variants.

That is, the increase in signal from including non-coding risk variants can outweigh the additional noise of non-coding neutral variants. Comparing the different tests, we noticed that the CMAT, WSS and collapsing test were equally powerful for the exon only model. However, the CMAT and the WSS were more robust to variant misspecification, and were therefore significantly more powerful when analyzing data representative of whole gene analysis. The CMAT provides similar power to the WSS and is computationally more efficient. As the WSS is based on ranking individuals, its computation time is bounded by the theoretical maximum of $O(n \log(n))$; the computation time of the CMAT is linear with sample size. This difference can be substantial when analyzing large sample sizes in genome-wide studies.

A pooling statistic that accepts probabilistic genotypes dramatically increases the range of possible rare variant study designs. Our simulations demonstrated the potential of including genotypes from both direct sequencing and imputation in the test statistic. As genotypes for rare variants are generally imputed with higher error rates than common variants it is important to propagate this uncertainty into the analysis using expected minor allele counts in the CMAT, as opposed to most likely genotype (unpublished data). Our simulation results show that substantial power can be gained by augmenting sequencing datasets with imputed samples. In particular, sequencing only a fraction of available individuals and imputing the remainder can recoup much of the power of a study that sequences all samples and provides a major cost reduction. Other methods for testing rare variants can likely be adapted to obtain a comparable gain of efficiency from imputed data. Note that we modeled sequencing an equal number of cases and controls, but more powerful sequencing strategies for observing risk alleles may exist, for example sequencing mainly cases[35].

We also provided an example of a rare variant analysis that does not require sequencing. Instead, rare variants can be imputed into existing GWAS datasets from publicly available reference panels. Presently, imputed rare variants are typically discarded because single marker tests have limited power to detect variants with low maf and the increased uncertainty in imputing rare variants[28]. Pooling these variants and testing their cumulative effect is more powerful and may uncover additional signals in the data. We used the haplotypes from the CEU samples in the 1000 Genomes Project to impute rare variants into the existing GAIN Psoriasis GWAS dataset. Our analysis shows that the CMAT can identify interesting genes that cannot be found by single marker tests. The identified gene (*SKIV2L*) contains multiple rare variants, none of which achieved genome-wide significance in a single marker test. *SKIV2L* resides in the HLA region of chromosome 6 which is thought to harbor

multiple psoriasis susceptibility genes. However, the biological interpretation is not clear. *SKIV2L* is not an obvious candidate for psoriasis. While *SKIV2L* may be a psoriasis locus, it is also conceivable that multiple rare variants in *SKIV2L* tag the same functional common variant in another gene and the observed signal might be the result of reverse synthetic association [16]. Further analysis is necessary to validate this finding. The analysis was limited by the size our reference panel containing only 112 haplotypes. Only 2889 genes contained 2 or more coding variants with $\text{maf} < 0.05$ in this panel and were thus eligible for the pooled analysis. Future releases from the 1000 Genomes Project should provide low coverage sequencing of 2500 individuals and deep exome resequencing of the same 2500 individuals[18]. This will increase the number of imputable rare variants, making this analysis method more powerful.

Accurate prediction of functionally relevant sites and appropriate weighting will reduce variant misspecification and may further improve power of pooling methods. The weighting scheme proposed by Madsen and Browning[41] is based on allele frequency and is most powerful for risk variants under relatively high purifying selection[56]. Alternatively, variants can be weighted according to predictions of molecular function. In practice, bioinformatic tools such as PolyPhen[62] and SIFT[52] are useful in predicting deleterious potential but are typically limited to coding variants. Determining functionality of non-coding variants is more difficult and although databases containing known phenotype altering non-coding variants exist (PupaSuite[64], for example), these are not applicable to novel variants. Instead, identifying conserved regulatory regions within non-coding portions of a gene will be crucial in determining which non-coding variants have phenotype altering potential and should be included in an analysis[40]. For this paper, rather than attempting to optimize weights for our specific disease model, we assumed very simple uniform weights and focused on the overall performance of our test with respect to variant misspecification and imputation. However we have included a general weighing term into the statistic allowing any desired scheme to be incorporated.

Our simulation results are based on several underlying assumptions. Like other methods, we assume that all rare variants pooled together have the same type of effect. That is, either all are causative, the likely model if risk variants are under purifying selection[58], or they are all protective. If this assumption is violated and causal and protective alleles are combined into a single statistic, pooling methods will lose power. Our results also depend on our disease model, specifically the range of allele frequencies and effect sizes for risk variants. The true frequency spectrum for risk alleles will depend on the strength of purifying selection at the locus and can

range from extremely rare family-specific mutations to so-called ‘goldilocks’ alleles that segregate at low frequency in the population[56]. We evaluated a combination of both models allowing frequencies between .01% and 5% for risk variants. However, we showed that our results also apply to analyses restricted to rarer variants between 0.1% and 1%. Since we are interested in variants that would not be detected by existing association methods, we assigned larger relative risks to rarer alleles. Our results therefore apply to this class of risk variants and do not generalize to extremely rare variants with Mendelian inheritance patterns. In particular, we note that our choice of disease model explains the poor performance of the private allele test which is best suited for testing highly penetrant Mendelian-like risk alleles segregating within families. We have included it in our analysis because it is currently one of the few statistical tests that has successfully provided evidence for rare variant associations.

In summary, the CMAT is a powerful and versatile tool for analyzing the contribution of rare variants to the heritability of common complex diseases. The test accounts for uncertainty in genotypes from imputation methods and can be used to reanalyze existing GWAS datasets.

3.5 Appendix

3.5.1 Empirical distributions for expected minor allele counts

We assume a set of cases and controls genotyped for a set of tagSNPs across a 1Mb segment that contains a 100kb region of interest. We assume N_{seq} cases and controls are randomly selected and sequenced at deep-coverage in the 100kb region. Variants observed among the sequenced samples in the region of interest are imputed into the remaining samples.

We created empirical distributions of expected minor allele counts for imputed genotypes assuming sequence data for $N_{seq} = 100, 200$ and 400 cases and controls and tagSNPs for the remainder of the sample. For each, we first simulated ten independent populations of ten-thousand 1Mb haplotypes using *cosi* and selected a set of tagSNPs for each region that mimicked real-world tagging properties[67]. For each 1Mb region, the 100 selected tagSNPs resulted in $\sim 78\%$ of the common variants having an $r^2 \geq 0.8$ with one of the selected tagSNPs, similar to the tagging properties of the Illumina HumanHap300 BeadChip SNP genotyping platform. From each population, we drew a random subset of 4,000 haplotypes and treated the first $2 \times N_{seq}$ as sequenced in the middle 100kb region of interest (these sample sizes correspond to datasets with $N = 1000$ and N_{seq} sequenced cases and controls).

We statistically phased the $2 \times N_{seq}$ haplotypes across the entire 1Mb region. These phased haplotypes then served as a reference panel for imputing the variants observed in the middle 100kb into the remaining haplotypes. Phasing and imputation were performed using the software program MaCH[37]. MaCH includes a ‘states’ option that speeds computation by limiting the number of haplotypes considered at each iteration of phasing or imputation. Since our analysis focused on rare variants that may only appear on a few haplotypes, we did not use the states shortcut. This likely prolonged computation time but improved imputation accuracy.

We observed that imputation accuracy for rare variants was dependent on the total number of haplotypes in the reference panel ($2 \times N_{seq}$) as well as the number of times a variant was observed in the reference panel. Therefore, we created empirical sampling distributions by binning the observed expected minor allele counts (dosage) by true underlying genotype and the number of times the minor allele was observed in the reference panel. We pooled analogous distributions across all ten realizations to average over varying degrees of linkage disequilibrium. The distributions for true heterozygotes were bimodal with peaks at 1.0, the true dosage for a heterozygote, and 0.0, the true dosage for a major allele homozygote. As the minor allele is observed

more often in the reference panel, imputation was more accurate, indicated by the density of the peak at 0.0 shifting to larger dosage values. Table 2 summarizes these empirical distributions for $N_{seq} = 100$. The < 0.1 and ≥ 0.9 columns capture the density in the two peaks. The distributions for true major allele homozygotes consist of a point mass at 0.0 with a small amount of density just above 0.0. As the number of minor alleles observed in the reference panel increases, the density shifts slightly away from the point mass.

Minor Allele Count in Reference Haplotypes	Fraction of Heterozygote Minor Allele Dosages			
	< 0.1	$[0.1, 0.5)$	$[0.5, 0.9)$	≥ 0.9
1	0.729	0.120	0.063	0.088
2	0.331	0.188	0.133	0.349
3	0.291	0.169	0.128	0.413
4	0.199	0.170	0.162	0.469
5	0.327	0.176	0.162	0.335
6	0.255	0.180	0.136	0.428
7	0.166	0.149	0.132	0.553
8	0.203	0.195	0.179	0.422
9	0.091	0.114	0.128	0.667
10	0.100	0.159	0.195	0.546
11-20	0.061	0.094	0.118	0.727
21-30	0.043	0.054	0.092	0.811
31-40	0.016	0.039	0.081	0.865
41-50	0.016	0.051	0.100	0.834
51-60	0.006	0.023	0.050	0.921
61-70	0.011	0.039	0.087	0.863
71-80	0.009	0.026	0.072	0.893
81-90	0.007	0.017	0.054	0.923
91-100	0.005	0.019	0.065	0.911

Table 3.2: Summary of empirical distributions of minor allele dosage for true heterozygotes. Each distribution is conditional on the indicated minor allele count in the reference haplotypes. Here we report results for $N_{seq} = 100$ sequenced cases and controls.

3.6 Supplementary material

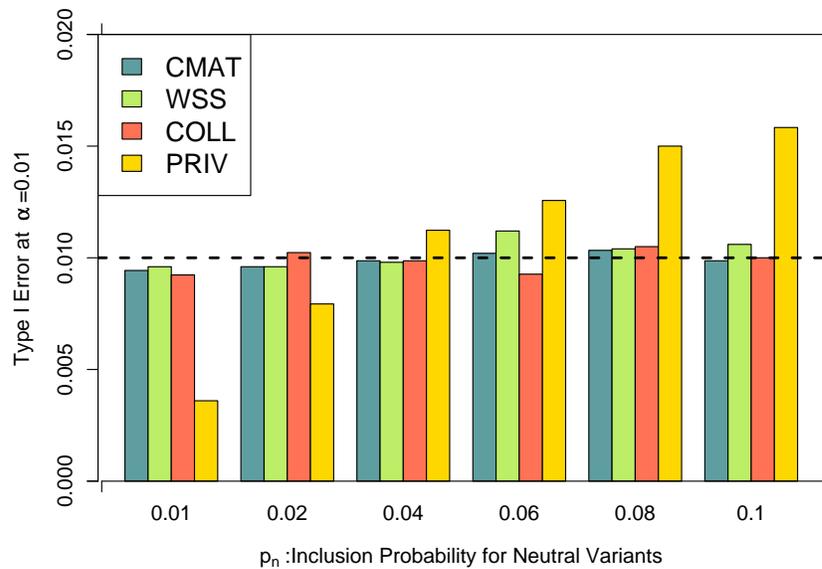


Figure 3.6: Type I error rates for the rare variant tests. Power is computed at $\alpha = 0.01$ on datasets with $N = 1000$ cases and controls at the indicated neutral inclusion rates (p_n). Type I error for the CMAT and collapsing test (COLL) is well-controlled across values of p_n . Type I error for the private allele test (PRIV) is initially conservative then increases with p_n becoming anti-conservative.

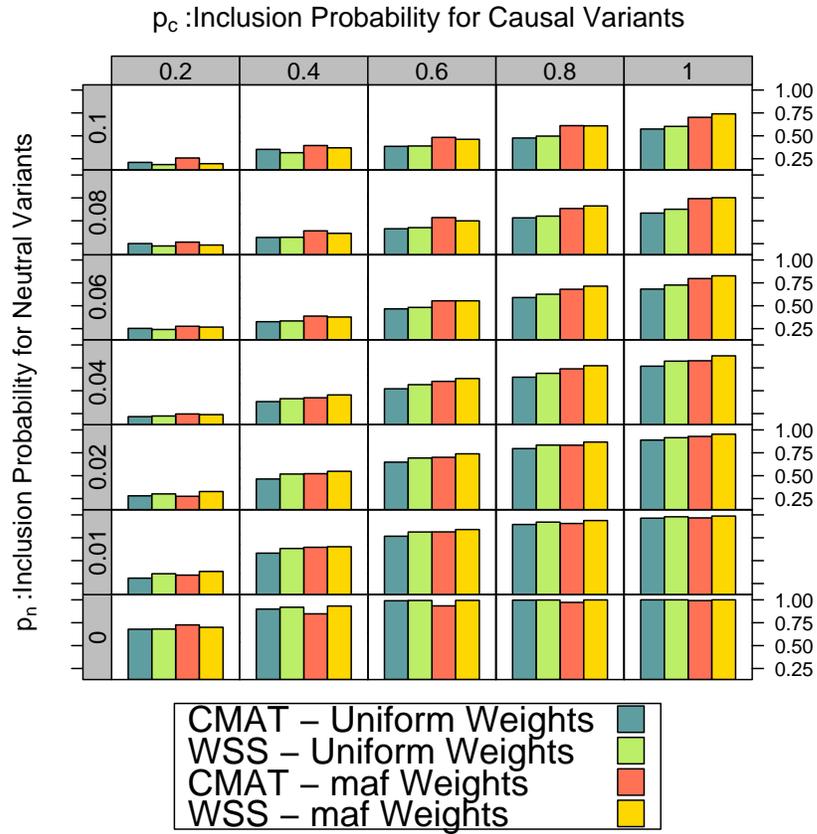


Figure 3.7: Power for CMAT and WSS using both uniform and maf-based weights described in Madsen and Browning. Conditional on weighting scheme, the CMAT and WSS have similar power across the grid. The maf-based weights correspond more closely to our disease model than do the simple uniform weights and therefore provided a more powerful analysis for both methods except when misspecification rates are highest.

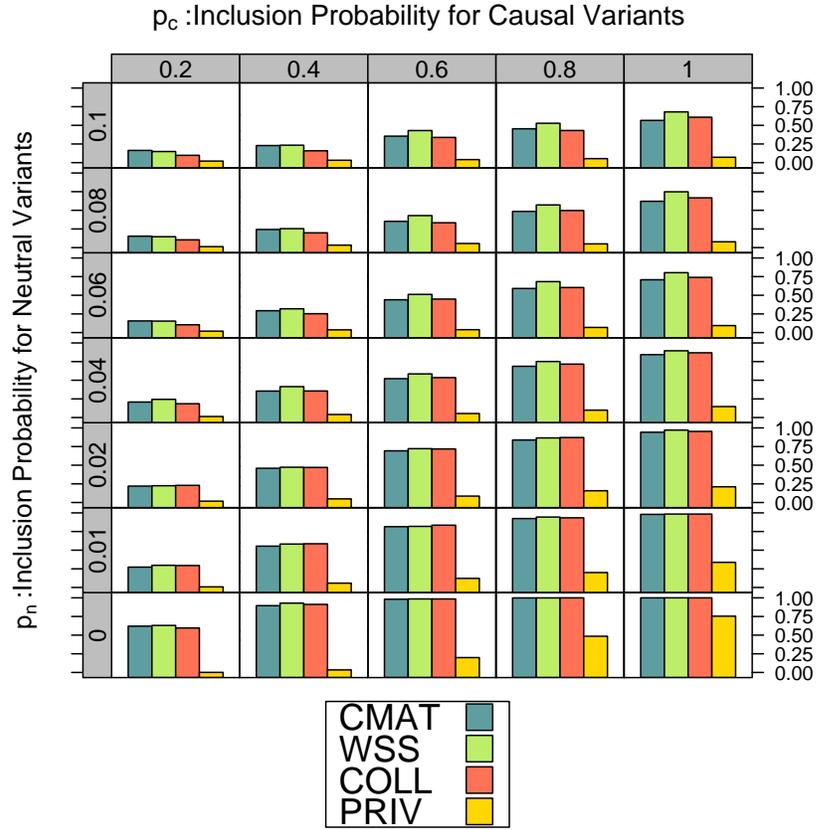


Figure 3.8: Power to analyze deep sequence datasets for minor allele cutoff $\beta = 1\%$. Each dataset contains exact genotypes for $N = 1000$ cases and controls based on $k = 15$ causative variants in the population. Along the vertical axis we vary the probability of (incorrectly) including a neutral variant (p_n) in the analysis and along the horizontal axis we vary the probability of (correctly) including a causative variant (p_c). The height of the bars in each cell indicates the power for the four tests at $\alpha = 0.01$. Here, the maximum allele frequency for risk variants (p_{max}) was set to 1%.

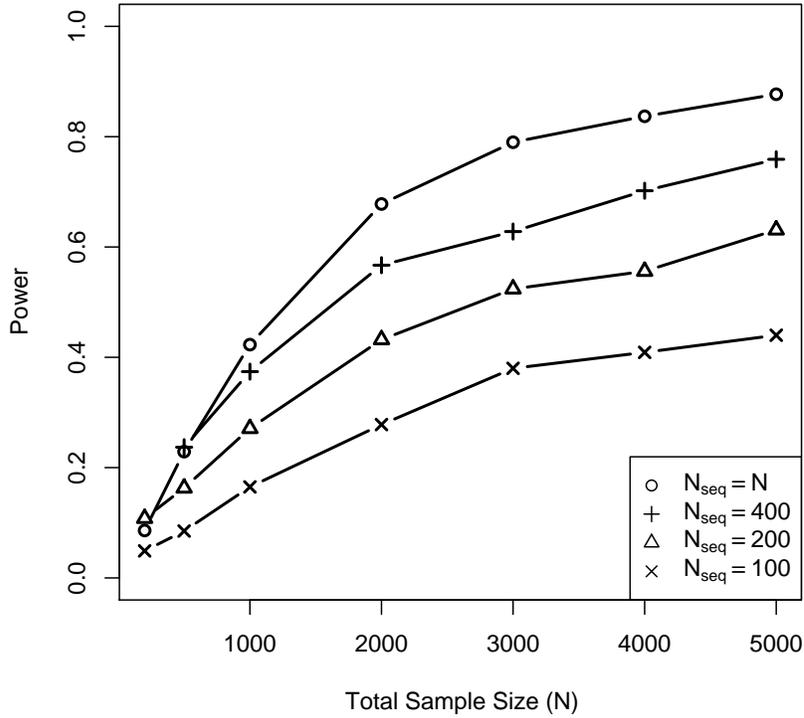


Figure 3.9: CMAT power for imputation datasets with minor allele cutoff $\beta = 1\%$. Datasets contain exact genotypes for N_{seq} cases and controls and probabilistic genotypes based on imputation for the remaining samples. The top line shows CMAT power for Deep Sequencing datasets ($N_{seq} = N$) and serves as an upper bound for power at a fixed total sample size N . Here, the maximum allele frequency for risk variants (p_{max}) was set to 1%. We report power at $\alpha = 0.01$ using the whole-gene inclusion threshold ($p_n = 0.1, p_c = 0.8$).

CHAPTER IV

A Coalescent Model for Genotype Imputation

4.1 Introduction

Genotype imputation, the estimation of genotypes at untyped markers using reference patterns of haplotype structure, has proven to be a powerful tool in modern genetic studies [38, 43]. Imputation is routinely used to increase the fraction of the genome covered in genome-wide association studies (GWAS) thereby increasing the power to detect risk variants through linkage disequilibrium (LD) mapping [5, 9, 24, 39, 44, 74]. Large-scale meta-analyses have been made possible by imputing a mutual set of markers into datasets genotyped on different platforms [4, 14, 82]. Imputation promises to be as important in future genetic studies that use sequencing technology to target rare variation. Algorithms that call genotypes from sequencing reads can be improved by including imputation methods that employ LD information, particularly for low pass sequencing [36, 39]. Power for rare variant burden tests of association can be increased by augmenting sequencing datasets with imputed samples; thus providing a balance between sequencing cost and statistical power [81].

Imputation procedures involve a set of target samples in which genotypes are to be imputed, a reference panel of phased haplotypes from which genotypes are copied and an algorithm for applying the copying procedure. Each target samples is genotyped for a set of single nucleotide polymorphisms (SNPs) whose genotypes serve as a scaffold for the haplotypes of the sample; the SNPs typed in the target samples are often selected based on patterns of LD to optimize power for LD mapping and imputation accuracy [61]. Haplotypes in the reference panel are either more densely typed than those in the targets, or fully sequenced. Algorithms for imputing the genotypes of markers untyped in the target samples [7, 39, 44, 60, 68] typically employ a Hidden Markov Model that uses the genotypes from the SNP-scaffold of each target sample to choose the haplotypes from the reference panel that are most similar. Genotypes

are then imputed into the target by copying the reference haplotypes that provide the best match.

Imputation accuracy is known to depend on several factors, including reference panel size, genetic diversity of the target population, and genetic similarity of the reference panel and the imputation targets [2, 6, 27, 29]. For example, imputation is likely to be most accurate when the reference panel is drawn from the same population as the target samples, or from one that is closely related [2, 26, 39]. Densely-typed haplotypes from the publicly available HapMap dataset [2] serve as the imputation reference panel in most current GWAS. Although these haplotypes are usually not derived from the same population as the GWAS target samples, they provide excellent imputation accuracy for common variants in many populations [2]. Reference panels based on sequenced haplotypes, from the 1000 Genomes Project for example [18], will allow imputation of rarer variants. These rarer variants, however, are often the result of more recent mutation events than common variants. As a result, recent population history between reference haplotypes and target samples may be more important for imputing rare variants. Next generation sequencing provides the possibility of creating custom-made reference panels by sequencing a subset of a large dataset to use as references for imputing variation into the remainder of the sample. Here we determine the improvement in imputation accuracy attained with custom-made reference panels compared to large, publicly available panels.

A thorough analysis on the effect of population subdivision on imputation accuracy requires a wide range of study variables, such as reference panel size and relation to the targets, to be considered. Currently, imputation accuracy is assessed by masking a subset of known genotypes in a dataset and imputing genotypes at these sites using an available imputation software program. Accuracy is inferred by comparing the imputed genotypes with the true genotypes. Both real SNP data and simulated data have served as the true underlying genotypes in this analysis design. Real data has the advantage that it reflects true LD patterns and population structures in humans. However, the scope of investigation is limited by the amount of real data that is available. Simulated datasets allow a wider range of scenarios to be considered. Regardless of the underlying data type, the masking method requires performing a large number of imputation experiments, each at a discrete set of parameter combinations; a time-consuming procedure that restricts prediction of imputation behavior across a range of study variables. Developing a method to obtain fast and accurate estimates of imputation accuracy would allow a more detailed analysis of reference panel characteristics that impact imputation accuracy.

In this paper, we propose a theoretical model of imputation based on coalescent theory [30]. Our imputation model considers a simple rule that mirrors imputation algorithms, investigating imputation accuracy as a function of reference panel size and population demography from which target and reference haplotypes are derived. Our imputation rule relies on the premise that, for a given target haplotype, an imputation algorithm will ideally choose as the template the reference haplotype with the fewest sequence differences from the target. In our model, we take the reference haplotype that is most closely related to the target, in terms of coalescence time, to serve as the template since it will have, on average, the fewest sequence differences from the target. Using this rule, we derive and analyze several equations based on coalescent theory to quantify imputation accuracy.

A feature of the coalescent framework is its ability to model complex population demographics, allowing a range of imputation study designs to be evaluated. Here, we use our model to investigate the effect of population subdivision between reference and target haplotypes on imputation accuracy. Given a set of target samples, we consider two potential reference panels. The first is an “internal” reference panel of sequenced haplotypes drawn from the same population as the target samples, and the second is an “external” reference panel of sequenced haplotypes from a distinct population. In practice, the external reference panel could be from a publicly available sequencing dataset such as the HapMap or 1000 Genomes Projects and the internal reference panel could be a custom-made set of sequenced haplotypes from the same dataset as the targets. Our model predicts that internal reference panels, even when they are considerably smaller than a competing external reference panel, are nearly always the optimal choice for imputation. The relative improvement in imputation accuracy for using an internal panel varies according to both the divergence time and growth rates for the populations. As divergence time increases, small internal reference panels can produce fewer imputation errors than much larger external panels. However, exponential growth attenuates the effect of the divergence time, improving the relative performance for external reference panels.

4.2 Coalescent model for genotype imputation

The coalescent is a stochastic model that describes the genealogical history of a set of haplotype lineages [30]. A genealogy is defined by a series of coalescence events in which pairs of lineages coalesce to form single lineages. The basic coalescent model assumes a single homogeneous population with constant size. Given n lineages in

a population with constant effective size of N diploid individuals, the waiting time until the next coalescence event is exponentially distributed with rate $n(n-1)/2$, in units of $2N$ generations. Independent of the time of the coalescent event, one of the $\binom{n}{2}$ possible pairs of lineages is chosen to coalesce. The process proceeds iteratively until only one lineage remains. This basic coalescent model can be extended to accommodate more complicated modeling assumptions, such as population structure and non-constant population size.

The coalescent algorithm described above produces a tree structure that defines the genealogical history for a set of haplotypes sampled in the present. The lengths of the branches of the tree indicate the times between coalescent events. Conditional on this tree, segregating mutations can be incorporated into the model by placing mutation events on the tree branches. This placement is typically done according to a Poisson process, so that the rate at which mutation events accumulate is directly proportional to coalescent time. Thus, as the coalescence time between two haplotypes increases, the expected number of segregating sites between the two sequences will also increase.

In this paper, we use a coalescent framework to model genotype imputation at a non-recombining genetic locus intended to represent a short region along a chromosome. We assume that genotypes for a single haplotype T are to be imputed. We define a reference panel to be a set of sequenced or densely genotyped haplotypes, not including T but presumed to be representative of T . We assume that one haplotype from the reference panel will be chosen as an imputation template and that all alleles from the template haplotype will be copied onto T . Assuming that mutations accumulate in direct proportion to coalescent time, the reference haplotypes with the fewest sequence differences from T are the descendants of the lineage with which T first coalesces. If we consider an imputation algorithm that always selects the haplotype in the reference panel with the fewest sequence differences from T (or one such haplotype in case of a tie), under the assumption that mutations accumulate in direct proportion to time, the procedure used by the algorithm is equivalent to choosing the reference haplotype with the closest genealogical history to T . Thus, jointly modeling the genealogy of the target haplotype T and the reference haplotypes in a coalescent framework can allow us to study the accuracy with which genotypes of T can be imputed.

By including haplotypes from multiple reference panels in our model, we can compare the performance of reference panels with different sizes or populations of origin. Here, we consider a scenario with two possible reference panels for imputing

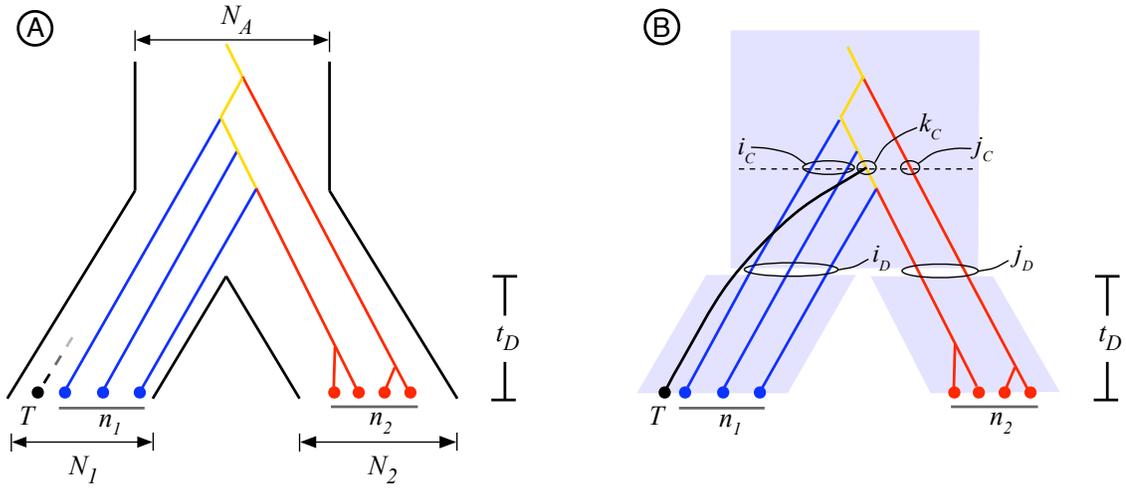


Figure 4.1: Two population coalescent model for imputation reference panel selection. **A** Two populations, labeled 1 and 2, of size N_1 and N_2 diploid individuals, respectively, diverge from an ancestral population of size N_A at time t_D . A single haplotype T for which genotypes at untyped markers are to be imputed is sampled from population 1. We consider two possible reference panels for imputing T : an “internal” reference panel of n_1 haplotypes sampled from population 1 and an “external” reference panel of n_2 haplotypes sampled from population 2. If T first coalesces with a type-1 lineage (blue), then the internal reference panel is optimal for imputing T (event C_1). The external reference panel is optimal (event C_2) if T first coalesces with a lineage of type 2 (red). Finally, if T first coalesces with a type 1-2 lineage (yellow), then the two reference panels are equivalent for imputing T (event C_{12}). **B** To compute the probability of optimality for each reference panel we condition on \mathcal{D} (the event that T coalesces before the divergence), the quantities i_D and j_D (the number of lineages originating in populations 1 and 2, respectively, that remain at the time of divergence), and i_C, j_C and k_C (the number of type 1, type 2 and type 1-2 lineages remaining at the instant when T first coalesces). In the realization pictured, T does not coalesce before the divergence time (event \mathcal{D}^C) and $i_D = 3, j_D = 2, i_C = 2$ and $j_C = k_C = 1$. Since T first coalesce with a type 1-2 lineage in the figure (event C_{12}) the two reference panels are equivalent for imputing T .

T . The first is an “internal” reference panel consisting of n_1 haplotypes sampled from the same underlying population as T . The second is an “external” reference panel consisting of n_2 haplotypes sampled from a distinct population. Defining the “optimal panel” as that which produces highest imputation accuracy for imputing T , we compute the probability of optimality for both the internal and external reference panels and quantify the relative gain in imputation accuracy obtained by using the optimal rather than the suboptimal panel.

To address these questions, we model the genealogical history of T and the reference haplotypes using a two-population coalescent model of divergence [66, 75](Figure 4.1A). Let the two populations be labeled 1 and 2, and assume that T is sampled from population 1. We assume that these populations diverged from an ancestral population at time t_D in the past and that no migration has occurred between the descendant populations. Therefore, more recently than the divergence time ($t < t_D$), a lineage can coalesce only with other lineages from the same population. This assumption provides a reasonable representation for pairs of populations that are geographically isolated. More anciently than the split ($t > t_D$), all remaining lineages are assumed to belong to a homogeneous ancestral population, and any two lineages are allowed to coalesce, regardless of the populations from which they originate. We assume constant effective population sizes of N_1 diploid individuals for population 1, N_2 for population 2, and N_A for the ancestral population.

At time $t = 0$, corresponding to the present, n_1 reference haplotypes in addition to the single target haplotype T are sampled from population 1 and n_2 reference haplotypes are sampled from population 2. The divergence time t_D and the reference panel sizes n_1 and n_2 are treated as model parameters. We refer to a lineage that has descendants only in population 1 as a lineage of type 1. Similarly, we refer to a lineage with descendants only in population 2 as a lineage of type 2. A lineage with descendants in both populations is a lineage of type 1-2. Based on the assumption that mutations accumulate in direct proportion to coalescent time, the best haplotypes for imputing T among all available reference haplotypes are the descendants of the lineage with which T first coalesces. Thus, the internal reference panel is optimal if T first coalesces with a lineage of type 1 and the external reference panel is optimal if T first coalesces with a lineage of type 2. If T first coalesces with a lineage of type 1-2, then the two reference panels are equally appropriate for imputing T and we say that they are both optimal.

The flexibility of the coalescent allows us to easily extend the model to include changes in population size. To model such changes, we let $N_1(t)$, $N_2(t)$ and $N_A(t)$ be

functions that define the size of populations 1, 2 and the ancestral populations, respectively, at time t . Given the utility of population expansion models for explaining properties of human genetic variation [13, 67], we consider a model of exponential growth in populations 1 and 2 (Fig. 4.2). Assuming a constant-sized ancestral population, we set $N_1(t) = N_1 e^{-\alpha_1 t}$ and $N_2(t) = N_2 e^{-\alpha_2 t}$ for $t \in [0, t_D]$ and $\alpha_1, \alpha_2 > 0$. We compare the results from this exponential growth model to those of the constant-size model to determine the potential effects of human exponential growth on imputation reference panel selection.

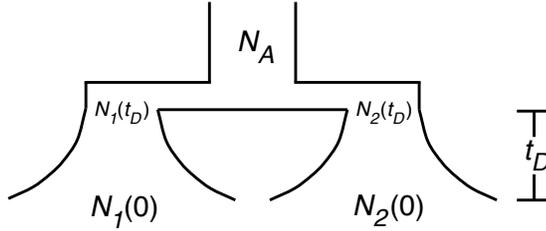


Figure 4.2: The two-population coalescent model of divergence assuming exponential growth in the descendant populations. The sizes of populations 1 and 2 change over time according to $N_1(t) = N_1 e^{-\alpha_1 t}$ and $N_2(t) = N_2 e^{-\alpha_2 t}$, respectively, for $t \in [0, t_D]$. The quantities $\alpha_1, \alpha_2 > 0$ represent growth rates and N_1, N_2 represent the sizes of populations 1 and 2 in the present. At time t_D , populations 1 and 2 merge instantaneously into the ancestral population which has a constant size N_A .

4.3 Methods

We use our coalescent model of genotype imputation to determine reference panel optimality and to quantify the differences in imputation accuracy between potential reference panels. For the problem of imputing the target T , we first derive the probability of optimality for both an internal reference panel of n_1 haplotypes and an external reference panel of n_2 haplotypes, sampled from populations with a divergence time of t_D (in units of N_A generations). Let C_1 be the event that the imputation target T first coalesces with a lineage of type 1, let C_2 be the event that T first coalesces with a lineage of type 2, and let C_{12} be the event that T first coalesces with a lineage of type 1-2. It follows from our definition of optimality that $P(C_1)$, the probability that target T first coalesces with a lineage of type 1, is the probability that the internal reference panel is optimal for imputing T . Similarly, $P(C_2)$ is the probability that the

external reference panel is optimal for imputing T and $P(C_{12})$ is the probability that the two reference panels are both optimal, with equal expected imputation accuracy.

In the case that exactly one reference panel is optimal, it is of interest to quantify the improvement in imputation accuracy that results from using the optimal as opposed to the suboptimal reference panel. Assuming that mutations follow a Poisson process, the expected number of sites incorrectly imputed in T is proportional to the time of coalescence between the target and the imputation template, the reference haplotype from which genotypes are copied. We have assumed that an imputation program will select the reference haplotype that has minimal coalescence time with T , thereby minimizing the expected number of incorrectly imputed sites. It follows that expected imputation accuracy for a given reference panel can be quantified by using the expected time that T first coalesces with a haplotype from that reference panel. In the following sections, we derive several measurements of imputation error, all based on expected coalescence times, to quantify the improvement in accuracy for one rather than the other of two potential reference panels.

4.3.1 Derivation of reference panel optimality probabilities

We consider two approaches for obtaining optimality probabilities, an exact computation and a recursive computation.

4.3.1.1 Exact computations

The events C_1 , C_2 and C_{12} correspond to the internal reference panel being optimal for imputing T , the external reference panel being optimal and the two reference panels being equally appropriate, respectively. To compute the probabilities for each of these events, we partition the coalescent model into three components: the separate periods in populations 1 and 2 more recently than the divergence, and the period in which only the ancestral population is present (Figure 4.1B). Because we assume that no migration occurs between populations 1 and 2 following their divergence from the ancestral population, the events occurring in populations 1 and 2 more recently than the divergence are independent and can be considered separately. Conditional on the number of lineages from populations 1 and 2 at the divergence time, the coalescent events in the ancestral population are independent of the events that occur more recently than the divergence time.

First, we consider the genealogy of haplotypes in population 1 from the present back to the divergence time t_D . Define \mathcal{D} to be the event that lineage T coalesces

more recently than time t_D and \mathcal{D}^c to be the event that T does not coalesce by t_D . Note that if T coalesces before time t_D , then the lineage with which it first coalesces can only have descendants in population 1 and must therefore be of type 1. It follows that if the event \mathcal{D} occurs, then C_1 also occurs and $P(C_1, \mathcal{D}) = P(\mathcal{D})$. If, however, T does not coalesce before the divergence time, it enters the ancestral population, where it can also coalesce with lineages of types 2 and 1-2. In the latter scenario we must consider the coalescent events in population 2 and the ancestral population.

The coalescent process in the ancestral population is dependent on the numbers of lineages from populations 1 and 2 that survive at the divergence time. Let $i_D, 1 \leq i_D \leq n_1$, denote the number of lineages, other than T , that originate from population 1 and survive to enter the ancestral population at the divergence time t_D . Similarly, let $j_D, 1 \leq j_D \leq n_2$, be the number of lineages originating from population 2 that survive at the divergence time t_D . Because the coalescent processes in populations 1 and 2 are independent, j_D is independent of both the quantity i_D and the event \mathcal{D} .

By conditioning on the coalescent history of T and the number of lineages from each population remaining at the divergence time t_D , we can write the quantity $P(C_1)$ as follows

$$\begin{aligned}
P(C_1) &= P(C_1, \mathcal{D}) + P(C_1, \mathcal{D}^c) \\
&= P(\mathcal{D}) + \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} P(C_1, \mathcal{D}^c, i_D, j_D) \\
&= P(\mathcal{D}) + \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} P(C_1 | \mathcal{D}^c, i_D, j_D) P(\mathcal{D}^c, i_D, j_D) \\
&= P(\mathcal{D}) + \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} P(C_1 | \mathcal{D}^c, i_D, j_D) P(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2), \tag{4.1}
\end{aligned}$$

where the last equality follows from independence between populations 1 and 2 and $h_{n,l}(t; N)$ is the probability that n lineages sampled from a diploid population with effective size N coalesce down to l lineages at time t . Tavaré [76] demonstrated that

$$h_{n,l}(t; N) = \sum_{m=l}^n \frac{(2m-1)(-1)^{m-l} l_{(m-1)} n_{[m]}}{l!(m-l)!n_{(l)}} e^{-\binom{m}{2} \frac{t}{2N}} \tag{4.2}$$

where $n_{[m]} = n(n-1)\cdots(n-m+1)$ and $n_{(m)} = n(n+1)\cdots(n+m-1)$.

To obtain the probability $P(\mathcal{D}^c, i_D)$, let $\mathcal{A}_n^l(t; N)$ be the event that n lineages in a diploid population of effective size N coalesce down to l lineages at time t and let

$I_{n,l} = \binom{n}{2} \binom{n-1}{2} \cdots \binom{l+1}{2} = n!(n-1)!/[2^{n-l}l!(l-1)!]$ be the number of ways that the n lineages can coalesce down to l lineages [66]. Then $P(\mathcal{D}^c, i_D)$ is derived by noting that for \mathcal{D}^c and i_D to both occur, the $n_1 + 1$ total lineages originating in population 1 must coalesce down to $i_D + 1$ lineages by the divergence time t_D . This can occur in I_{n_1+1, i_D+1} ways. In I_{n_1, i_D} of these, the n_1 reference haplotypes coalesce down to i_D lineages without T also coalescing. Thus,

$$\begin{aligned}
P(\mathcal{D}^c, i_D) &= P(\mathcal{D}^c, \mathcal{A}_{n_1+1}^{i_D+1}(t_D; N)) \\
&= P(\mathcal{D}^c | \mathcal{A}_{n_1+1}^{i_D+1}(t_D; N)) P(\mathcal{A}_{n_1+1}^{i_D+1}(t_D; N)) \\
&= \frac{I_{n_1, i_D}}{I_{n_1+1, i_D+1}} P(\mathcal{A}_{n_1+1}^{i_D+1}(t_D; N)) \\
&= \frac{i_D(i_D + 1)}{n_1(n_1 + 1)} h_{n_1+1, i_D+1}(t_D; N_1).
\end{aligned} \tag{4.3}$$

The probability $P(\mathcal{D}^c)$ is obtained by summing over all possible values of i_D

$$P(\mathcal{D}^c) = \sum_{i_D=1}^{n_1} P(\mathcal{D}^c, i_D) \tag{4.4}$$

and the probability $P(\mathcal{D})$ in Equation (4.1) can be computed as

$$P(\mathcal{D}) = 1 - P(\mathcal{D}^c) = 1 - \sum_{i_D=1}^{n_1} P(\mathcal{D}^c, i_D). \tag{4.5}$$

The final term to derive in Equation (4.1), $P(C_1 | \mathcal{D}^c, i_D, j_D)$, is the probability that T first coalesces with a lineage of type 1 assuming that, in addition to T , i_D lines from population 1 and j_D lines from population 2 survive to the ancestral population.

To derive $P(C_1 | \mathcal{D}^c, i_D, j_D)$ in closed form, let i_C , j_C , and k_C be the numbers of lineages of type 1, 2, and 1-2, respectively, remaining at the instant when T first coalesces. The probability $P(C_1 | \mathcal{D}^c, i_D, j_D)$ is computed by summing over all possible values of i_C , j_C , and k_C ,

$$P(C_1 | \mathcal{D}^c, i_D, j_D) = \sum_{k_C=0}^{\min\{i_D, j_D\}} \sum_{i_C=\delta_{k_C,0}}^{\min\{i_D, i_D-k_C\}} \sum_{j_C=\delta_{k_C,0}}^{\min\{j_D, j_D-k_C\}} P(C_1, i_C, j_C, k_C | \mathcal{D}^c, i_D, j_D), \tag{4.6}$$

where $\delta_{i,j} = 1$ if $i = j$ and $\delta_{i,j} = 0$ otherwise.

To derive the probability $P(C_1, i_C, j_C, k_C | \mathcal{D}^c, i_D, j_D)$, let $N(i_D, j_D \rightarrow i, j, k)$ be

the number of ways in which i_D lineages of type 1 and j_D lineages of type 2 can coalesce down to i , j , and k lineages of types 1, 2, and 1-2, respectively. Then $P(C_1, i_C, j_C, k_C | \mathcal{D}^c, i_D, j_D)$ is given by

$$P(C_1, i_C, j_C, k_C | \mathcal{D}^c, i_D, j_D) = \frac{N(i_D, j_D \rightarrow i_C, j_C, k_C) i_C}{I_{i_D+j_D+1, i_C+j_C+k_C}}. \quad (4.7)$$

The quantity $N(i_D, j_D \rightarrow i, j, k)$ is derived in Appendix 4.6.1.

The probability of optimality for the external reference panel, $P(C_2)$, and the probability that the two reference panels are equally appropriate, $P(C_{12})$, are computed in a similar manner to (4.1). The one notable exception is that because \mathcal{D} implies that the event C_1 has occurred, $P(C_2, \mathcal{D}) = P(C_{12}, \mathcal{D}) = 0$. Thus, the probability that the external reference panel is optimal can be written as

$$P(C_2) = \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} P(C_2 | \mathcal{D}^c, i_D, j_D) P(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2), \quad (4.8)$$

and the probability that the two reference panels are both optimal is

$$P(C_{12}) = \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} P(C_{12} | \mathcal{D}^c, i_D, j_D) P(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2). \quad (4.9)$$

4.3.1.2 Recursive computations

The closed-form equation for $P(C_1 | \mathcal{D}^c, i_D, j_D)$ is computationally intensive for large i_D and j_D . In this section, we derive a more efficient recursive algorithm. Assume that at some time $t > t_D$, in addition to lineage T , i lineages of type 1, j lineages of type 2 and k lineages of type 1-2 exist in the ancestral population. Conditional on this configuration, let $\tilde{P}(C_1 | i, k, j)$ denote the probability that T first coalesces with a lineage of type 1. We construct a recursive equation for $\tilde{P}(C_1 | i, k, j)$ by conditioning on the lineage pair involved in the next coalescent event and considering its effect on the subsequent coalescent process (Table 4.1).

Let $m = i + j + k + 1$ be the total number of lineages remaining, and note that each of the $\binom{m}{2}$ pairs of lineages is equally likely to coalesce. There are nine unique pairs of lineage types that can coalesce in the next event. For each pair of types, we compute the probability that the pair will coalesce and then, conditional on their coalescence, we compute the subsequent probability that T first coalesces with a lineage of type 1. If the next coalescent event involves T and a lineage of type 1, an event that occurs

Lineage Pair For Next Coalescent Event	Resulting Lineage	Number of Ways Event Can Occur	$P(C_1 \text{Event})$
$T,1$	-	i	1
$T,2$	-	j	0
$T,1-2$	-	k	0
1,1	1	$\binom{i}{2}$	$\tilde{P}(C_1 i-1, k, j)$
1,2	1-2	ij	$\tilde{P}(C_1 i-1, k+1, j-1)$
1,1-2	1-2	ik	$\tilde{P}(C_1 i-1, k, j)$
2,1-2	1-2	jk	$\tilde{P}(C_1 i, k, j-1)$
2,2	2	$\binom{j}{2}$	$\tilde{P}(C_1 i, k, j-1)$
1-2,1-2	1-2	$\binom{k}{2}$	$\tilde{P}(C_1 i, k-1, j)$

Table 4.1: Derivation of the recursion $\tilde{P}(C_1|i, k, j)$. Assume that, in addition to lineage T , i lineages of type 1, j lineages of type 2 and k lineages of type 1-2 exist in the ancestral population at some time $t > t_D$. Conditional on this configuration, let $\tilde{P}(C_1|i, k, j)$ denote the probability that T first coalesces with a lineage of type 1. Column 1 lists each possible lineage pair for the next coalescent event. Column 2 gives the resulting lineage type for the coalescent event. Column 3 contains the number of ways each event can occur. Column 4 gives the probability that T first coalesces with a lineage of type 1, conditional on the pair of lineages in column 1 being the next to coalesce. The recursive equation for $\tilde{P}(C_1|i, k, j)$ is attained by conditioning on all the possible lineage pairs for the next coalescent event.

with probability $i/\binom{m}{2} = \frac{2i}{m(m-1)}$, then the event C_1 occurs. Alternatively, if the next coalescent event occurs between T and either a lineage of type 2 or 1-2, events that occur with probabilities $\frac{2j}{m(m-1)}$ and $\frac{2k}{m(m-1)}$, respectively, then the event C_1 cannot occur. For the remaining lineage pairs, T is not involved in the next coalescent event and the probability of the event C_1 depends on the lineage pair that is involved in the event.

If two lineages of type 1 coalesce, an event that occurs with probability $\binom{i}{2}/\binom{m}{2} = \frac{i(i-1)}{m(m-1)}$, then the number of type 1 lineages is reduced by one. Thus, the probability of event C_1 , conditional on two lineages of type 1 coalescing is $\tilde{P}(C_1|i-1, k, j)$. Following this logic, if two lineages of type 2 coalesce, an event that occurs with probability $\frac{j(j-1)}{m(m-1)}$, then the conditional probability of event C_1 is $\tilde{P}(C_1|i, k, j-1)$. Similarly, if two lineages of type 1-2 coalesce, an event that occurs with probability $\frac{k(k-1)}{m(m-1)}$, then the conditional probability of event C_1 is $\tilde{P}(C_1|i, k-1, j)$. If a lineage of type 1 coalesces with lineage of type 2, then the number of type 1-2 lineages increases by one while the numbers of type 1 and type 2 lineages each reduces by one. This occurs with probability $\frac{2ij}{m(m-1)}$ and the conditional probability of event C_1 is then $\tilde{P}(C_1|i-1, k+1, j-1)$.

Finally, if the next coalescent event involves a type 1-2 lineage and either a type 1 or type 2 lineage the resulting lineage will be of type 1-2. Thus, such an event has the effect of reducing either the number of type 1 or type 2 lineages by one. A coalescence between a type 1-2 and a type 1 lineage occurs with probability $\frac{2ik}{m(m-1)}$ and, conditional on this, event C_1 occurs with probability $\tilde{P}(C_1|i-1, k, j)$. Similarly, the probability that the next coalescent event involves a type 2 and a type 1-2 lineage is $\frac{2jk}{m(m-1)}$. The probability of C_1 following this event is $\tilde{P}(C_1|i, k, j-1)$.

By conditioning on the possible lineage pairs for the next coalescent event, we obtain the following recursive equation,

$$\begin{aligned} \tilde{P}(C_1|i, k, j) &= \frac{2i}{m(m-1)} + \frac{i(i-1) + 2ik}{m(m-1)} \tilde{P}(C_1|i-1, k, j) \\ &\quad + \frac{2ij}{m(m-1)} \tilde{P}(C_1|i-1, k+1, j-1) \\ &\quad + \frac{j(j-1) + 2jk}{m(m-1)} \tilde{P}(C_1|i, k, j-1) + \frac{k(k-1)}{m(m-1)} \tilde{P}(C_1|i, k-1, j). \end{aligned} \tag{4.10}$$

Equation (4.10) holds for $i > 0, k, j \geq 0$. $\tilde{P}(C_1|0, k, j) = 0$ for all $k, j \geq 0$ because there must be at least one lineage of type 1 for event C_1 to occur. The recursion is

incorporated into equation (4.1) by replacing $P(C_1|\mathcal{D}^c, i_D, j_D)$ with $\tilde{P}(C_1|i_D, 0, j_D)$.

The terms $P(C_2|\mathcal{D}^c, i_D, j_D)$ in Eq. (4.8) and $Pr(C_{12}|\mathcal{D}^c, i_D, j_D)$ in Eq. (4.9) can be evaluated in a recursive fashion following the logic used to obtain Eq. (4.10). Let $\tilde{P}(C_2|i, k, j)$ be the probability that T first coalesces with a lineage of type 2. Then

$$\begin{aligned}\tilde{P}(C_2|i, k, j) &= \frac{2j}{m(m-1)} + \frac{i(i-1) + 2ik}{m(m-1)} \tilde{P}(C_2|i-1, k, j) \\ &\quad + \frac{2ij}{m(m-1)} \tilde{P}(C_2|i-1, k+1, j-1) \\ &\quad + \frac{j(j-1) + 2jk}{m(m-1)} \tilde{P}(C_2|i, k, j-1) + \frac{k(k-1)}{m(m-1)} \tilde{P}(C_2|i, k-1, j).\end{aligned}\tag{4.11}$$

Equation (4.11) holds for $i, k \geq 0, j > 0$. $\tilde{P}(C_1|i, k, 0) = 0$ for all $k, j \geq 0$ because at least one lineage of type 2 must be present for event C_2 to occur. Eq. (4.11) can replace $P(C_2|\mathcal{D}^c, i_D, j_D)$ in Eq. (4.8).

Finally, conditional on i, j , and k , let $\tilde{P}(C_{12}|i, j, k)$ be the probability that T first coalesces with a lineage of type 1-2. $\tilde{P}(C_{12}|i, j, k)$ can be written in recursive form as

$$\begin{aligned}\tilde{P}(C_{12}|i, k, j) &= \frac{2k}{m(m-1)} + \frac{i(i-1) + 2ik}{m(m-1)} \tilde{P}(C_{12}|i-1, k, j) \\ &\quad + \frac{2ij}{m(m-1)} \tilde{P}(C_{12}|i-1, k+1, j-1) \\ &\quad + \frac{j(j-1) + 2jk}{m(m-1)} \tilde{P}(C_{12}|i, k, j-1) + \frac{k(k-1)}{m(m-1)} \tilde{P}(C_{12}|i, k-1, j).\end{aligned}\tag{4.12}$$

The boundary condition for Eq. (4.12) is

$$\tilde{P}(C_{12}|i, 0, 0) = \tilde{P}(C_{12}|0, 0, j) = \tilde{P}(C_{12}|0, 0, 0) = 0,$$

since at least one lineage of type 1 and one lineage of type 2 must remain to allow the potential for a type 1-2 lineage in the future. The recursion is incorporated into Eq. (4.9) by replacing $P(C_{12}|T_D^c, i_D, j_D)$ with $\tilde{P}(C_{12}|i_D, 0, j_D)$.

4.3.2 Derivation of expected coalescent times

In this section, we quantify and compare imputation accuracy for internal and external reference panels. Let S_1 be the number of sites that are incorrectly imputed when using an internal reference panel and let S_2 be the number of sites incorrectly

imputed when using an external reference panel. In the following, we compute the expected number of incorrectly imputed sites, $E[S_1]$ and $E[S_2]$, assuming that mutation events accumulate in direct proportion to coalescent time.

Let the random variable T_1 be the waiting time until lineage T first coalesces with a lineage that has descendants in the internal reference panel, that is, a type 1 or type 1-2 lineage. Let T_2 be the waiting time until T first coalesces with a lineage that has descendants in the external reference panel, that is, a type 2 or type 1-2 lineage. Here, T_1 and T_2 are measured in units of generations. We assume that an imputation algorithm will select as the imputation template one of the reference haplotypes whose coalescence time with the target T is the shortest (Figure 4.3). Thus, the total branch length separating the target from the template is $2T_1$ when the internal reference panel is used, and $2T_2$ when the external reference panel is used. Under our model, mutations that occur along the branches between the target haplotype and the template haplotype will be incorrectly imputed.

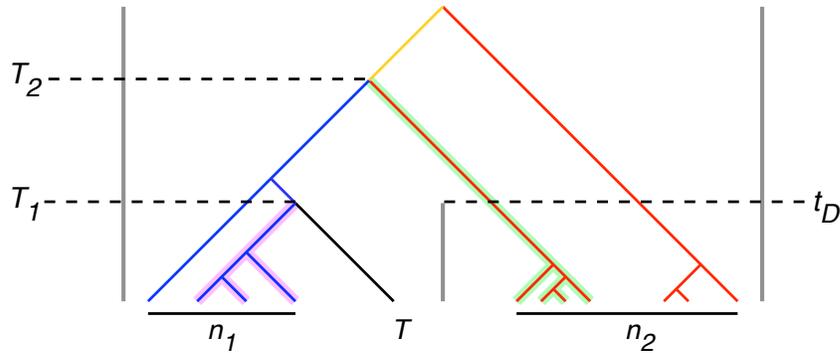


Figure 4.3: Coalescent times between the target T and the reference panels. T_1 indicates the time at which the target haplotype T first coalesces with a type 1 (blue) or type 1-2 lineage (yellow). We choose one of the descendant reference haplotypes from that coalescent event (highlighted in purple) to be the template from the internal reference panel. We assume when using the internal panel mutations that result in incorrectly imputed sites arise in direct proportion to $2T_1$, the coalescent time separating the target T and the templates from the internal panel. Similarly, T_2 is the time at which the target haplotype T first coalesces with a type 2 (red) or type 1-2 lineage (yellow) and $2T_2$ is the coalescent time between T and potential templates from the external reference panel (highlighted in green). We use $E[T_2 - T_1]$ to measure the expected difference in number of incorrectly imputed sites between using the external and internal reference panels.

We model mutation events using the infinite sites model [77] and assume the

number of mutations that occur along a branch of length t in the genealogy follows a Poisson distribution with mean $\mu\ell t$, where μ is the per-base per-generation mutation rate and ℓ is the length of the target haplotype in units of bases. Therefore, the expected number of sites incorrectly imputed when the internal reference panel is used is given by

$$E[S_1] = E[E[S_1|T_1]] = E[2\mu\ell T_1] = 2\mu\ell E[T_1]. \quad (4.13)$$

Similarly, the expected number of sites incorrectly imputed based on the external reference panel is

$$E[S_2] = 2\mu\ell E[T_2]. \quad (4.14)$$

It follows that the expected difference in number of sites incorrectly imputed between the external and internal reference panels is

$$E[S_2 - S_1] = 2\mu\ell E[T_2 - T_1] = 2\mu\ell \{E[T_2] - E[T_1]\} \quad (4.15)$$

so that, up to a constant, deriving $E[T_1]$ and $E[T_2]$ is sufficient to determine the expected difference.

4.3.2.1 Derivation of $E[T_1]$

To compute the expected waiting time $E[T_1]$ until T first coalesces with a lineage with descendants in the internal reference panel, we condition on the population in which lineage T first coalesces.

As in Section 4.3.1, let \mathcal{D} be the event that lineage T first coalesces in population 1 and let \mathcal{D}^c be the complement of event \mathcal{D} . Conditioning on event \mathcal{D} , we have

$$E[T_1] = E[T_1|\mathcal{D}]P(\mathcal{D}) + E[T_1|\mathcal{D}^c]P(\mathcal{D}^c). \quad (4.16)$$

Here, $P(\mathcal{D})$ and $P(\mathcal{D}^c)$ are obtained using Equations (4.5) and (4.4). We now compute the expectations $E[T_1|\mathcal{D}]$ and $E[T_1|\mathcal{D}^c]$.

To obtain the expected waiting time $E[T_1|\mathcal{D}]$ until lineage T first coalesces, given that it coalesces in population 1, we can integrate the conditional survival function

$S_{T_1|\mathcal{D}}(t)$ of T_1 given \mathcal{D} . The conditional survival function is defined by

$$\begin{aligned}
S_{T_1|\mathcal{D}}(t) &= P(T_1 \geq t|\mathcal{D}) \\
&= 1 - P(T_1 < t|\mathcal{D}) \\
&= 1 - P(T_1 < t, \mathcal{D})/P(\mathcal{D}) \\
&= 1 - P(T_1 < \min\{t, t_D\})/P(\mathcal{D}) \\
&= 1 - \frac{1}{P(\mathcal{D})} \sum_{i=2}^{n_1+1} P(T_1 < \min\{t, t_D\} | \mathcal{A}_{n_1+1}^i(\min\{t, t_D\})) P(\mathcal{A}_{n_1+1}^i(\min\{t, t_D\})) \\
&= 1 - \frac{1}{P(\mathcal{D})} \sum_{i=2}^{n_1+1} \left[1 - \frac{I_{n_1, i-1}}{I_{n_1+1, i}} \right] h_{n_1+1, i}(\min\{t, t_D\}; N_1) \\
&= 1 - \frac{1}{P(\mathcal{D})} \sum_{i=2}^{n_1+1} \left[1 - \frac{i(i-1)}{n_1(n_1+1)} \right] h_{n_1+1, i}(\min\{t, t_D\}; N_1). \tag{4.17}
\end{aligned}$$

In the sixth equality, the probability that lineage T does not coalesce when the sampled $n_1 + 1$ lineages coalesce to i lineages is given by $I_{n_1, i-1}/I_{n_1+1, i}$, by the same argument used to derive Equation (4.3). Therefore, the probability that lineage T does coalesce is given by $1 - I_{n_1, i-1}/I_{n_1+1, i}$.

To obtain the expectation $E[T_1|\mathcal{D}]$, we integrate $S_{T_1|\mathcal{D}}(t)$

$$\begin{aligned}
E[T_1|\mathcal{D}] &= \int_{t=0}^{t_D} S_{T_1|\mathcal{D}}(t) dt \\
&= \int_{t=0}^{t_D} \left[1 - \frac{1}{Pr(\mathcal{D})} \sum_{i=2}^{n_1+1} \left[1 - \frac{i(i-1)}{n_1(n_1+1)} \right] h_{n_1+1, i}(\min\{t, t_D\}; N_1) \right] dt. \tag{4.18}
\end{aligned}$$

The term $Pr(\mathcal{D})$ in Equation (4.18) is given by Equation (4.5). Although the integral in Equation (4.18) can be carried out analytically, we present the formula in its current form because numerical instabilities require that the integral be performed numerically using an asymptotic approximation to the term $h_{n_1+1, i}(\min\{t, t_D\})$. In this case, we use the asymptotic approximation of Griffiths [63]. Furthermore, it is easier to modify Equation (4.18) from the form given to account for exponential growth.

The quantity $E[T_1|\mathcal{D}^c]$ is the expected time until lineage T first coalesces in the ancestral population with a lineage with descendants in population 1, given that it does not coalesce in population 1. This expected time can be found by conditioning

on the number i_D of lineages of type 1 that remain at the divergence time:

$$\begin{aligned} E[T_1|\mathcal{D}^c] &= \sum_{i_D=1}^{n_1} E[T_1|i_D, \mathcal{D}^c] Pr(i_D|\mathcal{D}^c) \\ &= \sum_{i_D=1}^{n_1} \frac{4N_A}{i_D + 1} \frac{Pr(\mathcal{D}^c, i_D)}{Pr(\mathcal{D}^c)}. \end{aligned} \quad (4.19)$$

Here, we used the fact that $4N/(n + 1)$ is the expected time until a lineage coalesces with any of n other lineages in a population of diploid size N . $Pr(\mathcal{D}^c, i_D)$ and $Pr(\mathcal{D}^c)$ are found using equations (4.3) and (4.4), respectively. The quantity $E[T_1]$ is then obtained by plugging Equations (4.18), (4.19), (4.5), and (4.4) into Equation (4.16).

4.3.2.2 The expectation $E[T_2]$

The expected coalescence time $E[T_2]$ between T and the best template in the external reference panel is much simpler to derive. To compute $E[T_2]$, we simply condition on the number j_D of lineages of type-2 that remain at the divergence time.

$$\begin{aligned} E[T_2] &= \sum_{j_D=1}^{n_2} E[T_2|j_D] h_{n_2, j_D}(t_D) \\ &= \sum_{j_D=1}^{n_2} \frac{4N_A}{j_D + 1} h_{n_2, j_D}(t_D; N_2), \end{aligned} \quad (4.20)$$

where, as in Equation (4.19), we have used the fact that the expected waiting time until lineage T coalesces with j_D other lineages in the ancestral population is given by $4N_A/(j_D + 1)$. This completes the derivation of the expectations $E[T_1]$ and $E[T_2]$ in the case in which populations 1 and 2 are of constant size.

4.3.3 Derivations of probabilities and expectations under exponential growth

Let $N_1(t)$ and $N_2(t)$ be the sizes of populations 1 and 2 at time $t < t_D$. We model exponential growth in populations 1 and 2 between the present and the divergence by setting $N_1(t) = N_1(0)e^{-\alpha_1 t}$ and $N_2(t) = N_2(0)e^{-\alpha_2 t}$ (Figure 4.2). Thus far, our equations have been derived assuming constant population sizes. Changes in population size affect the time scale at which coalescent events within that population occur. Thus, to account for changing population sizes in the exponential growth model, we use a transformation method that scales coalescent time in a constant-sized popula-

tion to obtain coalescent rates that apply to an exponentially growing population [53]. The equations derived in the previous sections can then be applied to the exponential growth model simply by rescaling coalescent time.

Counting from time $t = 0$, let t denote coalescence time in a population of size $N(t) = N(0)e^{-\alpha t}$, $t \in [0, \infty)$. Let t' denote the transformed time in a population of constant size one, where the size of one is chosen for simplicity. Following [53], the transformation $g(t; N(0), \alpha)$ that converts time t in the growing population to time t' in the population of constant size one is given by

$$g(t; N(0), \alpha) = \int_0^t 1/N(z) dz = \begin{cases} \frac{e^{\alpha t} - 1}{N(0)\alpha}, & \text{if } \alpha \neq 0, \\ t, & \text{otherwise.} \end{cases} \quad (4.21)$$

All of the changes under exponential growth are accomplished by modifying the probability $h_{n,k}(t; N)$ that n lineages coalesce to k lineages in time t in a population of diploid size N (Equation 4.2). Under exponential growth, the distribution of the number of lineages remaining at time t_D in a population with size at time t given by $N(t) = N(0)e^{-\alpha t}$ is the same as the distribution of the number of lineages remaining at time $t'_D = g(t_D; N(0), \alpha)$ in a population of constant size 1. Thus, under growth rate α , Equation (4.2) becomes

$$h_{n,k}(t; N(0), \alpha) = \sum_{i=k}^n \frac{(2i-1)(-1)^{i-k} k_{(i-1)} n_{[i]}}{k!(i-k)!n_{(i)}} e^{-\binom{i}{2}g(t; N(0), \alpha)}. \quad (4.22)$$

By changing $h_{n,k}(t; N) \rightarrow h_{n,k}(t; N(0), \alpha)$ to accommodate growth, all of the equations in Sections 4.3.1 and 4.3.2 are easily reformulated to account for exponential growth. The modified equations and their dependencies are summarized in the appendix (section 4.6.2).

4.4 Results

4.4.1 Consistency of computations

We derived exact closed-form (Section 4.3.1.1) and recursive equations (Section 4.3.1.2) for computing the probabilities $P(C_1)$, $P(C_2)$ and $P(C_{12})$. Both the exact and recursive computations require the function $h_{n,k}(t; N)$ to be evaluated. The form given by Tavaré (Eq. 4.2) is numerically unstable for small t and large n . Griffiths showed that, as $t \rightarrow 0$ and $n \rightarrow \infty$, $h_{n,k}(t; N)$ converges to a normal distribution ([63], Thm. 2). Therefore, for $n < 40$ we use Eq. 4.2 for $h_{n,k}(t; N)$ and Griffiths'

normal approximation for $n \geq 40$.

Table 4.2 gives values of $P(C_1)$ computed using both the exact and recursive forms at two divergence times (t_D) and for different reference panel sizes (n_1 and n_2). For larger values of n_1 or n_2 , closed-form derivation of $P(C_1)$ becomes computationally unattainable. The probabilities are identical for the exact and recursive methods at parameter values where a comparison is possible. We also find that these analytic values agree with estimates based on coalescent simulations of our model. Since the exact, recursive and simulation computations agree, we employ the recursive form to allow large reference panel sizes to be considered.

n_1	n_2	$t_D = 0.01$			$t_D = 0.05$		
		Closed-form	Recursion	Simulation	Closed-form	Recursion	Simulation
1	1	0.3400	0.3400	0.3405	0.3658	0.3658	0.3659
1	5	0.1221	0.1221	0.1220	0.1648	0.1648	0.1650
1	10	0.0726	0.0726	0.0727	0.1191	0.1191	0.1190
5	5	0.4069	0.4069	0.4047	0.5083	0.5083	0.4996
5	10	0.2807	0.2807	0.2788	0.4164	0.4164	0.4094
10	10	0.4392	0.4392	0.4377	0.6051	0.6051	0.5993
50	50	—	0.6068	0.6055	—	0.8789	0.8770
50	100	—	0.5378	0.5369	—	0.8680	0.8666
100	100	—	0.7268	0.7262	—	0.9494	0.9498

Table 4.2: $P(C_1)$ computed analytically using closed form and recursive equations and estimated using coalescent simulations.

4.4.2 Constant-sized populations

We first present results for equal, constant-sized populations ($N_1 = N_2 = N_A$). For a range of reference panel sizes n_1 and n_2 and at several divergence times t_D (in units of $2N_A$ generation), we give the optimality probability for an internal reference panel, $P(C_1)$, and the expected difference in number of incorrectly imputed sites between the external and internal reference panels, $E[S_2 - S_1]$ (Figure 4.4). Positive values for $E[S_2 - S_1]$ indicate the additional number of sites that are incorrectly imputed, on average, by using the external reference panel instead of the internal panel. A negative value indicates that the external reference panel will result in fewer imputation errors and a value of zero indicates the same number of expected imputation errors using the internal and external panels. $E[S_2 - S_1]$ is reported in units of the scaled population mutation rate for the imputed region.

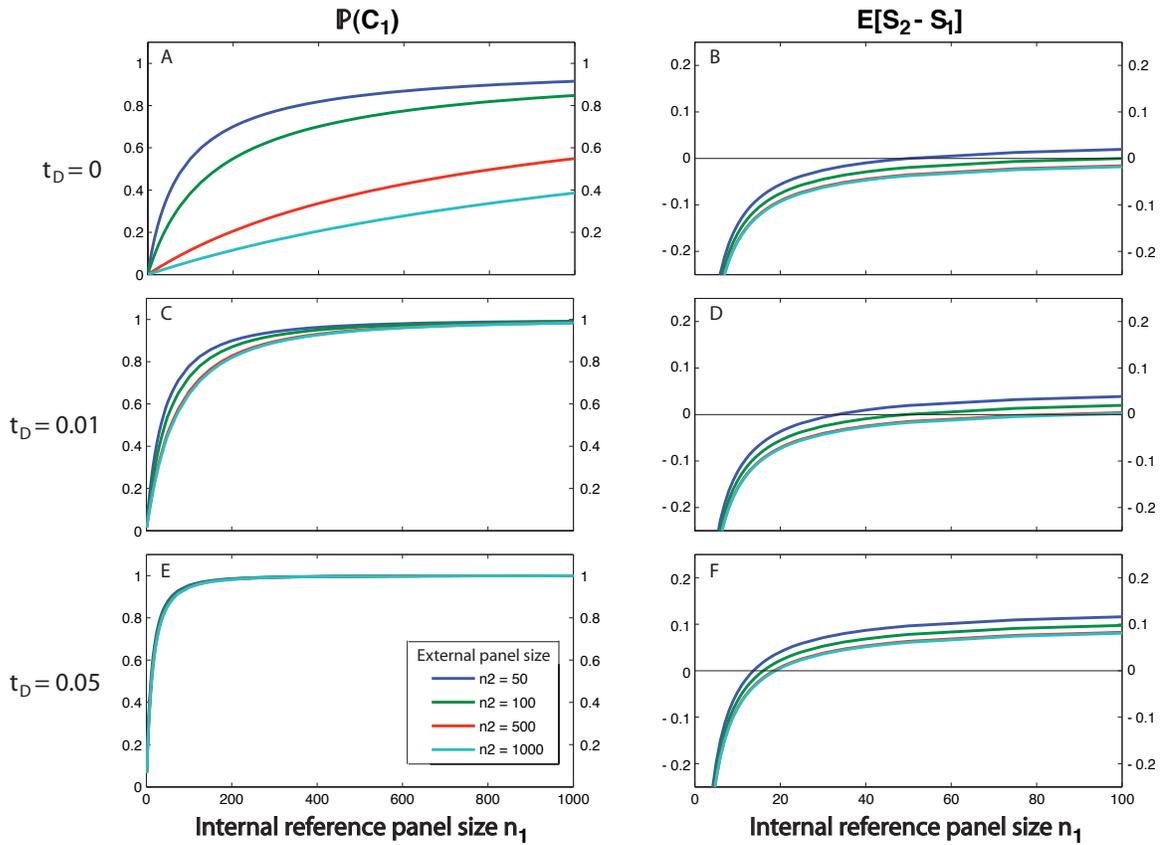


Figure 4.4: Imputation performance for the constant-size two population model. **(A,C,E)** The probability $P(C_1)$ that the internal reference panel is optimal, computed for divergence times of $t_D = 0, 0.01$, and 0.05 . **(B,D,F)** The expected difference in the number of sites incorrectly imputed $E[T_2 - T_1]$ when imputation is performed using the external reference panel rather than the internal reference panel, for divergence times of $t_D = 0, 0.01$, and 0.05 . $E[T_2 - T_1]$ is reported in units of the population scaled mutation for the imputed region.

A divergence time of $t_D = 0$ corresponds to no subdivision between the populations and is included to allow comparison with models that contain a non-zero divergence. In our model, the divergence time is approximately F_{st} , a measure of population differentiation. We present a divergence time of $t_D = 0.01$ to represent differentiation observed between two European populations and $t_D = 0.05$ to represent more distantly-related human populations [51].

For a divergence time of $t_D = 0$, the internal and external reference panels are both from the same population as the target and the only difference between them is size. Not surprisingly, given two reference panels from the same population, the larger of the two is more likely to be optimal (Fig 4.4 A). Larger reference panels are also likely to result in fewer sites incorrectly imputed (Fig 4.4 B). Increases in reference panel size initially result in large improvements for both $P(C_1)$ and $E[S_2 - S_1]$ compared to a competing reference panel of fixed size. However, we observe an asymptotic behavior indicating a diminishing return for increasing reference panel size beyond a certain point. This can be seen in Figure (4.4B) where, for a fixed reference panel size n_1 , increasing the competing reference panel size n_2 from 50 to 100 and then from 100 to 500 leads noticeable improvements in accuracy. However, increasing n_2 beyond 500 up to 1000 results in only a very modest improvement.

When the divergence time between populations 1 and 2 is nonzero, the internal reference panel is nearly always optimal for a large portion of the parameter space of n_1 and n_2 (Fig 4.4 B, C, E, F). For $t_D = 0.01$, $P(C_1)$ increases sharply with n_1 for all values of n_2 considered. We observe that even for $n_2 \gg n_1$, the probability that the internal reference panel is optimal can be quite large. For example, $P(C_1) = 0.647$ for an internal reference panel of $n_1 = 100$ haplotypes and an external reference panel ten times larger with $n_2 = 1000$ haplotypes. Once the internal reference panel contains $n_1 = 500$ haplotypes, it is nearly always optimal ($P(C_1) \approx 1$) regardless of the size of the external reference panel. Increasing divergence time t_D to 0.05 continues the trends observed for $t_D = 0.01$. Here, the internal reference panel is nearly always optimal for $n_1 \geq 200$. In fact, for sufficiently large n_1 , the probability of optimality for the internal reference panel is nearly independent of the size of the external reference panel.

With a nonzero divergence time, we also observe that internal reference panels that are smaller than a competing external reference panel can result in fewer expected imputation errors. The relative sizes of internal and external reference panels that provide similar accuracy depends on the divergence time (Fig 4.4 D, F). For $t_D = 0.01$, an internal reference panel of $n_1 = 50$ results in approximately the same number of

incorrectly imputed sites as an external panel of size $n_2 = 100$. As the internal reference panel becomes larger, it takes substantially more haplotypes in the external reference panel to achieve similar accuracy. For example, 1000 haplotypes are required in an external panel to provide the same accuracy as an internal reference panel of only $n_1 = 80$ haplotypes. When the divergence time increases to $t_D = 0.05$, we observe that an internal reference panel as small as $n_1 = 20$ haplotypes can provide better imputation accuracy than much larger external reference panels. Further increases in the size of the internal reference panel continue to improve accuracy. However after approximately $n_1 = 60$ haplotypes, the increase in accuracy for additional internal reference haplotypes begins to diminish.

4.4.3 Exponentially growing populations

We next computed $P(C_1)$ and $E[S_2 - S_1]$ assuming exponential growth in populations 1 and 2 beginning at the divergence and continuing to the present. Here we assume that both populations have size N_A at the divergence $t = t_D$ and size $100N_A$ in the present time $t = 0$. For the exponential growth model, we present results for divergence times of $t_D = 0.01, 0.05$ and 0.1 (in units of $2N_A$ generation). We include lines for both the exponential growth and constant-sized models in each plot (Fig 4.5) to allow comparison of results between the two models.

At each divergence time, compared to the constant size model, exponential growth reduces the probability $P(C_1)$ that the internal reference panel is optimal (Fig 4.5 A,C,E). The difference in expected number of incorrectly imputed sites $E[S_2 - S_1]$ also decreases under the exponential growth model (Fig 4.5 B,D,F). The reduction is greatest for the smallest divergence time $t_D = 0.01$, for which the large external reference panels ($n_2 = 500, 1000$) are either more optimal or equally appropriate versus internal panels of $n_1 = 300$ or less haplotypes. Also for $t_D = 0.01$, $E[S_2 - S_1]$ is negative for these large external reference panels compared to internal reference panels of up to $n_1 = 225$ haplotypes. This indicates that for exponentially growing populations separated by small divergence times, there is an advantage to using the large external reference panels compared to smaller and even moderately sized internal panels.

For a divergence time of $t_D = 0.05$, exponential growth still reduces $P(C_1)$ compared to the constant-size model but here a large external panel is only optimal over smaller internal reference panels ($n_1 \leq 50$). Once an internal reference panel contains 200 haplotypes it is still optimal more than 80% of the time compared to large external panels (Fig 4.5 C). In terms of $E[S_2 - S_1]$, under the exponential growth

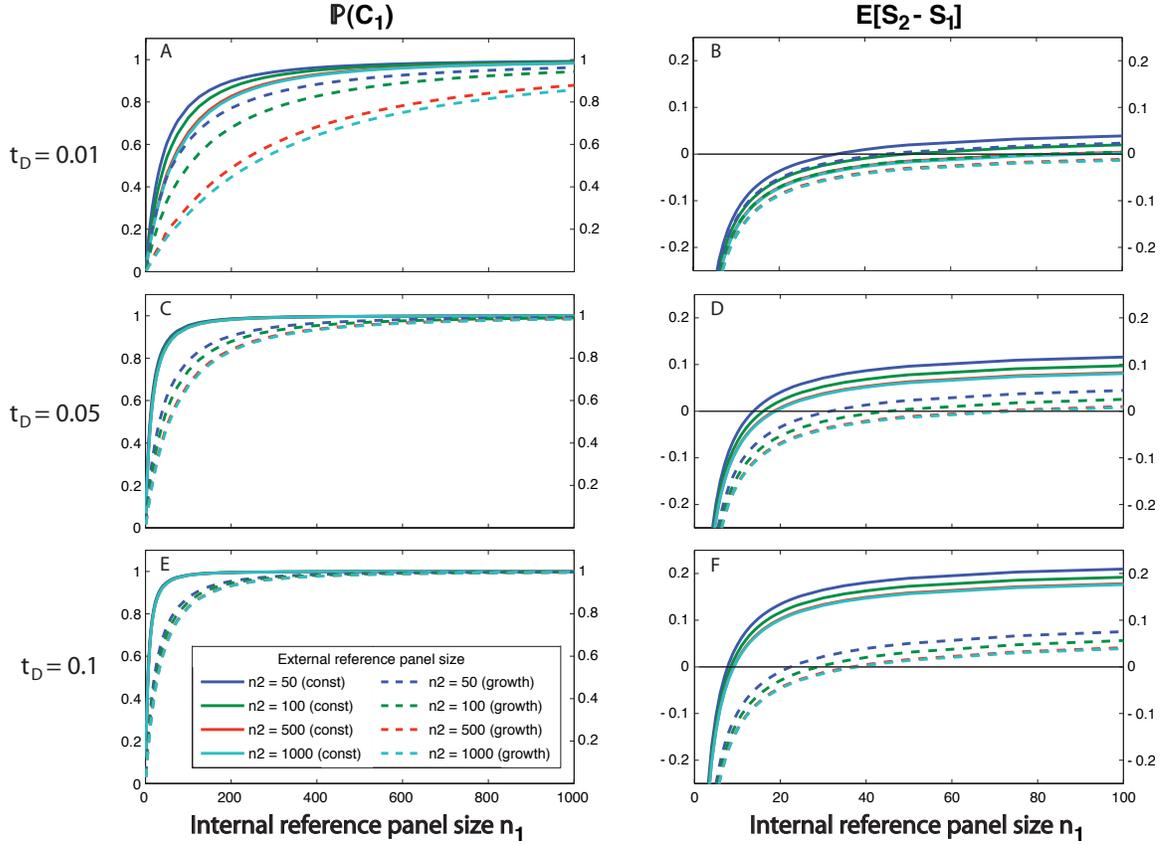


Figure 4.5: Imputation performance for the exponential growth two-population model. **(A,C,E)** The probability $P(C_1)$ that the internal reference panel is optimal, computed for divergence times of $t_D = 0.01, 0.05$ and 0.1 . Values for the exponential growth model are plotted with dashed lines and, for comparison, the equivalent values for a constant-size model are shown with solid lines. **(B,D,F)** The expected difference $E[T_2 - T_1]$ in the number of sites incorrectly imputed when imputation is performed using the external reference panel rather than the internal reference panel, for divergence times of $t_D = 0, 0.01$, and 0.05 . $E[T_2 - T_1]$ is reported in units of the population scaled mutation rate for the imputed region.

model at $t_D = 0.05$, an internal panel of $n_1 = 70$ haplotypes gives the same expected number of incorrectly imputed sites as an external panel with $n_2 = 500$ or 1000 haplotypes (Fig 4.5 D). The results represent a dramatic change from the constant-size model in which, for equivalently sized panels, the internal panel offers a considerable improvement in accuracy over the large external panels.

The trend of decreasing both $P(C_1)$ and $E[S_2 - S_1]$ continues for $t_D = 0.1$ with even larger reductions observed (Fig 4.5 E, F). However, for a divergence time this large, the values for $P(C_1)$ and $E[S_2 - S_1]$ under the constant model favor internal panels to such an extent that relatively small internal panels ($n_1 = 40$) provide better imputation performance than a large external panel.

4.5 Discussion

We have introduced a theoretic model of genotype imputation accuracy, employing the coalescent framework to model the ancestry of reference and imputation target haplotypes, taking the reference haplotype with the most similar genealogical history to the target to be the template for imputation and using the expected coalescence time between the target and the template haplotype to predict the expected number of incorrectly imputed sites in the target. Framing imputation in a coalescent model has two major benefits. First, the coalescent allows analytic equations to be derived for imputation accuracy, providing a computationally fast method for predictive imputation analysis across a continuous range of study design variables. Second, the coalescent allows complex modeling of the population histories, so that virtually any relationship between the reference and target populations can be considered.

Here, we used the model to determine the effects of population subdivision on imputation accuracy. In particular, we compared the performance of internal reference panels drawn from the same source population as the target to external reference panels drawn from distinct, yet closely related populations. In the specific model we considered, the populations containing the internal and external panels diverged from an ancestral population in the past, after which, no migration has occurred between the two populations. The model predicts that even when an internal reference panel is considerably smaller than an external reference panel, the internal panel is nearly always optimal in the sense that it contains the haplotype with the closest genealogical history to the target. As divergence time between the populations containing the reference panels increases, it becomes nearly impossible for even very large external reference panels (1000 haplotypes) to be optimal compared to an internal reference

panel. This result was observed both for the constant size and exponential growth population models, with exponential growth providing a slight improvement in the probability of optimality for external panels when compared to the constant size model.

Although an internal reference panel is nearly always optimal, the relative improvement in imputation accuracy for using an internal panel varies according to divergence time between the reference panel populations. For constant-sized populations, we observed that very large external reference panels can provide the same accuracy as a modestly sized internal reference panel for small divergence times. As divergence time increases, a small internal reference panel will provide more accurate imputation than a large external panel. However, if the populations experience exponential growth following the divergence, the magnitude of the expected number of additional errors is significantly reduced when compared to constant sized populations. Thus, exponential growth attenuates the effect of the divergence time, improving the relative performance of an external reference panel in terms of both optimality probability and imputation accuracy.

The results from our model have implications for imputation strategies in future population-based genetic association studies. Currently, large publicly-available datasets such as HapMap and the 1000 Genomes Project are commonly used for imputation reference panels. However, next generation sequencing technology will allow investigators to create custom-made internal reference panels from the same source population as their study sample. Our results suggest that such a custom-made reference panel will nearly always improve imputation accuracy even if it is considerably smaller than an existing reference panel based on the 1000 Genomes Project. The quantitative improvement, however, may be small if the existing panel is very large and not too distantly related to the imputation targets in the study sample. The advantage of an internal reference panel will be in its ability to accurately impute rare population-specific variants that exist in the study sample but do not exist in a reference panel from an external population.

We have assumed no migration between the present-day populations in our model. Relaxing this assumption will likely lead to changes in our results. Notably, migration allows the target haplotype to coalesce with a lineage ancestral to the external reference panel more recently than the divergence time. This coalescence can occur if either the target migrates to the population containing the external reference panel or if an ancestor of an external reference haplotype migrates to the population containing the target. Therefore, including migration in the model is likely to reduce

the expected coalescence time between the target and an ancestor of the external reference panel, improving the probability of optimality and accuracy for an external reference panel. The extent of the improvement is unclear and requires further investigation.

We note that our analysis for expected number of incorrectly imputed sites assumes perfect information for the haplotypes in the reference panel, including both perfect phasing and that all sites are correctly called. Perfect phasing would likely require information from parental haplotypes and is not realistic for most studies. The assumption that all sites are correctly called is also not realistic for most sequencing experiments, especially for low-coverage sequencing. In practice, these assumptions may not be met, and differences in the data quality between competing reference panels can lead to imputation results that differ from our predictions. If the source of error can be specified by a model, it can potentially be included in our analysis. For example, the probability that a site in a reference haplotype is correctly called in a sequencing experiment is dependent upon sequencing depth, allele frequency of the variant and the specific calling algorithm (personal communication, Shyam Gopalakrishnan). Estimates of these probabilities based on empirical data or simulations can be incorporated into our equations to compute the expected number of sites incorrectly imputed conditional on the probability distribution for sites to be called in the reference panel.

Finally, to mimic computational imputation algorithms, we used the rule that the reference haplotype with minimal coalescent time between itself and the target serves as the imputation template. Thus, the only source of imputation error in our model is in mutations that occur after the coalescence of the target and the template. In reality this is likely a best-case scenario since additional sources of error may result from the imputation software. For example, our model assumes that the entire length of the target haplotype is imputed using the same reference haplotype. However, due to recombination events in the past, a real target haplotype is likely to be composed of multiple segments, each with a unique reference haplotype that provides the best template for imputation. Our model, therefore, implicitly assumes that an imputation algorithm will always correctly jump between reference haplotypes when imputing a target. It is not clear how realistic this assumption is in practice, and provides an additional source of imputation error that is not treated in our model.

Genotype imputation is a valuable tool in genetic studies of complex disease. Optimizing imputation accuracy is important for obtaining valid analysis results based on the imputed data. In this chapter, we have introduced a coalescent model of

genotype imputation and showed that it can be used to predict accuracy for a range of imputation study designs. We have demonstrated that the model can be used to compare the performance of competing reference panels, and we anticipate that it will be useful in optimizing imputation performance in future genetic studies.

4.6 Appendix

4.6.1 The quantity $N(i_D, j_D \rightarrow i_C, j_C, k_C)$

Here we derive the number of ways $N(i_D, j_D \rightarrow i_C, j_C, k_C)$ in which i_D lineages of type 1 and j_D lineages of type 2 can coalesce down to i_C , j_C , and k_C lineages of types 1, 2, and 1-2 at the moment when lineage T first coalesces. This quantity is used to obtain the closed-forms of the probabilities $Pr(C_1)$, $Pr(C_2)$, and $Pr(C_{12})$ (Equations 4.1, 4.8, and 4.9).

To derive an expression for $N(i_D, j_D \rightarrow i_C, j_C, k_C)$, note that if k_C lineages of type 1-2 remain at the time immediately before the first coalescence involving lineage T , then at least k_C lineages of type 1, and at least k_C lineages of type 2 must combine together to produce these lineages. Let i_D^* and j_D^* be the numbers of lineages of type 1 and type 2, respectively, that combine to create the k_C lineages of type 1-2 (figure 4.6). Specifically, let $i_{D_r}^*$ lineages of type 1 and $j_{D_r}^*$ lineages of type 2 combine to make the r th lineage of type 1-2. The possible values of $i_{D_1}^*, \dots, i_{D_{k_C}}^*$ are given by all possible partitions of i_D^* objects into k_C nonempty subsets. Similarly the possible values of the $j_{D_1}^*, \dots, j_{D_{k_C}}^*$ are given by all possible partitions of j_D^* objects into k_C nonempty subsets.

Let $P(n, k)$ denote the number of partitions of an integer n into k positive integers. Let $\pi^q(n, k) = (\pi_1^q(n, k), \dots, \pi_k^q(n, k))$ denote the q th partition of this kind with $\pi_r^q(n, k)$ denoting the r th part in the partition. We can permute the k parts of the q th partition in $k!$ ways. Denote the z th permutation of partition q by $\pi^{(q,z)}(n, k) = (\pi_1^{(q,z)}(n, k), \dots, \pi_k^{(q,z)}(n, k))$. For simplicity of notation, denote the number of labeled histories among a set of n lineages by $F(n) \equiv I_{n,1}$. Then the quantity $N(i_D, j_D \rightarrow i_C, k_C, j_C)$ is given by

$$\begin{aligned}
 & N(i_D, j_D \rightarrow i_C, k_C, j_C) \\
 &= \sum_{i_D^*=k_C}^{i_D-i_C} \sum_{j_D^*=k_C}^{j_D-j_C} \binom{i_D}{i_D^*} \binom{j_D}{j_D^*} \sum_{\eta=1}^{P(i_D^*, k_C)} \sum_{\gamma=1}^{P(j_D^*, k_C)} \alpha(i_D^*, k_C, \eta) \alpha(j_D^*, k_C, \gamma) R(i_D^*, j_D^*, k_C, \eta, \gamma) \\
 & \quad \times I_{i_D-i_D^*, i_C} I_{j_D-j_D^*, j_C} \binom{i_D + j_D - (i_C + k_C + j_C)}{i_D - i_D^*, i_D^* + j_D^* - k_C, j_D - j_D^*}. \tag{4.23}
 \end{aligned}$$

In Equation (4.23), the quantity $\alpha(n, k, q)$ is the number of ways to form the q th partition $\pi^q(n, k) = (\pi_1^q(n, k), \dots, \pi_k^q(n, k))$ of size k nonempty parts of n objects. To obtain an expression for $\alpha(n, k, q)$, define $a(\varphi; \pi^q(n, k))$ to be the number of parts of

the partition $\pi^q(n, k)$ that are of size φ . Then $\alpha(n, k, q)$ is given by

$$\alpha(n, k, q) = \frac{\binom{n}{\pi_1^q(n, k), \dots, \pi_k^q(n, k)}}{\prod_{\varphi=1}^k a(\varphi; \pi^q(n, k))!}, \quad (4.24)$$

where $\binom{n}{\pi_1^q(n, k), \dots, \pi_k^q(n, k)}$ is the number of ways to choose the elements in each part, and $\prod_{\varphi=1}^k a(\varphi; \pi^q(n, k))!$ is the factor by which we overcount the ways to choose the elements of the parts due to the fact that, for each way of distributing k elements into r parts of equal size, there are $r! - 1$ other ways to distribute the elements in which the same elements are grouped together.

The quantity $R(i, j, k, \eta, \gamma)$ in Equation (4.23) is the number of labeled histories for the $i_D^* + j_D^*$ lineages that ultimately coalesce to form the k_C lineages of type 1-2. For given values of i_D^* and j_D^* , consider a particular partition of the i_D^* lineages into k_C nonempty parts, and a particular partition of the j_D^* lineages into k_C nonempty parts. Each one of the k_C lineages of type 1-2 is made by combining one part from the partition of the i_D^* lineages with a part from the partition of the j_D^* lineages. To find all possible ways to pair up parts, we fix the indices of the parts of the j_D^* lineages and we permute the parts of the i_D^* lineages with respect to them. There are $k_C!$ ways to pair up the parts, and these ways are indexed by z . For the z th way of permuting the parts, the lineages in part $\pi_r^{(\eta, z)}(i_D^*, k_C)$ combine with the lineages in part $\pi_r^\gamma(j_D^*, k_C)$ to produce the r th lineage of type 1-2.

The r th pair of parts of lineages undergoes $\pi_r^{(\eta, z)}(i_D^*, k_C) + \pi_r^\gamma(j_D^*, k_C) - 1$ coalescent events on its way down to a single lineage. Thus, there are

$$\begin{aligned} & W(i_D^*, j_D^*, k_C, \pi^{(\eta, z)}, \pi^\gamma) \\ & \equiv \binom{i_D^* + j_D^* - k_C}{\pi_1^{(\eta, z)}(i_D^*, k_C) + \pi_1^\gamma(j_D^*, k_C) - 1, \dots, \pi_{k_C}^{(\eta, z)}(i_D^*, k_C) + \pi_{k_C}^\gamma(j_D^*, k_C) - 1} \end{aligned} \quad (4.25)$$

possible ways to order the coalescent events among all pairs of partitions.

Since there are $F(\pi_r^{(\eta, z)}(i_D^*, k_C) + \pi_r^\gamma(j_D^*, k_C))$ labeled histories for the r th pair of parts as they coalesce down to form the r th lineage of type 1-2, there are

$$\begin{aligned} & \prod_{r=1}^{k_C} F(\pi_r^{(\eta, z)}(i_D^*, k_C) + \pi_r^\gamma(j_D^*, k_C)) \times \\ & \binom{i_D^* + j_D^* - k_C}{\pi_1^{(\eta, z)}(i_D^*, k_C) + \pi_1^\gamma(j_D^*, k_C) - 1, \dots, \pi_{k_C}^{(\eta, z)}(i_D^*, k_C) + \pi_{k_C}^\gamma(j_D^*, k_C) - 1} \end{aligned}$$

labeled histories for all of the i_D^* and j_D^* lineages when paired in this way. Finally,

we sum over all $k_C!$ possible ways to permute the partitions of the i_D^* lineages with respect to the partitions of the j_D^* lineages. This gives

$$R(i_D^*, j_D^*, k_C, \eta, \gamma) = \sum_{z=1}^{k_C!} W(i_D^*, j_D^*, k_C, \pi^{(\eta, z)}, \pi^\gamma) \prod_{r=1}^{k_c} F(\pi_r^{(\eta, z)}(i_D^*, k_C) + \pi_r^\gamma(j_D^*, k_C)). \quad (4.26)$$

We have separately considered three parts of the labeled history of the lineages in the ancestral population: 1) the labeled history of the $i_D - i_D^*$ lineages that coalesce to form lineages of type 1, 2) the labeled history of the $i_D^* + j_D^*$ lineages that coalesce to form lineages of type 1-2, and 3) the labeled history of the $j_D - j_D^*$ lineages that coalesce to form lineages of type 2. To integrate these histories into one full history for all lineages, we must only consider how the coalescence times in each of these histories relate to the coalescence times in the other histories. The final quantity in Equation (4.23) is the number of ways to interweave the coalescent events in these labeled histories.

There are $i_D - i_D^* - i_C$ coalescence events among the lineages that are ultimately of type 1, there are $j_D - j_D^* - j_C$ coalescence events among the lineages that are ultimately of type 2, and there are $i_D^* + j_D^* - k_C$ coalescence events among the lineages that are ultimately of type 1-2. The number of ways to interweave the coalescence times for each of these histories is equal to the number of ways to choose which of the $i_D + j_D - i_C - j_C - k_C$ total coalescent events correspond to events within each of these different histories. This is given by

$$\binom{i_D + j_D - (i_C + k_C + j_C)}{i_D - i_D^*, i_D^* + j_D^* - k_C, j_D - j_D^*}. \quad (4.27)$$

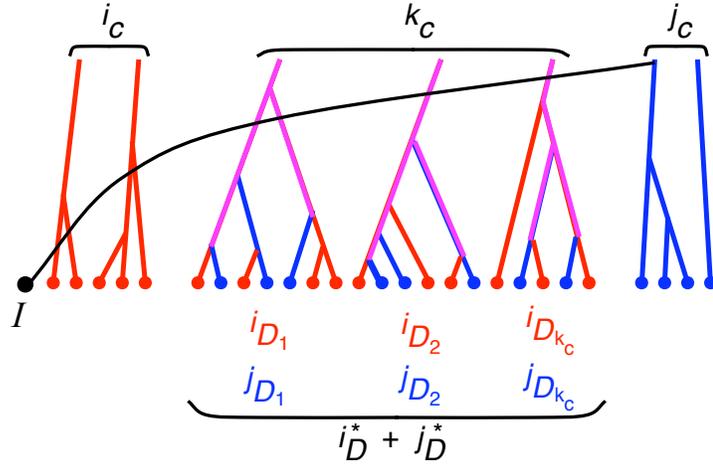


Figure 4.6: An illustration of $N(i_D, j_D \rightarrow i_C, j_C, k_C)$. One possible way in which $i_D = 15$ lineages of type 1, $j_D = 12$ lineages of type 2, and one lineage I can coalesce down to $i_C = 2$ lineages of type 1, $k_C = 3$ of type 1-2, and $j_C = 2$ lineages of type 2, with the final coalescence occurring between I and a lineage of type 2. Lineages from G_1 are in red, lineages from G_2 are in blue, and lineages with descendants in both G_1 and G_2 are in purple. Here $i_D^* = 10$ is the number of lineages of type 1 that coalesce with $j_D^* = 8$ lineages of type 2 to produce the $k_C = 3$ lineages of type 1-2. $i_{D_1}^* = 3$ is the number of lineages of type 1 that combine with $j_{D_1}^* = 4$ lineages of type 2 to create the first lineage of type 1-2. In general, $i_{D_r}^*$ is the number of lineages of type 1 that coalesce with $j_{D_r}^*$ lineages of type 2 to produce the r th lineage of type 1-2.

4.6.2 Table of equations

	Quantity	Dependencies
1	$g(t; N(0), \alpha) = \begin{cases} \frac{e^{\alpha t} - 1}{N(0)\alpha}, & \text{if } \alpha \neq 0, \\ t, & \text{otherwise.} \end{cases}$	
2	$h_{n,k}(t; N(0), \alpha) = \sum_{i=k}^n \frac{(2i-1)(-1)^{i-k} k \binom{i-1}{n} n! e^{-\binom{i}{2} g(t; N(0), \alpha)}}{k!(i-k)! n \binom{i}{n}}$	
3	$\mathbb{P}(\mathcal{D}^c, i_D) = \frac{i_D(i_D+1)}{n_1(n_1+1)} h_{n_1+1, i_D+1}(t_D; N_1(0), \alpha_1)$	2
4	$\mathbb{P}(\mathcal{D}^c) = \sum_{i_D=1}^{n_1} \mathbb{P}(\mathcal{D}^c, i_D)$	3
5	$\mathbb{P}(\mathcal{D}) = 1 - \sum_{i_D=1}^{n_1} \mathbb{P}(\mathcal{D}^c, i_D)$	3
6	$\begin{aligned} \tilde{\mathbb{P}}(C_1 i, j, k) &= \frac{2i}{m(m-1)} + \frac{i(i-1)+2ik}{m(m-1)} \tilde{P}(C_1 i-1, k, j) \\ &\quad + \frac{2ij}{m(m-1)} \tilde{P}(C_1 i-1, k+1, j-1) \\ &\quad + \frac{j(j-1)+2jk}{m(m-1)} \tilde{P}(C_1 i, k, j-1) + \frac{k(k-1)}{m(m-1)} \tilde{P}(C_1 i, k-1, j) \end{aligned}$	
7	$\begin{aligned} \tilde{P}(C_2 i, k, j) &= \frac{2j}{m(m-1)} + \frac{i(i-1)+2ik}{m(m-1)} \tilde{P}(C_2 i-1, k, j) \\ &\quad + \frac{2ij}{m(m-1)} \tilde{P}(C_2 i-1, k+1, j-1) \\ &\quad + \frac{j(j-1)+2jk}{m(m-1)} \tilde{P}(C_2 i, k, j-1) + \frac{k(k-1)}{m(m-1)} \tilde{P}(C_2 i, k-1, j) \end{aligned}$	
8	$\begin{aligned} \tilde{P}(C_{12} i, k, j) &= \frac{2k}{m(m-1)} + \frac{i(i-1)+2ik}{m(m-1)} \tilde{P}(C_{12} i-1, k, j) \\ &\quad + \frac{2ij}{m(m-1)} \tilde{P}(C_{12} i-1, k+1, j-1) \\ &\quad + \frac{j(j-1)+2jk}{m(m-1)} \tilde{P}(C_{12} i, k, j-1) + \frac{k(k-1)}{m(m-1)} \tilde{P}(C_{12} i, k-1, j) \end{aligned}$	
9	$\mathbb{P}(C_1) = \mathbb{P}(\mathcal{D}) + \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} \mathbb{P}(C_1 \mathcal{D}^c, i_D, j_D) \mathbb{P}(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2(0), \alpha_2)$	6, 5, 3, 2
10	$\mathbb{P}(C_2) = \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} \mathbb{P}(C_2 \mathcal{D}^c, i_D, j_D) \mathbb{P}(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2(0), \alpha_2)$	7, 3, 2
11	$\mathbb{P}(C_{12}) = \sum_{i_D=1}^{n_1} \sum_{j_D=1}^{n_2} \mathbb{P}(C_{12} \mathcal{D}^c, i_D, j_D) \mathbb{P}(\mathcal{D}^c, i_D) h_{n_2, j_D}(t_D; N_2(0), \alpha_2)$	8, 3, 2
12	$S_{T_1 \mathcal{D}}(t) = 1 - \frac{1}{Pr(\mathcal{D})} \sum_{i=2}^{n_1+1} \left[1 - \frac{i(i-1)}{n_1(n_1+1)} \right] h_{n_1+1, i}(\min\{t, t_D\}; N_1(0), \alpha_1)$	5, 2
13	$E[T_1 \mathcal{D}] = \int_{t=0}^{t_D} \left[1 - \frac{1}{Pr(\mathcal{D})} \sum_{i=2}^{n_1+1} \left[1 - \frac{i(i-1)}{n_1(n_1+1)} \right] h_{n_1+1, i}(\min\{t, t_D\}; N_1(0), \alpha_1) \right] dt.$	5, 2
14	$E[T_1 \mathcal{D}^c] = \sum_{i_D=1}^{n_1} \frac{4N_A}{i_D+1} \frac{Pr(\mathcal{D}^c, i_D)}{Pr(\mathcal{D}^c)}$	4, 3
15	$E[T_2] = \sum_{j_D=1}^{n_2} \frac{4N_A}{j_D+1} h_{n_2, j_D}(t_D; N_2(0), \alpha_2)$	2

Table 4.3: Summary of all derived equations and their dependencies. Formulas for the case in which populations 1 and/or 2 are of constant size are obtained by setting α_1 and/or α_2 equal to 0.

CHAPTER V

Discussion

This dissertation has addressed several current problems in the field of human statistical genetics. The projects were diverse, but had a unifying theme of developing statistical methods to complement the Genome-Wide Association Study (GWAS) and further our understanding of complex disease. In this section we discuss the contributions of this dissertation and we address future questions and challenges.

The first project dealt with follow-up testing to characterize the true genotype-phenotype relationship for risk variants implicated by a GWAS. We expected and observed a bias in genetic and interaction effects estimated in a naive fashion due to the Winner's Curse phenomenon. Less expected and more problematic was the observation that a follow-up hypothesis test for interaction had an elevated type I error when the initial GWAS test did not model a potential interaction. Thus, while the Winner's Curse has previously been blamed for replication failure of true risk variants, here we show that it can also cause false positives. Gene-environment interactions are already notoriously difficult to identify and replicate and our results reinforce the need for care in the application and interpretation of tests for interaction.

We introduced the partial likelihood Markov Chain Monte Carlo (MCMC) algorithm to obtain bias-reduced estimates of gene-environment effect parameters from a likelihood function that conditions on the initial significant GWAS test. The partial likelihood MCMC uses a simulation-rejection step to account for the intractable portion of the conditional likelihood function while the tractable portion is utilized in a Metropolis-Hastings-like step. The algorithm has broader applications in statistical genetics such as gene-gene interactions, and below we describe how it might be applied to parameter estimation from burden-style tests like the Cumulative Minor Allele Test (CMAT). The generality of the algorithm allows application to any scenario likely to suffer from a Winner's Curse effect, that is, parameters estimated from a dataset that had low power to detect an association. The only requirements for the partial

likelihood MCMC are knowledge of the initial test that was performed and the unconditional likelihood function from the data are derived. Beyond the ascertainment-correction problems considered here, the partial likelihood MCMC may find further applications where inference is required on a likelihood function that consists of both tractable and intractable portions.

The second project was an investigation of statistical methods and study designs to identify rare genetic variants that contribute to disease risk. By looking at the cumulative effect of multiple rare variants, our CMAT burden statistic is more powerful than individually applying GWAS-style single marker tests. More importantly, we used the CMAT to investigate several practical challenges in the application of burden tests to real sequence data, including neutral variation dramatically outnumbering risk variants and datasets plagued by population stratification. We also showed that genotype imputation, a vital tool for current GWAS, can be applied in a similar manner to increase power for burden tests of rare variants.

The explosion of data from next-generation sequencing and the unprecedented access it provides to rare variation will provide a range of new research problems in the field of statistical genetics, several following directly from the projects considered here. The development of novel burden tests is a very active area of research; already our CMAT has been improved upon by methods that allow combinations of risk and protective variants and continuous outcomes and covariates [50, 78]. However, the application of these methods to real data has been limited thus far and performance outside of simulated data, often with idealized conditions, is not yet known. Predictably, many of the same challenges that complicate GWAS will need to be addressed for burden tests. Also, new unforeseen problems are likely to develop and require solutions.

We touched on the issue of population stratification in burden tests in the second chapter. Although the example we provided was somewhat extreme, it demonstrated the significant problem that stratification can cause in burden tests. We showed that stratification can be controlled for if population of origin is known, but in practice, this will need to be inferred from the data. It is not immediately clear that existing methods to control for stratification in single marker GWAS [15, 57] will be sufficient for burden tests. Stratification in single marker tests is the result of differences in allele frequencies at individual markers between populations. Existing correction methods use the single-site allele frequency differences from across the genome to assess stratification in the dataset and apply a global correction to all tests. In contrast, stratification in burden tests is caused by differential sampling of

cases and controls from populations containing different proportions of rare variation. Already, data from the 1000 Genomes Project suggests substantial rare variant diversity between populations, highlighting the serious risk for stratification in burden tests [18]. Patterns of rare variation can differ between populations for a variety of reasons unrelated to the phenotype of interest. Differences in population history including expansion, bottlenecks and migration can influence rare variant diversity on a genome-wide level. Alternatively, selective pressures that regulate rare variant distributions can differ between populations and act in a more localized gene-based level. Further, the cumulative nature of burden tests implies that subtle differences at single sites can be magnified when pooled together. The combination of these factors may render correction methods based on genome-wide patterns of single site differences insufficient for burden tests. Moreover, existing methods rely on extracting information on diversity from common variants and may not adequately reflect diversity at the rare variant level. New correction methods designed specifically for burden tests will be required to account for the multiple sources of confounding described here. It is possible that the burden statistic for each gene will need to be individually corrected using information from both genome-wide and gene-level rare variant distributions.

Another foreseeable challenge for rare variant association studies will be determining the effect size of genes identified by burden tests. Whole genome sequencing will allow burden tests for the more than 20,000 genes in the human genome and correction for multiple testing will open the door for a Winner's Curse effect on estimates of effect sizes for rare variants. A natural effect quantity to report is the risk explained by total variation at the locus. Extending existing analytical methods to estimate locus-effect may be difficult due to the complicated nature in which many burden tests combine information across sites. Further, burden test statistics can be composed of causative, protective and neutral variants creating a complex mix of genetic effects. Our ascertainment-correction method described in the first chapter may be a useful tool to handle the challenge of estimating effect size from burden tests. The partial likelihood MCMC does not require that the effect estimate correspond to the burden test statistic used to declare association. This is important because the statistic used to summarize rare variant burden will almost certainly not be indicative of effect size.

Finally, we examined the role that new sequencing technology can play in improving the existing tool of imputation. We observed that custom reference panels made directly from the study population of interest will be valuable for improving

imputation, especially for rare variants. However, we also noted that under certain conditions the large public reference panels will provide similar and even improved accuracy compared to that of smaller custom panels. In the chapter, we did not explicitly consider imputation in the setting of disease-gene mapping and perhaps underrated the major advantage custom reference panels can provide. We briefly mentioned in the discussion of that chapter that custom reference panels allow variants unique to the study sample to be observed and imputed. In our coalescent model these were population-specific variants. However, if the study sample contains individuals ascertained for a specific phenotype, the unique variants in a custom panel may include novel risk variants that could not be directly analyzed if the study sample were imputed using a public reference panel. This is particularly important because novel rare variants may not be tagged well enough by the public panels to enable powerful linkage disequilibrium mapping. Thus, the benefit of the custom reference panel may go beyond simply improving imputation accuracy to also allowing analysis of novel risk variants that would otherwise be missed.

Several interesting questions can be posed for the design of custom reference panels in a case-control setting, including strategies for choosing which individuals from the study sample to sequence for the reference panel. For example, in our CMAT simulations we assumed custom reference panels of equal numbers of cases and controls to match the balanced design of the larger sample. An alternative strategy is to sequence a larger fraction of cases to enrich the custom reference panel for risk variants and potentially increase association testing power. However, a potential pitfall of this strategy is that variants appearing in the reference panel at dramatically different frequencies than in the population are prone to severe imputation error (unpublished data) that could bias association test results. A more thorough analysis of reference panel design in a gene-mapping setting would help to reveal optimal strategies for designing imputation-based analyses.

Despite the best efforts of this dissertation, it is clear that numerous challenges remain. The tools developed here are only a small step toward the complete understanding of the genetics of complex disease, but we anticipate that this work will be built upon and lead to further advances in the field of human statistical genetics.

BIBLIOGRAPHY

- [1] H Akaike. Information theory and the maximum likelihood principle. *2nd International Symposium in Information Theory*, 1973.
- [2] DM Altshuler, RA Gibbs, L Peltonen, DM Altshuler, RA Gibbs, E Dermitzakis, SF Schaffner, and F Yu. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010.
- [3] S Asthana, WS Noble, G Kryukov, CE Grant, S Sunyaev, and JA Stamatoyannopoulos. Widely distributed noncoding purifying selection in the human genome. *Proceedings of the National Academy of Sciences*, 104(30):12410–12415, July 2007.
- [4] JC Barrett, S Hansoul, DL Nicolae, JH Cho, and RH Duerr. Genome-wide association defines more than 30 distinct susceptibility loci for crohn’s disease. *Nat Genet*, 40, 2007.
- [5] T Becker, A Flaquer, FF Brockschmidt, C Herold, and M Steffens. Evaluation of potential power gain with imputed genotypes in genome-wide association studies. *Hum Hered*, 68, 2009.
- [6] BL Browning and SR Browning. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*, 84, 2009.
- [7] SR Browning and BL Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81, 2007.
- [8] SJ Chanock, T Manolio, M Boehnke, E Boerwinkle, DJ Hunter, G Thomas, JN Hirschhorn, G Abecasis, D Altshuler, and JE Bailey-Wilson. Replicating genotype-phenotype associations. *Nature*, 447(7145):655–660, 2007.

- [9] AG Clark and J Li. Conjuring SNPs to detect associations. *Nat Genet*, 39, 2007.
- [10] D Clayton and PM McKeigue. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet*, 358(9290):1356–1360, 2001.
- [11] JC Cohen, A Pertsemlidis, S Fahmi, S Esmail, GL Vega, SM Grundy, and HH Hobbs. Multiple rare variants in *npc1l1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci USA*, 103(6):1810–1815, 2006.
- [12] JC Cohen, A Pertsemlidis, IK Kotowski, R Graham, CK Garcia, and HH Hobbs. Low ldl cholesterol in individuals of african descent resulting from frequent non-sense mutations in *pcsk9*. *Nat Genet*, 37(2):161–165, 2005.
- [13] A Coventry, LM Bull-Otterson, X Liu, AG Clark, and TJ Maxwell. Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nat Commun*, 1, 2010.
- [14] PI De Bakker, MAR Ferreira, X Jia, BM Neale, S Raychaudhury, and BF Voight. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*, 29, 2008.
- [15] B Devlin and K Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.
- [16] S Dickson, K Wang, I Krantz, H Hakonarson, and D Goldstein. Rare variants create synthetic genome-wide associations. *PLoS Biol*, 8(e1000294), 2010.
- [17] P Donnelly. Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728–731, 2008.
- [18] RM Durbin, GR Abecasis, DL Altshuler, A Auton, LD Brooks, RM Durbin, RA Gibbs, ME Hurles, and GA McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.
- [19] ES Emison, AS McCallion, CS Kashuk, RT Bush, E Grice, S Lin, ME Portnoy, DJ Cutler, ED Green, and A Chakravarti. A common sex-dependent mutation in a ret enhancer underlies hirschsprung disease risk. *Nature*, 434:857–863, 2005.
- [20] DB Goldstein. Common genetic variation and human traits. *The New England journal of medicine*, 360(17):1696–1698, April 2009.

- [21] HHH Göring, JD Terwilliger, and J Blangero. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet*, 69(6):1357–1369, 2001.
- [22] IP Gorlov, OY Gorlova, SR Sunyaev, MR Spitz, and CI Amos. Shifting paradigm of association studies: Value of rare single-nucleotide polymorphisms. *Am J Hum Genet*, 82(1):100 – 112, 2008.
- [23] G Haller, DG Torgerson, C Ober, and EE Thompson. Sequencing the il4 locus in african americans implicates rare noncoding variants in asthma susceptibility. *Journal of Allergy and Clinical Immunology*, 124(6):1204 – 1209.e9, 2009.
- [24] K Hao, E Chudin, J McElwee, and E Schadt. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet*, 10, 2009.
- [25] LA Hindorff, P Sethupathy, HA Junkins, EM Ramos, JP Mehta, FS Collins, and TA Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.
- [26] L Huang, M Jakobsson, TJ Pemberton, M Ibrahim, T Nyambo, S Omar, JK Pritchard, SA Tishkoff, and NA Rosenberg. Haplotype variation and genotype imputation in african populations. –, 00:000–000, Submitted.
- [27] L Huang, Y Li, AB Singleton, JA Hardy, G Abecasis, NA Rosenberg, and P Scheet. Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet*, 84:235–250, 2009.
- [28] L Huang, C Wang, and NA Rosenberg. The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am J Hum Genet*, 85(5):692–698, November 2009.
- [29] L Jostins, KI Morley, and JC Barrett. Imputation of low-frequency variants using the hapmap3 benefits from large, diverse reference sets. *European journal of human genetics EJHG*, 19(6):662–666, 2011.
- [30] JFC Kingman. The coalescent. *Stoch Proc Appl*, 13:235–248, 1982.
- [31] P Kraft and CA Haiman. Gwas identifies a common breast cancer risk allele among brca1 carriers. *Nature Genet*, 42(10):819–820, 2010.

- [32] GV Kryukov, A Shpunt, JA Stamatoyannopoulos, and SR Sunyaev. Power of deep, all-exon resequencing for discovery of human trait genes. *Proceedings of the National Academy of Sciences*, 106(10):3871–3876, 2009.
- [33] J Lazary, A Lazary, X Gonda, A Benko, E Molnar, G Juhasz, and G Bagdy. New evidence for the association of the serotonin transporter gene (slc6a4) haplotypes, threatening life events, and depressive phenotype. *Biological Psychiatry*, 64(6):498–504, 2008.
- [34] B Li and SM Leal. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet*, 83(3):311 – 321, 2008.
- [35] B Li and SM Leal. Discovery of rare variants via sequencing: Implications for the design of complex trait association studies. *PLoS Genet*, 5(5):e1000481, 05 2009.
- [36] Y Li, C Sidore, HM Kang, M Boehnke, and G Abecasis. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Research*, 21(6):940–951, 2011.
- [37] Y Li, C Willer, S Sanna, and G Abecasis. Genotype imputation. *Annual Review of Genomics and Human Genetics*, 10(1):387–406, 2009. PMID: 19715440.
- [38] Y Li, C Willer, S Sanna, and GR Abecasis. Genotype imputation. *Annu Rev Genomics Hum Genet*, 10:387–406, 2009.
- [39] Y Li, CJ Willer, J Ding, P Scheet, and GR Abecasis. MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiol*, 34:816–834, 2010.
- [40] D Lomelin, E Jorgenson, and N Risch. Human genetic variation recognizes functional elements in noncoding sequence. *Genome Research*, 20(3):311–319, 2010.
- [41] BE Madsen and SR Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, 02 2009.
- [42] TA Manolio, FS Collins, NJ Cox, DB Goldstein, LA Hindorff, DJ Hunter, MI McCarthy, EM Ramos, LR Cardon, A Chakravarti, JH Cho, AE Guttmacher,

- A Kong, L Kruglyak, E Mardis, CN Rotimi, M Slatkin, D Valle, AS Whittemore, M Boehnke, AG Clark, EE Eichler, G Gibson, JL Haines, TFC Mackay, SA McCarroll, and PM Visscher. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.
- [43] J Marchini and B Howie. Genotype imputation for genome-wide association studies. *Nature Rev Genet*, 11:499–511, 2010.
- [44] J Marchini, B Howie, S Myers, G McVean, and P Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet*, 39:906–913, 2007.
- [45] P Marjoram, J Molitor, V Plagnol, and S Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 100(26):15324–15328, 2003.
- [46] MI McCarthy, GR Abecasis, LR Cardon, DB Goldstein, J Little, JPA Ioannidis, and JN Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–369, 2008.
- [47] ML Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, December 2009.
- [48] B Mukherjee, J Ahn, SB Gruber, G Rennert, V Moreno, and N Chatterjee. Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genetic Epidemiol*, 32(7):615–626, 2008.
- [49] RP Nair, KC Duffin, C Helms, J Ding, PE Stuart, D Goldgar, JE Gudjonsson, Y Li, T Tejasvi, BJ Feng, A Ruether, S Schreiber, M Weichenthal, D Gladman, P Rahman, SJ Schrodi, S Prahalad, SL Guthery, J Fischer, W Liao, PY Kwok, A Menter, GM Lathrop, CA Wise, AB Begovich, JJ Voorhees, JT Elder, GG Krueger, A Bowcock, and GR Abecasis. Genome-wide scan reveals association of psoriasis with *il-23* and *nf- κ b* pathways. *Nat Genet*, 41(2):199–204, 2009.
- [50] BM Neale, MA Rivas, BF Voight, D Altshuler, B Devlin, M Orho-Melander, S Kathiresan, SM Purcell, K Roeder, and MJ Daly. Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3):e1001322, 2011.
- [51] M Nelis, T Esko, R Mägi, F Zimprich, A Zimprich, D Toncheva, S Karachanak, T Piskckov, I Balack, L Peltonen, E Jakkula, K Rehnström, M Lathrop, S Heath,

- P Galan, S Schreiber, T Meitinger, A Pfeufer, H Wichmann, B Melegh, N Polgr, D Toniolo, P Gasparini, P D'Adamo, J Klovins, L Nikitina-Zake, V Kucinskas, J Kasnauskiene, J Lubinski, T Debniak, S Limborska, A Khrunin, X Estivill, R Rabionet, S Marsal, A Juli, SE Antonarakis, S Deutsch, C Borel, H Attar, M Gagnebin, M Macek, M Krawczak, M Remm, and A Metspalu. Genetic structure of europeans: A view from the northeast. *PLoS ONE*, 4:e5472, 05 2009.
- [52] PC Ng and S Henikoff. SIFT: predicting amino acid changes that affect protein function. *Nucl Acids Res*, 31(13):3812–3814, 2003.
- [53] M Nordborg. Coalescent theory. In D.J. Balding, M.J. Bishop, and Cannings C., editors, *Handbook of Statistical Genetics*, pages 179–212. John Wiley & Sons, Inc., Chichester, U.K., 2001.
- [54] E Pauws, GE Moore, and P Stanier. A functional haplotype variant in the tbx22 promoter is associated with cleft palate and ankyloglossia. *Journal of medical genetics*, 46(8):555–561, August 2009.
- [55] WW Piegorsch, CR Weinberg, and JA Taylor. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Statistics in Medicine*, 13(2):153–162, 1994.
- [56] AL Price, GV Kryukov, PIW de Bakker, SM Purcell, J Staples, LJ Wei, and SR Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet*, 86(6):832 – 838, 2010.
- [57] AL Price, NJ Patterson, RM Plenge, ME Weinblatt, NA Shadick, and D Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.
- [58] JK Pritchard. Are rare variants responsible for susceptibility to complex diseases. *Am J Hum Genet*, 69:124–137, 2001.
- [59] JK Pritchard and NJ Cox. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet*, 11:2417–2423, 2002.
- [60] P Purcell, B Neale, K Todd Brown, L Thomas, MAR Ferreira, D Bender, J Maller, P Sklar, PIW de Bakker, MJ Daly, , and PC Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81:559–575, 2007.

- [61] Z Qin, S Gopalakrishnan, and G Abecasis. An efficient comprehensive search algorithm for tagsnp selection using linkage disequilibrium criteria. *Bioinformatics*, 22(2):220–225, 2006.
- [62] V Ramensky, P Bork, and S Sunyaev. Human non-synonymous SNPs: server and survey. *Nucl Acids Res*, 30(17):3894–3900, 2002.
- [63] Griffiths RC. Asymptotic line-of-descent distributions. *J Math Biology*, 21:67–75, 1984.
- [64] J Reumers, L Conde, I Medina, S Maurer-Stroh, J Van Durme, J Dopazo, F Rousseau, and J Schymkowitz. Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. *Nucl Acids Res*, 36(suppl_1):D825–829, 2008.
- [65] N Risch and K Merikangas. The future of genetic studies of complex human diseases. *Science*, 13(5281):1516–1517, 1996.
- [66] NA Rosenberg. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, 57:1465–1477, 2003.
- [67] SF Schaffner, C Foo, S Gabriel, D Reich, MJ Daly, and D Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11):1576–1583, 2005.
- [68] P Scheet and M Stephens. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78:629–644, 2006.
- [69] LJ Scott, KL Mohlke, LL Bonnycastle, CJ Willer, Y Li, WL Duren, MR Erdos, HM Stringham, PS Chines, AU Jackson, and et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–1345, 2007.
- [70] SG Self, RH Mauritsen, and J Ohara. Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 48(1):31, 1992.
- [71] ST Sherry, MH Ward, M Kholodov, J Baker, L Phan, EM Smigielski, and K Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucl Acids Res*, 29(1):308–311, 2001.

- [72] D Siegmund. Upward bias in estimation of genetic effects. *Am J Hum Genet*, 71(5):1183–1188, 2002.
- [73] P Soronen, H M Ollila, M Antila, K Silander, O M Palo, T Kiesepp, J Linnqvist, L Peltonen, A Tuulio-Henriksson, T Partonen, and et al. Replication of gwas of bipolar disorder: association of snps near *cdh7* with bipolar disorder and visual processing. *Molecular Psychiatry*, 15(1):4–6, 2010.
- [74] CA Spencer, Z Su, P Donnelly, and J Marchini. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 2009.
- [75] N Takahata and M Nei. Gene genealogy and variance of interpopulation nucleotide differences. *Genetics*, 110:325–344, 1985.
- [76] S Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor Popul Biol*, 26:119–164, 1984.
- [77] J. Wakeley. *Coalescent theory*. Roberts & Company Publishers, Greenwood Village, CO, 2008.
- [78] MC Wu, S Lee, T Cai, Y Li, M Boehnke, and X Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*.
- [79] R Xiao and M Boehnke. Quantifying and correcting for the winner’s curse in genetic association studies. *Genetic Epid*, 33(5):453–462, 2009.
- [80] J Yang, B Benyamin, BP McEvoy, S Gordon, AK Henders, DR Nyholt, PA Madden, AC Heath, NG Martin, and GW Montgomery. Common snps explain a large proportion of the heritability for human height. *Nature Genet*, 42(7):565–569, 2010.
- [81] M Zawistowski, S Gopalakrishnan, J Ding, Y Li, S Grimm, and S Zöllner. Extending rare-variant testing strategies: Analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet*, 87(5):604 – 617, 2010.
- [82] E Zeggini, LJ Scott, R Saxena, and BF for the Diabetes Gene Replication And Meta-analysis (DIAGRAM) Consortium. Voight. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40:638–645, 2008.

- [83] H Zhong and RL Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics Oxford England*, 9(4):621–634, 2008.
- [84] H Zhong and RL Prentice. Correcting winner’s curse in odds ratios from genomewide association findings for major complex human diseases. *Genetic Epid*, 34(1):78–91, 2010.
- [85] S Zöllner and JK Pritchard. Overcoming the winners curse: Estimating penetrance parameters from case-control data. *Am J Hum Genet*, 80(4):605–615, 2007.