ABSTRACT

Developing Statistical Methods for Incorporating Complexity in Association Studies

Cameron Palmer

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with hundreds of human traits. Yet the common variant model tested by traditional GWAS only provides an incomplete explanation for the known genetic heritability of many traits. Many divergent methods have been proposed to address the shortcomings of GWAS, including most notably the extension of association methods into rarer variants through whole exome and whole genome sequencing. GWAS methods feature numerous simplifications designed for feasibility and ease of use, as opposed to statistical rigor. Furthermore, no systematic quantification of the performance of GWAS across all traits exists. Beyond improving the utility of data that already exist, a more thorough understanding of the performance of GWAS on common variants may elucidate flaws not in the method but rather in its implementation, which may pose a continued or growing threat to the utility of rare variant association studies now underway.

This thesis focuses on systematic evaluation and incremental improvement of GWAS modeling. We collect a rich dataset containing standardized association results from all GWAS conducted on quantitative human traits, finding that while the majority of published significant results in the field do not disclose sufficient information to determine whether the results are actually valid, those that do replicate precisely in concordance with their statistical power when conducted in samples of similar ancestry and reporting accurate per-locus sample sizes. We then look to the inability of effectively all existing association methods to handle missingness in genetic data, and show that adapting missingness theory from statistics can both increase power and provide a flexible framework for extending most existing tools with minimal effort. We finally undertake novel variant association in a schizophrenia cohort from a bottleneck population. We find that the study itself is confounded by nonrandom population sampling and identity-by-descent, manifesting as batch effects correlated with outcome that remain in novel variants after all sample-wide quality control. On the whole, these results emphasize both the past and present utility and reliability of the GWAS model, as well as the extent to which lessons from the GWAS era must inform genetic studies moving forward.
# Contents

<table>
<thead>
<tr>
<th>List of Figures</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>vii</td>
</tr>
</tbody>
</table>

## 1 Introduction

## 2 Statistical Correction of the Winner’s Curse Explains Replication Variability in Quantitative Trait Genome-Wide Association Studies

2.1 Abstract .......................... 8

2.2 Summary .................................. 9

2.3 Introduction .......................... 9

2.4 Results ................................ 12

2.4.1 Paper Quality Control ................. 12

2.4.2 Paper Characteristics ................. 13

2.4.3 Replication Rates, by Paper .......... 16

2.4.4 Sample Size and Ancestry Explain Replication Inconsistency ............. 19

2.4.5 Functional Enrichment in Replicated Variants .................. 23

2.5 Discussion .......................... 28

2.6 Conclusion .......................... 35

2.7 Methods .......................... 35

2.7.1 Notation .......................... 35

2.7.2 Data Collection .................... 36
2.7.3 Winner’s Curse Correction ........................................... 37
2.7.4 Independence of Loci ................................................. 38
2.7.5 Definition of Replication ............................................ 38
2.7.6 Power to Replicate .................................................. 39
2.7.7 Poisson Binomial Distribution ....................................... 40
2.7.8 Sample Ancestries .................................................... 41
2.7.9 Functional Annotation ............................................... 41

3 Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation 42
3.1 Abstract ................................................................. 42
3.2 Summary ............................................................... 43
3.3 Introduction ........................................................... 43
3.4 Results ................................................................. 45
  3.4.1 Imputation Accuracy by Genotype, Frequency, and Quality .......... 45
  3.4.2 Imputation Probability Consistency .................................. 49
  3.4.3 Signal Enrichment with Multiple Imputation ......................... 51
  3.4.4 MI Changes the Null Distribution of Variants ....................... 51
  3.4.5 Extensibility ....................................................... 54
3.5 Discussion ............................................................ 54
3.6 Conclusion ............................................................ 61
3.7 Methods ............................................................... 61
  3.7.1 Study Datasets ..................................................... 61
  3.7.2 Sample Quality Control ........................................... 62
  3.7.3 Prephasing and Genotype Imputation ................................ 63
  3.7.4 Standard Analysis Methods ........................................ 63
  3.7.5 Statistical Missingness Theory .................................... 64
  3.7.6 Multiple Imputation ............................................... 65
  3.7.7 Meta-Analysis of Multiple Imputation .............................. 66

4 Detecting Novel Variation in an Ashkenazi Jewish Schizophrenia Cohort 67
4.1 Abstract ............................................................... 67
List of Figures

2.1 Schematic of the Winner’s Curse 11
2.2 GWAS Replication by Paper and Journal 17
2.3 GWAS Replication by Paper and Publication Date 18
2.4 Receiver Operating Characteristic for Debiasing Process 20
2.5 Effect of Ancestry and Sample Size on Nominal Replication Prediction 21
2.6 Effect of Ancestry and Sample Size on Bonferroni Replication Prediction 22
2.7 GWAS Replication by Signal Strength 24
2.8 GWAS Replication by Signal Strength: All Studies 26
2.9 GWAS Replication by Signal Strength: Replication in Same Ancestry 26
2.10 GWAS Replication by Signal Strength: Replication in Different Ancestry 27
2.11 GWAS Replication by Signal Strength: Study in European Ancestry 27
2.12 Chromosomal Distribution of Analyzed Loci 29
2.13 Chromosomal Distribution of Bonferroni-replicated Loci 30
3.1 IMPUTE2 Dosage Performance by $r^2$ and Frequency 46
3.2 minimac3 Dosage Performance by $r^2$ and Frequency 47
3.3 IMPUTE2 Quality Distribution 47
3.4 minimac3 Quality Distribution 48
3.5 IMPUTE2 Best-Guess Performance by $r^2$ and Frequency 48
3.6 IMPUTE2 Probability Consistency by RSQ 49
3.7 IMPUTE2 Probability Consistency by Frequency 50
3.8 BMI AUROC Improvement with MI 52
3.9 Null Distribution Shifts 53
3.10 Mean Convergence for MI ........................................ 58
3.11 Variance Convergence for MI ..................................... 58

4.1 Genome-wide All-Variant Quality Associated with Outcome .......... 71
4.2 IBS by Schizophrenia Affectation Status, All Variants ................. 72
4.3 Genome-wide Pass-Variant Quality Shows Reduced Association with Outcome  .... 73
4.4 Genome-wide Full-QC Variants Show No Association with Outcome .... 74
4.5 IBS by Schizophrenia Affectation Status, Filtered Variants ............. 75
4.6 Novel Variants Remain Associated with Outcome ........................ 76
4.7 Novel Loss of Function Singletons Stratified by Cohort ................... 78
4.8 Novel Synonymous Singletons Stratified by Cohort ........................ 79
4.9 Comparative Allele Count Distributions ................................... 83
4.10 Null Inflation Under Fisher Model .................................... 84
4.11 Rare Variant Allele Loading in Imbalanced Sample ........................ 86
List of Tables

1.1 Example GWAS Replication Counts ........................................... 4
2.1 Publication Venue for Significant, Replicated qtGWAS .................. 13
2.2 Publication Venue for Significant qtGWAS ................................. 14
2.3 QC Process for All qtGWAS Papers .......................................... 15
2.4 Ancestry Distribution of All GWAS Samples ................................. 16
2.5 Replication Prediction by Strength of Signal .............................. 25
4.1 Distribution of Novel Variants by Functional Class ..................... 70
4.2 Distribution of Doubletons by Functional Class ........................ 80
4.3 Fisher/CMH Test for Novel Functional Variants .......................... 81
I would like to acknowledge advisor Dr. Itsik Pe'er; Committee Chair Dr. Tuuli Lappalainen; Committee Members Dr. Shuang Wang, Dr. Iuliana Ionita-Laza, and Dr. Guy Sella, for their guidance and evaluation of this project.

For the use of their genetic and trait data in the imputation quality study (Chapter 3), I would like to acknowledge the anonymous individuals who participated in the Health ABC (Health, Aging, and Body Composition) and NFBC66 (Northern Finnish Birth Cohort, birth year 1966) studies. For the use of their genetic and trait data in the schizophrenia rare variant study (Chapter 4), I would like to acknowledge the anonymous individuals who participated in the TAGC (Ashkenazi Genome Consortium) study.
Chapter 1

Introduction

The past decade of human genetics has been dominated by the success of the genome-wide association study (GWAS). The GWAS model prioritizes the common single nucleotide polymorphism (SNP) as the most likely causal agent of heritable variation in complex traits. While importantly a model born of convenience, the Common Trait (or Disease)/Common Variant hypothesis [1] has proven a remarkably effective representation of genetic causality. Tens of thousands of associations have been detected for hundreds of common traits [2, 3]. The technology continues to improve in both genome coverage and price; modern GWAS bear little resemblance to the original studies in the mid–2000s, and continue to expand into traits and populations left out of the initial wave of study.

The inability of GWAS to consistently explain the entirety of genetic heritability [4, 5, 6] of traits has been attributed to many different possible causes. The simplest explanation suggests that the complex genetic architecture of polygenic human traits manifests in potentially thousands of common variants of low effect. Under this model, missing heritability is simply a function of insufficient power: collecting more samples will eventually address the issue. Evidence supporting this explanation has been most thoroughly characterized for common anthropometric traits (height, body mass index as a proxy for adiposity, etc.): statistical models [7, 8] project the total genetic heritability of these traits captured by GWAS to be more than 80% of the total heritability for the trait. These models remain, however, projections: the actual genetic causes have not been specifically identified, but their behavior has been characterized en masse. The implications of the tendency of GWAS to characterize patterns but not individual targets will be discussed more in Chapter 2.
Perhaps the most widely accepted explanation for Missing Heritability is simply that GWAS are simply looking in the wrong place: the Common Trait/Common Variant hypothesis ignores the substantial contribution of rare genetic variation to common heritable traits. The contribution of rare variants to human traits is now well-established in many reliable studies (for example among many others, [9, 10, 11]). This evidence is largely the motivation behind and goal of the transition of genetics from array-based technology to exome and genome sequencing.

Additional limitations of the GWAS model abound. Association studies carry with them a series of simplifying assumptions. In the vast majority of studies, a cross-sectional trait is tested for marginal association with a single common SNP by linear or logistic regression. The only major deviation from this model has come with the era of rare variant testing, as minor allele counts for rare variants (below 1% of alleles present in a sample) are insufficient for marginal testing, and thus must be combined in some form of burden test (discussed further in Chapter 4). Testing independent sites for a single collection of a phenotype leads to acceptable statistical complexity when from $5 \cdot 10^5$ to $4 \cdot 10^7$ variants are tested in an individual study. Yet that is the only reason for such assumptions: there is no particular reason to believe that this model is an accurate general representation of meaningful biological characteristics or processes. The heritability estimated from a genetic model is only as accurate as the model itself.

It is important to note that the projections of explained heritability in [7, 8] differ substantially from the metrics cited in [12], when the Missing Heritability issue first reached prominence at the beginning of this decade. For human height, the proportion of explained genetic heritability was cited as approximately 6.25% in 2010; as of 2015, that number was closer to 85%. There are important caveats to the 2015 number: closer to 25% of explained genetic heritability is attributable to specific loci, and the remainder is based on statistical models of bulk behaviors of the rest of the genome. Furthermore, some genetic heritability is still unexplained, and this single result is based on the combined effort and resources of a nontrivial percentage of the entire field. Yet it is difficult to describe this change as anything but promising, and certainly supportive of the general model of GWAS at large.

This dissertation can be framed in light of this consideration: a reevaluation of the field of GWAS, with the benefit of improved statistics and retrospect, shows that the field at large is in fact substantially more effective at explaining human genetic heritability than first thought. We ask what other aspects of GWAS might be significantly improved with consideration and standardization. We
An early concern in the design of this study was to provide metrics regarding the reliability of the GWAS model: to quantify the vague assessment of whether a GWAS study is typically rigorous and informative. The very idea of assessing the “reliability” of GWAS is complicated, and requires some details of the GWAS study design. The standard genome-wide association study has a two-stage design. In the first (discovery) stage, some number of cohorts, consisting of human subjects with phenotype and epidemiological covariate data, as well as DNA assayed on one of many possible GWAS platforms, are tested for association between each assayed variant in turn and a phenotype of interest. In the second (replication) stage, some proportion of strongly associated variants are tested in an “independent” set of cohorts. The precise definition of replication varies from study to study (see Chapter 2). The use of Bonferroni corrected p-values for significance testing is prevalent, wherein replication is defined as either:

1. replication p-value exceeding Bonferroni corrected $\alpha$ for number of attempted replications; or

2. meta-analyzed discovery and replication p-value exceeding Bonferroni corrected $\alpha$ for number of attempted discoveries

Regardless of the precise definitions used, the discovery process in GWAS is subject to regression to the mean [13], colloquially called the “Winner’s Curse.” In short, the process by which the most significant variants are selected for either replication or publication induces an upward bias in the estimated magnitude of effect of the “winning” variants, as the sampling error distribution for variants near the selection cutoff are probabilistically truncated. While significant theoretical work has been conducted to create methods of correcting discovery effect estimates [14, 15, 16, 17, 18, 19, 20], these have not enjoyed widespread adoption by the field at large. Rather, studies rely on large sample sizes and the relative cost efficiency of in silico replication lookups to accommodate aberrantly high replication failure. Thus do we see results such as those in Table 1.1.

What is ultimately deceptive about these extreme examples is that there is nothing intrinsically wrong with these studies. The power calculations are inherently biased, as in none of these cases (nor indeed in any of the other 327 papers from which these data are taken) did the authors attempt to prospectively adjust for the Winner’s Curse. As such, there is no evidence from these data that these GWAS were inefficient at replication. Nevertheless, the psychological impact of failing to replicate present these results in the form of three vignettes: methods and aspects of the association study methodology upon which we can significantly improve with formal analysis.
Table 1.1: Exemplar replication performances from actual studies. PMID: PubMed ID code for particular study. Attempted: number of variants brought forward from discovery into replication. Expected: MLE number of replications based on reported power to replicate at Bonferroni threshold from the published study. Observed: number of replications amongst the attempted variants, computed as number of variants surpassing Bonferroni-corrected replication p-value. See Chapter 2 for more on these data.

<table>
<thead>
<tr>
<th>PMID</th>
<th>Attempted</th>
<th>Expected</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>18846228</td>
<td>21</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>20010834</td>
<td>30</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>20887962</td>
<td>32</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>22021425</td>
<td>84</td>
<td>81</td>
<td>5</td>
</tr>
<tr>
<td>23669352</td>
<td>67</td>
<td>59</td>
<td>3</td>
</tr>
</tbody>
</table>

thousands of apparent signals should not be underestimated. Such results may leave the reader with the impression that GWAS consistently fail to replicate, and these papers make no direct effort to address this concern.

The unfortunate result of these observations is that it is simply not possible to make a general statement about whether GWAS perform in practice as they are expected to in theory. To make such a statement, one would need to harvest GWAS discovery and replication data from hundreds of studies across dozens of journals, standardize the contents, apply a Winner’s Curse correction algorithm to those data, and then using uniform definitions of replication, fit models for how well such studies were predicted to replicate according to their apparent statistical power. No such undertaking has been attempted, to our knowledge: at most, the studies showcasing methods to correct the Winner’s Curse typically deploy their method on no more than six selected GWAS [17, 20], which even ignoring selection bias does not represent a sample sufficiently powered to render conclusions about the field as a whole.

Thus, as discussed at length in Chapter 2, we have undertaken just the sort of analysis described above. We have focused in particular on quantitative (normally-distributed) traits, for which less theory has been developed and which have received negligible attention in existing evaluation studies. We show two major results. The first is that the field at large underreplicates using naïve estimates of power based on cursed discovery data, but that underreplication is completely abrogated by (1) correcting the Winner’s Curse, and excluding studies that (2) conduct discovery and replication in cohorts of distinct ancestry, as well as those that (3) incorrectly report per-variant sample sizes. While the first two factors are widely acknowledged, incorrect sample sizes (in particular, uniformly
overestimated sample sizes based on the *maximum possible sample size*) are a component of our second major finding from the study: more than half of published quantitative trait GWAS studies report insufficient information to actually prove their replication performance. Lack of uniform reporting standards in the field have created a vast corpus of unproven claims. This is an urgent problem that needs to be addressed.

Moving beyond quantification of existing results, we consider particular components of the GWAS study design that invite reconsideration and rigorous treatment. An integral component of any GWAS of the last decade is the process of genotype imputation. As discussed at length in Chapter 3, imputation is the process of probabilistically estimating untyped variants in a sample using external reference data. Imputation methods have received substantial attention in the GWAS era ([21, 22], among dozens). Less widely discussed is the impact these methods have on the choice of downstream statistical tests. As a rule, statistical genetics methods are not capable of directly operating on probabilistic genotypes, without some form of approximation that reduces the available information concerning genotypic uncertainty. But calling genotypes from imputed probabilities is a problematic endeavor, as genotype imputation introduces characteristic biases and high error rates, as we report in some depth in Chapter 3.

Data missingness and uncertainty have been thoroughly characterized in the general statistics literature [23]. In that context, the method of Multiple Imputation (MI; upon which genotype imputation is in fact modeled) represents the formally correct method of handling probabilistic data in statistical tests. While there are periodic efforts to use MI in particular contexts, it has not been formally evaluated in the context of probabilistic genotypes. We therefore have fully characterized MI in the context of GWAS association. We also deem that the primary hurdle towards using MI in genomics is not theoretical but rather practical: each individual analyst and statistician needs to know of and implement MI from scratch, leading to the hodgepodge of approximate methods currently in use. To address this issue, we have created a software tool [24] that streamlines the process of integrating existing analysis tools with the MI regime. We hope that such a software interface may provide some impetus to adopting more appropriate missingness handling in future work.

Finally, any discussion of the state of association studies would be incomplete without consideration of rare variants. The initial wave of association studies made use of DNA microarrays, and as such entirely avoided consideration of variation below appreciable frequencies in reference popu-
lations. Yet the advent of the Missing Heritability problem [4, 5, 6], combined with ever-decreasing costs of exome and genome sequencing, rapidly moved rare variation to the forefront of statistical genetics.

Even with the technical restriction on sequencing alleviated, rare variation offers unique challenges for both quality control and statistical analysis. The process of sequence calling relies on both prior probabilities of variation and observed numbers of reads of a given alternate allele [25]; both of these parameters are most difficult to specify at rare and novel loci. Furthermore, even assuming high quality calls, standard analysis tools such as regression cannot reliably estimate correlations between variation and outcome with so few observations of the minor allele. This observation has led to the proliferation of burden tests [26, 27, 28, 29], in which pools of rare variation are combined under various models to attempt to increase power to detect association.

Formal burden tests are exceedingly useful, but even so require substantial resources to generate sample sizes sufficient to create well-powered tests. There is the potential to lessen these demands by creative selection of study sample. Allele frequency may differ substantially between populations of restricted mating over short numbers of generations, independent of selective pressure. Rare variation in one population may be substantially more common, and thus easier to ascertain, in another population of different demographic history [30]. In particular, bottleneck populations, in which recent population contraction and expansion have greatly restricted the effective number of haplotypes in a reproductive group, offer a particularly attractive opportunity for efficient rare variant testing.

In Chapter 4, in the context of a larger consortium case/control study of schizophrenia in Ashkenazim, we conduct novel variant burden testing as a first approach towards characterizing the rare variant burden of this cohort. Other collaborators recently conducted a similar study [31] of Alzheimer’s disease in Ashkenazim, so we take their statistical model as our initial approach. While this reference study was able to detect a strong preferential burden of rare, functionally-annotated variants in their case samples, we find instead in this case a complicated series of stratification effects ultimately attributable to nonrandom sampling from hidden pedigrees and nonrandom sequencing of individuals with respect to both case status and relatedness of samples. Furthermore, in calibrating our statistical model, we find that published work on functional novel burden testing is subject to substantially miscalibrated Type I error, due to improprieties in selecting reference groups of novel variation to serve as a null control. While disappointing from the perspective of biological discovery,
the realization that effective IBD of samples may cause local distortions in singleton and doubleton call rates is important and provides essential guidance for sequencing study design in the future.

Taken as a whole, these vignettes of statistical genetics provide a meaningful contribution to the field. They broadly suggest that a large variety of highly technical issues remain to be resolved to improve the GWAS study model. One would hope that many of the observations made herein additionally contribute to the larger conversation surrounding statistical genetics as a paradigm for understanding trait variation.
Chapter 2

Statistical Correction of the Winner’s Curse Explains Replication Variability in Quantitative Trait Genome-Wide Association Studies

2.1 Abstract

Genome-wide association studies (GWAS) have identified hundreds of SNPs responsible for variation in human quantitative traits. However, genome-wide-significant associations often fail to replicate across independent cohorts, in apparent inconsistency with their apparent strong effects in discovery cohorts. This limited success of replication raises pervasive questions about the utility of the GWAS field. We identify all 332 studies of quantitative traits from the NHGRI-EBI GWAS Database with attempted replication. We find that the majority of studies provide insufficient data to evaluate replication rates. The remaining papers replicate significantly worse than expected ($p < 10^{-14}$), even when adjusting for regression-to-the-mean of effect size between discovery- and replication-
cohorts termed the Winner’s Curse ($p < 10^{-16}$). We show this is due in part to misreporting replication cohort-size as a maximum number, rather than per-locus one. In 39 studies accurately reporting per-locus cohort-size for attempted replication of 707 loci in samples with similar ancestry, replication rate matched expectation (predicted 458, observed 457, $p = 0.94$). In contrast, ancestry differences between replication and discovery (13 studies, 385 loci) cause the most highly-powered decile of loci to replicate worse than expected, due to difference in linkage disequilibrium.

2.2 Summary

The majority of associations between common genetic variation and human traits come from genome-wide association studies, which have analyzed millions of single-nucleotide polymorphisms in millions of samples. These kinds of studies pose serious statistical challenges to discovering new associations. Finite resources restrict the number of candidate associations that can brought forward into validation samples, introducing the need for a significance threshold. This threshold creates a phenomenon called the Winner’s Curse, in which candidate associations close to the discovery threshold are more likely to have biased overestimates of the variant’s true association in the sampled population. We survey all human quantitative trait association studies that validated at least one signal. We find the majority of these studies do not publish sufficient information to actually support their claims of replication. For studies that did, we computationally correct the Winner’s Curse and evaluate replication performance. While all variants combined replicate significantly less than expected, we find that the subset of studies that (1) perform both discovery and replication in samples of the same ancestry; and (2) report accurate per-variant sample sizes, replicate as expected. This study provides strong, rigorous evidence for the broad reliability of genome-wide association studies. We furthermore provide a model for more efficient selection of variants as candidates for replication, as selecting variants using cursed discovery data enriches for variants with little real evidence for trait association.

2.3 Introduction

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex human traits [3]. GWAS are most commonly two-stage designs, with a discovery
study followed up by (possibly several) internal replication studies on independent samples. Due to
the number of variants tested in the typical association study, replication is only attempted for a
small fraction of the discovered variants exceeding a p-value threshold adjusted for \(10^6\) independent
tests. The tradeoff between study power per-variant and resources, along with the strategy of testing
millions of variants for association, leads to study designs where many associated variants of low
effect size [32] are underpowered to be detected.

The Winner’s Curse (WC) is the systematic overestimation of effects ascertained by thresholding.
This phenomenon is induced by ascertainment of the most significant GWAS signals for reporting:
introducing a threshold on statistical significance means that the selected set of signals will preferen-
tially contain variants whose effects are overestimated in a particular study sample due to chance
noise (Figure 2.1). This tendency of studies to overestimate their association with a phenotype in the
discovery cohort might cause them to replicate at an unexpectedly low rate, increasing the apparent
unreliability of results from the field. This study relies on computationally correcting this biased
overestimate of effect size, in order to produce accurate estimates of the chances for replication.

Several models for directly estimating bias in effect estimates have been developed. Parametric
models, based predominantly on the theory established in [18], generate a maximum likelihood
estimation of the effect estimate based on the impact of introducing a p-value threshold into the
reported list of variants; thus, test statistics close to the threshold tend to be biased more severely
than those more substantially exceeding the threshold. Alternatively, nonparametric bootstrap
correction of the Winner’s Curse using individual-level genetic data [20] has been implemented.
Evaluation of these models for binary [17, 15] and quantitative [14] traits has been limited to
simulations and a small number of studies, without establishing the importance of WC-correction
to GWAS study design.

Further complicating matters, there is no single accepted standard for successful internal re-
 replication of a variant in a GWAS. Across the GWAS considered in this study we have observed
several definitions of replication. The variability of these definitions leads to differing standards of
“replicating signal” in the literature, and complicates an evaluation of replicability across the field.

Variants found to be trait-associated in GWAS are not necessarily causal [33], due to linkage
disequilibrium (LD) between common variants. Causal variants are expected to replicate, whereas
significantly-associated noncausal variants will only replicate if they remain linked to a causal variant
in a replication study. The predicted rate of replication for noncausal variants is not trivial, as in
Figure 2.1: **Schematic diagram representing the bias from the Winner’s Curse.** A SNP with fixed frequency, sample size, and phenotypic variance explained was repeatedly drawn at random, and the accompanying standard normal trait was simulated. The apparent effect of SNP on trait was estimated exclusively from simulations exceeding progressively more stringent replication $\alpha$. X-axis: replication $\alpha$; Y-axis: estimated SNP effect from linear regression. Two simulations with different effect estimates (horizontal lines) are shown. The height of the vertical lines corresponds to the average bias (curse) introduced by discovery prioritization.
general the causal variant in a locus is unknown and may not be assayed in the study. In particular, more GWAS now attempt discovery and replication in samples of distinct ancestries, which are expected to have substantially different LD patterns across much of the genome. Moreover, even when LD between a hidden causal variant and its observed proxy are comparable across replication and discovery, there remains an open question as to whether, and in what contexts, SNPs are expected to have comparable effect in different ancestral backgrounds; existing work, in particular using the same database from this study [34], has provided inconclusive results that may be confounded by both the Winner’s Curse and a preponderance of false positive variants.

In this study [35] we seek to evaluate the replicability of SNPs in genome-wide association studies across the field of human quantitative trait genetics. We specifically consider quantitative trait studies as they are underrepresented amongst theoretical work for correcting the Winner’s Curse, and represent a meaningful subset of the field (33% of papers considered) that is still sufficiently small that we may feasibly evaluate all existing studies. The NHGRI-EBI GWAS Catalog [2, 3] provides a reasonably complete database of publications claiming to report genome-wide significant associations between variants and human traits. We use this catalog as a tool to identify the vast majority of papers in the field. Using only summary data reported in these papers, we modeled the Winner’s Curse in all papers providing enough information to actually support their claims of replication. We recomputed their replication rates according to the nominal and Bonferroni standards of replication, thus introducing a standardized regime to make generalizations about replication efficiency across all studies. Together, we obtain reliable metrics to evaluate the state of human quantitative trait genetics as a reproducible scientific domain.

2.4 Results

2.4.1 Paper Quality Control

We considered all 332 GWAS papers for quantitative-traits in the database [2, 3] from journals we deemed pertinent to human genetics (see Table 2.1, Table 2.2) that attempted replication of discovered variants. We filtered this pool, requiring study design of strict thresholding, reports of data needed to calculate bias in effect sizes [18], and related consistency criteria (see Methods, Table 2.3). This reduced the pool to $k = 100$ post-QC papers (30%) for analysis.

The above counts consider each paper as a functional unit. In some cases, a single paper will
<table>
<thead>
<tr>
<th>Journal</th>
<th>Analyzed</th>
<th>Excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am J Hum Genet</td>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>Am J Med Genet B Neuropsychiatr Genet</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>BMC Med Genet</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Circ Cardiovasc Genet</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Front Genet</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Gene</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Genet Epidemiol</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Hum Genet</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Hum Mol Genet</td>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td>J Med Genet</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Nat Genet</td>
<td>26</td>
<td>48</td>
</tr>
<tr>
<td>PLoS Genet</td>
<td>18</td>
<td>31</td>
</tr>
<tr>
<td>PLoS One</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>Science</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>100</strong></td>
<td><strong>204</strong></td>
</tr>
</tbody>
</table>

Table 2.1: **Distribution of papers across journals, for journals that had at least one article with sufficient information for analysis.** The full distribution of all journals analyzed in the study, including those with all papers excluded, is in Table 2.2.

Publish multiple GWAS: that is, multiple phenotypes will be analyzed in the same paper. The 100 papers passing QC correspond to 134 “studies,” with 79 papers containing only a single study, and the remainder having fewer than 6 studies each. As these additional studies typically contribute a very small number of variants to our analysis, we proceed with the paper count as a more honest reflection of the scope of our analysis.

### 2.4.2 Paper Characteristics

The sum of discovery sample sizes across all analyzed papers reaches approximately 1.8 million non-unique individuals. The majority (88%) of this cumulative count have European ancestry, framing the analysis in the context of this group. This 6.7-fold over representation of European ancestry is part of uneven sampling of world populations in GWAS (Table 2.4).

The tally of variants these papers attempted to replicate lists 2691 non-unique variants, each passing the corresponding paper-specific p-value threshold in its discovery cohort. Many of these papers include linked variants on this list, introducing partial redundancies. We filtered dependent variants (Methods) to obtain 1652 loci for analysis, independent within each paper.
<table>
<thead>
<tr>
<th>Journal</th>
<th>Analyzed</th>
<th>Filtered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Am J Hum Genet</td>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>Am J Med Genet B Neuropsychiatr Genet</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Ann Hum Genet</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BMC Genomics</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>BMC Med Genet</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>BMC Med Genomics</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Circ Cardiovasc Genet</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Eur J Hum Genet</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Front Genet</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>G3 (Bethesda)</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Gene</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Genes Brain Behav</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Genes Immun</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Genet Epidemiol</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Genomics</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Genomics Inform</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hum Genet</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Hum Mol Genet</td>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td>Immunogenetics</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>J Hum Genet</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>J Med Genet</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Nat Genet</td>
<td>26</td>
<td>48</td>
</tr>
<tr>
<td>Nature</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>PLoS Genet</td>
<td>18</td>
<td>31</td>
</tr>
<tr>
<td>PLoS One</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>Pharmacogenet Genomics</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pharmacogenomics</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Pharmacogenomics J</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Science</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Twin Res Hum Genet</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Total:</td>
<td>100</td>
<td>232</td>
</tr>
</tbody>
</table>

Table 2.2: Distribution of quantitative trait GWAS papers across journals, for journals that had at least one article annotated as "attempting replication" in the NHGRI-EBI GWAS Database as of 04 Feb 2016.
<table>
<thead>
<tr>
<th>Filter type</th>
<th>Filter subcategories</th>
<th>Papers removed</th>
<th>Papers remaining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative trait studies in GWAS Catalog</td>
<td></td>
<td>691</td>
<td>332</td>
</tr>
<tr>
<td>Complete data reporting for study</td>
<td>all attempted</td>
<td>121</td>
<td>211</td>
</tr>
<tr>
<td></td>
<td>replications reported</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Threshold replication model</td>
<td></td>
<td>31</td>
<td>180</td>
</tr>
<tr>
<td>Complete data reporting for variants</td>
<td>trait mean,</td>
<td>70</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>variant frequency,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sample size,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>effect size,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>standard error or p-value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical test restrictions</td>
<td>frequentist, additive,</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>same trait in both stages, etc.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: **Summary of quality control process applied to 332 candidate quantitative trait GWAS papers.** Filters are hierarchical, in the sense that a paper failing a criterion at one stage of the process was not evaluated for other criteria of lower priority. The 100 papers passing these filters were further annotated for other discrepancies that would interfere with but not entirely prevent debiasing calculations: allele frequency from reference population instead of actual cohort; allele frequency from one round but not both; maximum sample size reported instead of per-variant sample size, reflecting missingness; extreme low precision errors; etc.
Table 2.4: Ancestry distribution of samples included in GWAS. Rows are as follows: (1) “Totals”: number of samples of a given ancestry in analyzed papers, with redundancy between studies published multiple times; (2) “Rate in GWAS”: percentage of total samples considered that were of this ancestry; (3) “Rate in Population”: percentage of world’s population that is of this ancestry; (4) “Enrichment in GWAS”: relative over (or under) representation of ancestry in GWAS relative to its rate in the world. Ancestry labels are approximations with the standard correspondences to HapMap2 reference samples (European = CEU, East Asian = JPT+CHB, African = YRI); here, “African American” denotes samples reported with that nomenclature, which typically corresponds to 80:20 admixture between ancestral sub-Saharan African and Western European genetics [36]. All of these equivalences are oversimplifications but correspond to assumptions widely used in the field. Counts are computed from totals across all papers analyzed in this study, not adjusting for duplicate uses of the same datasets across multiple studies. Total sample sizes are maximum counts of samples assuming no per-genotype missingness is present. The totals are rounded to the nearest integer as several imputed studies reported nonintegral sample sizes. Row 3 percentages in world population are approximations based on demographic data from 2014-2015 [37, 38].

<table>
<thead>
<tr>
<th></th>
<th>European</th>
<th>East Asian</th>
<th>African</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Totals</td>
<td>1601628</td>
<td>135472</td>
<td>1226</td>
<td>80006</td>
</tr>
<tr>
<td>Rate in GWAS (percent)</td>
<td>88.08</td>
<td>7.45</td>
<td>0.07</td>
<td>4.4</td>
</tr>
<tr>
<td>Rate in Population (percent)</td>
<td>13.3</td>
<td>59.8</td>
<td>13.1</td>
<td>0.5</td>
</tr>
<tr>
<td>Enrichment in GWAS (percent)</td>
<td>670.8</td>
<td>12.5</td>
<td>0.51</td>
<td>821.6</td>
</tr>
</tbody>
</table>

2.4.3 Replication Rates, by Paper

At a nominal threshold $\alpha = 0.05$, we observe 793/1652 independent loci to replicate (48%) across 100 papers. Based on the raw effect sizes reported in the discovery cohort, we would have expected 1498 loci to replicate (90.7%), significantly more than observed (two-tailed Poisson binomial $p = 4.2 \cdot 10^{-15}$). Statistical correction of WC leads to a prediction of 888 replicated loci (53.8%), 7-fold closer but still significantly more than observed ($p < 3 \cdot 10^{-16}$). Replacing the nominal threshold by Bonferroni-adjusted thresholds ($\alpha = \frac{0.05}{\# \text{ loci attempted in a particular paper}}$), we observe 519 replicated loci (31.4%), significantly different than both raw ($p = 3.3 \cdot 10^{-14}$) and WC-corrected ($p = 9.0 \cdot 10^{-15}$) replication predictions of 1235 (74.8%) and 610 (36.9%) loci, respectively.

Predicting WC-corrected replication rates per paper (Poisson binomial distribution), we observe excess of papers both over- and under-performing their respective expectations (Figure 2.2A). This excess significantly correlates with publication venue (Figure 2.2B). Specifically, papers in higher impact journals tended to over-replicate, consistent with publication bias [39, 40, 41] (Discussion). We also consider weak correlation between paper replication behavior and date of publication in Figure 2.3.
Figure 2.2: **Expected and observed replication rate per publication, stratified by journal.**

Top panel (A): predicted versus expected replication for each paper. Each paper is flagged as being within 95% confidence of predicted replication rate under WC-corrected model (dots), greater (diamonds) or lower (Xs) than expectation. X-axis: predicted number of replications in a given paper, calculated as the sum across all loci of power to replicate based on WC-corrected discovery effect estimates. Y-axis: observed (jittered integer) number of replications in the paper. Colors correspond to journals. Replication is defined as a one-tailed replication p-value surpassing a per-paper Bonferroni threshold: $\frac{0.05}{\#\text{loci attempted in paper}}$. Confidence intervals defined as 95% confidence according to Poisson binomial draws from the WC-corrected power distribution. Bottom panel (B): distinct behaviors in journals depending on which set of papers is considered. Clusters correspond to paper quality (point shapes) from top panel; confidence intervals are 95% confidence intervals from the binomial distribution. Red lines are expected bar heights assuming that the observed paper data correspond to the WC-corrected model.
Figure 2.3: **Expected and observed replication rate per publication aligned with publication date.** Top panel (A): predicted versus expected replication for each paper, with each paper flagged as being within 95% confidence of predicted replication rate under WC-corrected model, greater than or less than expectation. X-axis: predicted number of replications in a given paper, calculated as the sum across all loci of power to replicate based on WC-corrected discovery effect estimates. Y-axis: observed number of replications in the paper. Colors correspond to publication dates binned into six month intervals, from 2008 to 2014. Point shapes to correspondence between observed and expected rates. Replication is defined as a one-tailed replication p-value surpassing a per-paper Bonferroni threshold: $\frac{0.05}{\text{# loci per paper}}$. Confidence intervals defined as 95% confidence according to Poisson binomial draws from the WC-corrected power distribution. Bottom panel (B): papers underperforming the WC-corrected model tend to have been published later; papers overperforming the WC-corrected model tend to have been published earlier. While these trends are significant, they are dominated by the large number of QC-passing papers from Nature Genetics, which possess an earlier average publication date as well as a replication rate significantly higher than expected based on reported power.
2.4.4 Sample Size and Ancestry Explain Replication Inconsistency

Few papers \((k = 13)\) discovered variants in one continental ancestry and attempted replication in another. This study design may hurt replication beyond WC due to population-specific effects, including linkage disequilibrium. Most \((48/87)\) remaining papers reported single sample size \(N\) for replication across all attempted variants, not reflecting different fractions of missing data for each variant. Note that this includes genotypes missing from association analysis, rather than unmeasured genotypes whose analysis was conducted within the study, even if through imputation. In particular, studies conducting meta-analysis may only obtain variant data from a subset of their contributing cohorts, leading to large discrepancies in effective sample size per locus. This exaggerated replication sample size overestimates power to replicate and thus inflates predicted replication rate.

The remaining 39 papers with 707 discovered loci both maintained continental ancestry across discovery and replication while also correctly reporting per-locus \(N\). At nominal threshold, 457 loci (64.6\%) replicate, consistent (Poisson binomial \(p = 0.94\)) with the WC-corrected prediction of 458 loci (64.8\%). Considering instead the more stringent Bonferroni correction, observed replication of 304 loci (43\%) was also consistent \((p = 0.14)\) with the 316 expected (44.7\%). In both cases, predicting replication without WC-correction fails \((all p < 10^{-14})\). Considering all thresholds across these papers, WC-correction significantly improved sensitivity over raw discovery estimates (ROC AUC 0.785 vs. 0.582, DeLong two-tailed \(p < 2 \cdot 10^{-16}\); see Figure 2.4). We thus hereafter consider only WC-corrected estimates.

The improved fit amongst these 39 remaining papers is not explained by reduction in power to reject fit: fit is more improved than chance expectation (based on simulations on subsets of variants with matched power to observed; nominal replication, \(p < 0.001\); Bonferroni replication, \(p < 0.001\)). Furthermore, both \(N\) and ancestry filters are required for good model fit (see Figure 2.5 and Figure 2.6).

We further tested the importance of per-locus sample size reporting by repeating the replication rate analysis on 39 papers/707 loci with correct ancestry and sample size, but instead using the maximum available sample size for each study. Correcting the Winner’s Curse using these aberrant sample sizes, the predicted rates of replication are no longer consistent with observed data (nominal replication: \(p = 0.0495\); Bonferroni replication: \(p = 0.000202\)). These results further support the conclusion that correct per-locus sample size reporting is essential for both accurate Winner’s Curse
Figure 2.4: Improved prediction of replication with Winner’s Curse-corrected discovery effect estimates. Plotted are receiver operating characteristics (ROC) for raw discovery and WC-corrected estimates of power to nominally replicate variants in individual studies. X-axis: false positive rate, computed as number of non-replicating variants having greater than a given power to replicate; Y-axis: true positive rate, computed as number of replicating variants having such power. AUC: 0.795 (WC-corrected) versus 0.579 (biased); DeLong’s two-tailed $p < 2 \cdot 10^{-16}$. Considers studies with same ancestry in both discovery and replication and correct per-locus sample sizes, as used in the main results.
Improvement in fit for nominal replication prediction requires both same ancestry replication and per-locus sample sizes. Each node corresponds to one method of subsetting the data: all papers; papers conducting discovery and replication in the same continental ancestry; papers reporting correct per-locus sample sizes; or papers doing both. Probabilities in nodes correspond to two-tailed Poisson binomial fit test for prediction of nominal replication rates after Winner’s Curse correction. Probabilities along edges correspond to the chance of randomly seeing improvement of fit between connected nodes at least as large as observed due to loss of power exclusively, based on 10000 simulated subsamplings from the source node matched on total predicted power to replicate. Relative position of nodes along vertical axis corresponds to number of loci removed in the subset.
Figure 2.6: Improvement in fit for Bonferroni replication prediction requires both same ancestry replication and per-locus sample sizes. Each node corresponds to one method of subsetting the data: all papers; papers conducting discovery and replication in the same continental ancestry; papers reporting correct per-locus sample sizes; or papers doing both. Probabilities in nodes correspond to two-tailed Poisson binomial fit test for prediction of Bonferroni replication rates after Winner’s Curse correction. Probabilities along edges correspond to the chance of randomly seeing improvement of fit between connected nodes at least as large as observed due to loss of power exclusively, based on 10000 simulated subsamplings from the source node matched on total predicted power to replicate. Relative position of nodes along vertical axis corresponds to number of loci removed in the subset.
We next investigated the relationship between the strength of a signal and its replication rate. We partitioned all loci across all 100 papers into deciles according to their observed replication p-value. We then used each variant’s power to replicate within its study to predict the replication rate within each decile. Note that, as we used deciles here, the observed and expected values should both be 10%, within confidence bounds and rounding error.

Across all variants, the predicted replication rate per bin was not significantly different from 10%, as expected, with the notable exception of the highest decile: the strongest signals tended to replicate significantly less than predicted (see Figure 2.7). This deviation primarily explains why the entire partition into deciles was significantly different than expected ($\chi^2$ goodness of fit $p < 10^{-4}$). As before, when restricting analysis to same-ancestry replication and reporting per-locus N (see Figures 2.8 through 2.11 for other subsets), replication rates became consistent with prediction, both jointly across all decile bins ($p = 0.67$) as well as within each (Table 2.5). Again, this is not simply lack of power to reject fit: the reduction in significance is beyond random expectation ($p < 0.01$). Several other partitions of the data approached good fit (Table 2.5), but no more than was expected due to reduction in power (all $p > 0.05$).

As before, when restricting analysis to same-ancestry replication and reporting per-locus N (see Figure 2.8 for all papers, Figure 2.9 for same ancestry studies, Figure 2.10 for different ancestry studies, and Figure 2.11 for the subset of same ancestry studies considering European samples), replication rates became consistent with prediction, both jointly across all decile bins ($p = 0.94$) as well as within each (Table 2.5). Again, this is not simply lack of power to reject fit: the reduction in significance is beyond random expectation ($p < 0.01$). Several other partitions of the data approached good fit (Table 2.5), but no more than was expected due to reduction in power (all $p > 0.05$).

### 2.4.5 Functional Enrichment in Replicated Variants

Finally, we evaluated enrichment of functional annotations in detected and replicated variants. We restrict this analysis to 56 papers which imputed their discovery samples using the HapMap2 CEU reference panel. Variants in the CEU reference provide a null distribution for functional annotation. Amongst all 998 loci for which replication was attempted in these papers, the observed 29 nonsynonymous variants constitute 5X enrichment compared to expectation from HapMap2 (expected 6
Figure 2.7: Expected and observed rates of replication in replication deciles. All variants are sorted by replication p-value and partitioned into deciles; we then compute power to replicate the variants in each bin using effect estimates with or without the Winner’s Curse. Left panel (A): including all papers (WC-corrected $\chi^2$ goodness of fit $p < 10^{-4}$); right panel (B): including only papers conducting discovery and replication in the same continental ancestry per variant and reporting accurate per-locus $N$ (WC-corrected $\chi^2$ goodness of fit $p = 0.67$). Improvement of fit exceeds what is expected due to loss of power from subsetting data ($p < 0.01$). X-axis: upper p-value boundary of bin; Y-axis: predicted fraction of replication within corresponding bin based on power estimated from discovery data. Tracks correspond to predicted power to replicate using raw discovery (red) or WC-corrected (teal) effect estimates. Error bars correspond to 95% confidence intervals around mean replication rates as estimated across multiple loci.
<table>
<thead>
<tr>
<th>Ancestry</th>
<th>N Reported</th>
<th>Papers</th>
<th>Loci</th>
<th>$\chi^2 p$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>all</td>
<td>100</td>
<td>1652</td>
<td>&lt; .0001</td>
<td>.17</td>
<td>.15</td>
<td>.18</td>
<td>.058</td>
<td>.0014</td>
<td>.15</td>
<td>.22</td>
<td>.038</td>
<td>5.6E-4</td>
<td>7.4E-07</td>
</tr>
<tr>
<td></td>
<td>per-locus N</td>
<td>43</td>
<td>836</td>
<td>.0053</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>max N</td>
<td>57</td>
<td>816</td>
<td>&lt; .0001</td>
<td>.0072</td>
<td>.17</td>
<td>.81</td>
<td>.35</td>
<td>.02</td>
<td>.017</td>
<td>.1</td>
<td>.32</td>
<td>6.5E-5</td>
<td>1.8E-06</td>
</tr>
<tr>
<td>Same</td>
<td>all</td>
<td>87</td>
<td>1269</td>
<td>&lt; .0001</td>
<td>.02</td>
<td>.042</td>
<td>.16</td>
<td>.16</td>
<td>.21</td>
<td>.6</td>
<td>.92</td>
<td>.018</td>
<td>.0037</td>
<td>2.8E-04</td>
</tr>
<tr>
<td></td>
<td>per-locus N</td>
<td>39</td>
<td>707</td>
<td>.94</td>
<td>.41</td>
<td>.54</td>
<td>.13</td>
<td>.84</td>
<td>.3</td>
<td>.79</td>
<td>.74</td>
<td>.095</td>
<td>.51</td>
<td>.33</td>
</tr>
<tr>
<td></td>
<td>max N</td>
<td>48</td>
<td>562</td>
<td>&lt; .0001</td>
<td>2.1e-4</td>
<td>.011</td>
<td>.075</td>
<td>.051</td>
<td>.11</td>
<td>.22</td>
<td>.77</td>
<td>.091</td>
<td>1.4E-4</td>
<td>2.0E-15</td>
</tr>
<tr>
<td>Different</td>
<td>all</td>
<td>13</td>
<td>383</td>
<td>.043</td>
<td>.12</td>
<td>1.0</td>
<td>1.0</td>
<td>.15</td>
<td>.012</td>
<td>.41</td>
<td>.24</td>
<td>.45</td>
<td>.12</td>
<td>.0089</td>
</tr>
<tr>
<td></td>
<td>per-locus N</td>
<td>4</td>
<td>129</td>
<td>&lt; .0001</td>
<td>.88</td>
<td>.0082</td>
<td>.15</td>
<td>.25</td>
<td>.5</td>
<td>.15</td>
<td>.063</td>
<td>.78</td>
<td>1.0</td>
<td>3.0E-16</td>
</tr>
<tr>
<td></td>
<td>max N</td>
<td>9</td>
<td>254</td>
<td>.019</td>
<td>.018</td>
<td>.022</td>
<td>1.0</td>
<td>.012</td>
<td>.23</td>
<td>.17</td>
<td>.45</td>
<td>.13</td>
<td>1.0</td>
<td>.057</td>
</tr>
<tr>
<td>European</td>
<td>all</td>
<td>60</td>
<td>976</td>
<td>&lt; .0001</td>
<td>.14</td>
<td>.019</td>
<td>.86</td>
<td>.033</td>
<td>.19</td>
<td>.22</td>
<td>.54</td>
<td>.11</td>
<td>.16</td>
<td>.0079</td>
</tr>
<tr>
<td></td>
<td>per-locus N</td>
<td>29</td>
<td>646</td>
<td>.371</td>
<td>1.0</td>
<td>.88</td>
<td>.21</td>
<td>.59</td>
<td>.14</td>
<td>.35</td>
<td>.09</td>
<td>.45</td>
<td>.94</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td>max N</td>
<td>31</td>
<td>330</td>
<td>&lt; .0001</td>
<td>.022</td>
<td>.0068</td>
<td>.023</td>
<td>1.0</td>
<td>.49</td>
<td>.14</td>
<td>.076</td>
<td>1.0</td>
<td>.045</td>
<td>7.3E-09</td>
</tr>
</tbody>
</table>

Table 2.5: **Probability of observing true replication counts within decile bins according to WC-corrected power estimates.**

Data are partitioned by ancestry matching between discovery and replication as well as reporting of per-locus or max sample size. First three data columns are: number of papers in category, number of (independent) loci in category, $\chi^2$ goodness of fit p-value for nominal replication in subset of papers. For graphical example of decile bins, see Figure 2.7. Null hypothesis for decile p-values: power within decile according to WC-corrected replication effect estimates is true generative probability for each variant, according to Poisson binomial distribution (Methods). Bins for which null is likely false (Poisson binomial two-sided test $p < 0.01$) are shaded.


Figure 2.8: Expected and observed rates of replication at different replication thresholds, across all 100 papers, 1652 independent loci. X-axis: left boundary of bin on replication threshold; Y-axis: percentage of replication within corresponding bin. Tracks correspond to actual data in the replication study, predicted power to replicate using discovery effect estimates, and predicted power to replicate using WC-corrected discovery effect estimates. Loci brought forward to replication in papers were pruned to contain independent signals within (but not necessarily between) individual papers. Error bars correspond to 95% confidence intervals around mean replication rates as estimated across multiple loci. Left panel: including all papers; middle panel: including only papers with accurate per-locus sample sizes; right panel: including only papers with maximal sample sizes per variant. Including only maximum sample size regardless of per-variant missingness will uniformly inflate replication rate estimates.

Figure 2.9: Expected and observed rates of replication at different replication thresholds, across 87 papers conducting discovery and replication in the same ancestry, 1269 independent loci. X-axis: left boundary of bin on replication threshold; Y-axis: percentage of replication within corresponding bin. Tracks correspond to actual data in the replication study, predicted power to replicate using discovery effect estimates, and predicted power to replicate using WC-corrected discovery effect estimates. Loci brought forward to replication in papers were pruned to contain independent signals within (but not necessarily between) individual papers. Error bars correspond to 95% confidence intervals around mean replication rates as estimated across multiple loci. Left panel: including all papers; middle panel: including only papers with accurate per-locus sample sizes; right panel: including only papers with maximal sample sizes per variant. Including only maximum sample size regardless of per-variant missingness will uniformly inflate replication rate estimates.
Figure 2.10: Expected and observed rates of replication at different replication thresholds, across 13 papers conducting discovery and replication in different continental ancestries, 383 independent loci. X-axis: left boundary of bin on replication threshold; Y-axis: percentage of replication within corresponding bin. Tracks correspond to actual data in the replication study, predicted power to replicate using discovery effect estimates, and predicted power to replicate using WC-corrected discovery effect estimates. Loci brought forward to replication in papers were pruned to contain independent signals within (but not necessarily between) individual papers. Error bars correspond to 95% confidence intervals around mean replication rates as estimated across multiple loci. Left panel: including all papers; middle panel: including only papers with accurate per-locus sample sizes; right panel: including only papers with maximal sample sizes per variant. Including only maximum sample size regardless of per-variant missingness will uniformly inflate replication rate estimates.

Figure 2.11: Expected and observed rates of replication at different replication thresholds, across 60 papers conducting discovery and replication in individuals of European grandparental ancestry, 976 independent loci. X-axis: left boundary of bin on replication threshold; Y-axis: percentage of replication within corresponding bin. Tracks correspond to actual data in the replication study, predicted power to replicate using discovery effect estimates, and predicted power to replicate using WC-corrected discovery effect estimates. Loci brought forward to replication in papers were pruned to contain independent signals within (but not necessarily between) individual papers. Error bars correspond to 95% confidence intervals around mean replication rates as estimated across multiple loci. Left panel: including all papers; middle panel: including only papers with accurate per-locus sample sizes; right panel: including only papers with maximal sample sizes per variant. Including only maximum sample size regardless of per-variant missingness will uniformly inflate replication rate estimates.
loci; $p < 0.0001$ based on 10000 simulated resamplings of random variants matched on count and minor allele frequency). This is due to significant enrichment of genic SNPs amongst all replication candidates ($3.6X$, $p < 0.0001$), as well as an additional enrichment of nonsynonymous variants among them ($1.5X$, $p = 0.0003$). Variants reaching per-paper Bonferroni replication are further $1.8X$ enriched in nonsynonymous exonic variants, from $2.9\%$ across 998 attempted variants to $5.2\%$ in 443 replicated ones (Binomial test one-tailed $p = 0.0061$). This change is due to enrichment of exonic SNPs in replicated variants, with no further significant selection for functional variants ($p = 0.37$). These results are not being driven by particular outliers ($\chi^2$ goodness of fit $p = 0.44$; Methods). Analogous enrichment among nominally-replicated variants ($1.3X$) is not significant (Binomial one-tailed $p = 0.1447$).

2.5 Discussion

This study provides the first systematic evidence of the efficacy of internal replication in the field of quantitative trait genome-wide association studies. Overall, with important caveats, we find that the field as a whole publishes results that replicate in a manner consistent with their expected power to replicate; this seemingly argues against the possibility of systematic flaws in GWAS methodology. The two significant predictors of aberrant replication performance, beyond the Winner’s Curse itself, are (1) incorrectly reporting maximum sample size instead of per-variant sample sizes, reflecting locus-specific missingness; and (2) conducting replication in samples of different continental ancestry than those used in discovery. Corresponding to reporting error and linkage disequilibrium effects, these influences are not surprising. Yet we have shown (Figure 2.5, Figure 2.6) that, within the papers considered here, these factors are necessary and sufficient to explain all internal replication discrepancies. This result is both novel and reassuring.

Though we present data separately for papers violating one of the two consistency conditions (see Figure 2.9, Figure 2.10, Table 2.5), we do not present extensive analysis or conclusions for these substrata. Unfortunately, the number of papers in each bin becomes quite small, in particular the mere four papers with different ancestries in discovery and replication but correct per-locus sample sizes. With such small counts, given the large paper-level heterogeneity we observe in Figure 2.2, we hesitate to draw conclusions about these subsets. This prevents direct evaluation of the relative importance of ancestry and sample size in replication prediction; in practice, it is likely variable,
Figure 2.12: Chromosomal distribution of loci from 100 papers used in this study. Includes all loci brought forward from discovery, not filtered for independence. Data point shapes correspond to approximate continental ancestry: CEU=European, YRI=(sub-Saharan West African, ASN=East Asian. Individual sites are stacked horizontally when the same variant is tested in multiple contexts. Distribution of variants across chromosomes is nonrandom relative to distribution of SNPs in reference datasets ($p < 0.00001$). Generated using the software at [42].
Figure 2.13: Chromosomal distribution of replicating loci from 100 papers used in this study. Subset of data from Figure 2.12; includes only loci passing study-specific Bonferroni-corrected replication threshold. Generated using the software at [42].
dependent on the rate of missingness in a given study and the relative divergence between the ancestries considered.

Several strategies have been developed for accounting for the Winner’s Curse in reporting of signals. The use of multiple stage GWAS, in which samples are conceptually partitioned into (possibly several) “discovery” and “replication” phases for internal replication, may be considered an attempt at removing positive bias in effect estimates. The discovery samples are used to reduce the pool of candidate SNPs from $\sim 10^6$ to $\sim 10^1 - 10^3$, at which point replication samples are used to verify that the selected SNPs maintain their direction and approximate magnitude of effect in an independent sample. Unfortunately, in many studies that make use of the discovery and replication partition, the final reported results are not solely based on the replication sample. Most commonly citing the argument in [43], studies frequently meta-analyze effect estimates from discovery and replication for a given SNP. This joint estimate maintains the benefits of prioritization by discovery, namely in reducing the cost of the study by minimizing the number of variants assayed in the replication samples. However, this estimate incorporates the probabilistically biased estimate from discovery, possibly attenuated by the less-biased estimate from replication. Thus while the argument of [43] holds, stating that meta-analysis of two-stage studies maximizes power to detect variants, this increase in power comes at the cost of both increased false positive rate and significant bias in the estimate of effect at true, detected signals.

Our selection of the Winner’s Curse correction method of [18] is based on two considerations. First, and most importantly, we lack access to the raw genotype data behind the loci we consider, as is required in [20]. Moreover, the number of variants reported in each study is unpredictable: in some cases there is just a single variant reported, whereas in others the investigators considered several hundred. In the case of the former, we cannot reliably apply methods like [44], which intrinsically require summary data from a large number of the most associated variants in a study to generate an effect estimate distribution. Both of these limitations strongly call for a method such as [18], in which correction is applied individually to each variant using only summary data. Yet this remains a limitation of our study, as we are not practically able to evaluate alternate methods of WC correction.

This study exclusively addresses the performance of internal replication in quantitative trait GWAS. Yet other forms of replication are of just as much importance to the field. External replication, in which the results of one study are tested by independent investigators, is an important
metric of the reliability of a field. In other contexts, external replication is known to perform at much lower rates than internal replication, suggesting various forms of bias. There is furthermore the consideration of functional replication, here broadly meaning the extent to which meaningful biological insights can be derived from GWAS data. Both of these forms of replication are largely unaddressed by the current study, nor indeed are they considered by the majority of GWAS publications; this does not diminish their importance. It is entirely possible that our results concerning the correct performance of internal replication may coexist with extremely low rates of external replication. Yet internal replication itself remains an important component of the field, and one in need of proper characterization. This study emerged from a discussion of the performance of internal replication in GWAS, when we discovered that there were no available data to prove one way or the other whether the internal replication model was effective in practice. We hope that our analysis provides one measure of reassurance about the fundamental reliability of the GWAS model.

Perhaps the most unusual observation of this analysis is the substantial proportion of manuscripts in the field that do not provide enough information to actually allow independent validation of their results. While some of the filters applied in our QC pipeline were present simply for ease of modeling, at least 58% of papers we collected failed to include the minimal amount of reporting to fully prove their claims of replication. This situation is a failure both in data reporting by authors and by peer review in journals. Combined with variable definitions of replication, we suggest this accidental lack of transparency substantially contributes to perceived unreliability of statistical genetics within other scientific disciplines. A higher standard of reporting, that will not only enable computation of unbiased effect estimates, but also list them explicitly, may be beneficial for the field.

The collection of these studies provides metrics other than simply replication rate. A primary example is the tabulation of ancestries studied in each analyzed paper (Table 2.4). We observe, as expected, a significant overrepresentation of European ancestry samples across these studies (expected 13.3% based on global census data [37, 38], observed 88.1%). This is accompanied by a concomitant underrepresentation in the other two major continental ancestries available, (East) Asian and (Northwest) African; in particular, African ancestry samples represent a mere 0.51% of samples analyzed in these papers, consisting of a Nigerian sample of approximately 1100 individuals, and a study of several hundred Moroccan Arabs. Note that we are deliberately making vast simplifications of minimally hundreds of distinct ancestry groups from all the countries of these continents, as this is consistent with the approximations used in the field.
I highlight these numbers in an attempt to provide metrics for what should be an apparent characteristic of the field to anyone who has read the literature over the past decade: human genetics is a field that studies whites. There are various explanations for these discrepancies in population coverage, which are largely beyond the scope of this document. Yet without doubt, the bulk of scientific knowledge about complex trait genetics is directly applicable to, maximally, 13% of the global population. We should not shy away from this number or make excuses for it, but rather directly contemplate the implications of this statistical and cultural bias for the coming years of medical genetics. In particular, any impact of personalized medicine through individual genome sequencing will be stratified first by existing racial bias in scientific genetics, and then furthermore not just by the cost of the sequencing itself but also by the cost of the treatment suggested by the sequencing. Existing race, gender, and class disparities in access and quality of medical care and for-profit medical insurance coverage will intersect [45] with genomics sampling biases to further restrict the utility of the scientific endeavor to a privileged elite.

We detect significant evidence for publication bias, the preferential publication of results based on their perceived quality. In particular, as shown in Figure 2.2, journals of higher impact factor, most notably Nature Genetics, published studies that replicated more variants than expected based on their statistical power; the inverse relationship holds as well. While this may intuitively suggest that the most robust results are published in the best journals, there is no intrinsic reason that the results of a study that replicates according to its statistical power distribution should be less robust than those of a study that replicates more often than should be possible. Rather, as is often the case with publication bias in other contexts, we raise concerns that the competitive publication of GWAS is giving rise to a publication record with invalid statistical properties, an important consideration that is not widely appreciated at this time.

Somewhat surprising is the lack of a clear ranking bias, in which studies combine variants at a locus according to strength of association, thus biasing each locus’s indicator SNP beyond the standard Winner’s Curse. This process, which has been termed “LD clumping,” is reasonably common, but was not consistently reported as used in the papers we analyzed. In some cases, papers reported data for all variants at a locus, and we implemented our own version of LD clumping, by randomly selecting a variant at each locus and discarding all other variants within a conservative physical distance. This process may have somewhat attenuated any ranking bias in this dataset; but it is quite likely that some of the fit deviation observed in our dataset is attributable to ranking
bias, yet is simply not strong enough to create systematic significant deviation at the granularity of the tests we apply here.

The indirect method of data collection used in this study raises several difficult questions concerning data consistency. Due to the sheer volume of papers analyzed in the course of this study, we must assume some errors are included within our data: both in the form of flawed data collection on my part, and mistaken reporting from the individual papers that was missed in both peer review and our manual inspection. Of particular note, for several tests included in this study, we have assumed for our statistical models that these papers report complete sets of loci brought to replication. We have furthermore specifically removed papers that transparently report partial subsets of results. However, without access to raw SNP lists from the contributing GWAS, there is no method to directly verify this criterion. There are also concerns about the low precision of data typically reported in GWAS publications. While we are able to make bulk conclusions across many loci, calculations at individual loci are somewhat unreliable. In particular, we have attempted to recompute the apparent sample size per variant for studies that have reported maximum sample size only; however, low precision data have made the resulting sample sizes rather unreliable, even generating in many cases sample sizes larger than the original value. Future analyses of this kind would strongly benefit from access to more of the raw data from contributing studies, should the resources be available for such an undertaking.

It is important to note that the restriction of analysis to quantitative (normally distributed) traits limits the direct conclusions we may draw to those same studies. This leaves the remainder of GWAS, which typically study binary disease traits and make up approximately 67% of the NHGRI-EBI GWAS Catalog. The methods used in [18] were initially developed for case/control studies and operate on regression test statistics, meaning the approach we have taken here may be easily applied to case/control studies in a later analysis. However, the practicalities of data collection meant that it was not possible to more than double our data acquisition for this study. We see no particular reason to assume different conclusions will be drawn based on binary trait studies, and suggest that our conclusions may provide a reasonable starting point for the interested analyst.

This study is not designed to counter-productively single out individual papers or investigators. For transparency, the full citation list is included (S1 File). I directly disclose my own statistics among considered papers. I contributed to 12 papers (3.6%) in the initial pool, two passing QC (consistent with expectation, Binomial given overall rejection rate across all papers, $p = 0.7751$).
Seven of the ten removed papers provided incomplete data for replication, more than expected by chance (Binomial given rate of this error across all papers, \( p = 0.007 \)). This anecdotal observation of papers focusing on anthropometric traits suggests the consistency of stylistic conventions within a phenotypic field to translate into recurrent faults in data reporting.

2.6 Conclusion

The Winner’s Curse correction algorithm used here is based on a simple and fast method of generating unbiased effect estimates [18]. Our implementation [46] requires simple input parameters (replication threshold, SNP frequency, etc.) available from studies in the field with no paper-specific modifications required. This tool models a traditional two-stage GWAS design, as opposed to a paradigm of merging data from both study stages [43]. While strict staging is less powerful in detecting true associations, meta-analyzing discovery and replication results in effect estimates still subject to directional bias from discovery, and is thus not considered in our software.

This analysis provides the first systematic evidence that quantitative trait association studies as a whole are replicable at expected rates. The fairly lenient quality control required to generate such a result is instructive: papers conducting discovery and replication in populations of similar ancestry and reporting accurate sample sizes replicate according to their predicted power. That these criteria are met in only 12% of all successfully published papers indicates intrinsic flaws, not in the paradigm of GWAS, but rather in study design and reporting standards. Correction of discovery effects provides distinct advantages for any GWAS study. Most fundamentally, replication at expected level is a sanity check for the analyst. Furthermore, WC-correction allows rational and optimal prioritization of variants for replication. Finally, as a field, it is critical for GWAS to report correct, rather than inflated results.

2.7 Methods

2.7.1 Notation

We consider \( M \) independent loci brought forward to replication from all papers combined. Each individual paper \( x \) contributes \( M_x \) loci to this total. A variant has an estimate of effect \( \beta_{\text{obs}} \) on a given phenotype as well as a standard error of that estimate \( s \), both computed from some form of
linear regression. Significance, either for bringing variants forward from discovery, or for considering variants successfully replicated, is defined based on a p-value threshold $\alpha$. The corresponding test statistic $\frac{\hat{\beta}_{\text{obs}}}{s}$ is standard normally distributed; $\phi, \psi$ are thus the PDF and CDF of the standard normal distribution, respectively.

### 2.7.2 Data Collection

The NHGRI-EBI GWAS Catalog [2, 3] is an online resource that collects certain annotations for all SNPs reported as significantly associated with a human trait. As significant association and successful peer review are the only major criteria used for inclusion in this database, we used it as a reasonably unbiased source of papers in the field across a variety of phenotypes and journals. We restricted the articles selected from the database to fit our modeling requirements as follows. The papers selected must primarily:

1. study at least one quantitative trait;
2. be published in a journal with a primary focus on human genetics;
3. provide for both discovery and replication:
   - regression effect;
   - regression standard error;
   - allele frequency; and
   - sample size
4. provide data for all variants brought forward to replication; and
5. model a minimally two-stage (discovery and replication) study design with a p-value threshold used to select variants for replication.

The full list of filters and papers lost due to each criterion is shown in Table 2.3. Whenever possible, we made reasonable accommodations to the papers to attempt to include them in this study. We consider variants novelly discovered in each paper, as opposed to those previously reported for a trait in question, as those are the variants typically brought forward for replication. Papers conducting multiple GWAS (i.e., reporting multiple phenotypes tested in the same study sample)
had all novel discovered variants from all traits included in the analysis, and are conservatively reported as a single unit in this analysis. For studies that reported a single allele frequency per variant, as opposed to a distinct frequency for each of discovery and replication stages, we used that one frequency for both stage instead. Studies that did not report a variant-specific sample size, to accommodate for differential missingness at different sites, were assigned the maximum available sample size assuming no per-site missingness. These modifications will introduce noise into the final analysis, yet a large percentage of papers required at least one of these modifications and thus were permitted in the interest of representation and sufficient sample size.

Studies with different replication designs were compelled whenever possible into the traditional two-stage format we use here. Thus for studies that attempted multiple non-tiered replications, followed by a meta-analysis of all discovery and replication panels together, we conducted the replication study meta-analysis manually using standard error weighting in METAL [47]. Studies that conducted tiered replications were included with the first tier replication, in which all variants passing a threshold from discovery were tested, used for their replication study.

### 2.7.3 Winner’s Curse Correction

To perform bias estimation, we use an implementation of the model in another study [18]. The major benefit of this model is that it may be applied to variant summary statistics as opposed to raw genetic data. As the non-parametric method BRsquared [20] requires raw genetic data, we did not consider this alternative. The maximum likelihood model we use is as follows:

\[
\beta_{\text{obs}} = \beta_{\text{true}} + s \frac{\phi \left( \frac{\beta_{\text{true}} - c}{s} \right) - \phi \left( -\frac{\beta_{\text{true}} - c}{s} \right)}{\psi \left( \frac{\beta_{\text{true}} - c}{s} \right) + \psi \left( -\frac{\beta_{\text{true}} - c}{s} \right)}
\]

Here, \( \beta_{\text{obs}} \) is the (likely biased) effect estimate observed in discovery; \( \beta_{\text{true}} \) is the conceptual underlying unbiased effect of the variant in the source population; and \( c \) is the test statistic corresponding to the discovery \( \alpha \) threshold in a given study. The expected bias of the observed effect, \( E[\beta_{\text{obs}} - \beta_{\text{true}}] \), scales inversely with the distance between the observed test statistic and the cutoff applied to variants brought forward to replication. The bias can be solved using any standard zero-finding algorithm (for example, Brent’s method as implemented in C [48]). Note that in situations in which the observed test statistic far exceeds the \( \alpha \) threshold, each component of the bias in the above equation is dominated by one or the other of the paired terms; only when the statistic is close...
to the threshold (that is, when the expected bias is large) do both terms meaningfully contribute to the bias estimate.

2.7.4 Independence of Loci

To simplify predictions of replication efficiency, we considered an independent subset of all reported loci. As we lack direct access to the genetic samples used in these studies, we extracted a subset of the variants such that no two variants in a paper are situated within one megabase of any other. This is a very simple modification of the standard clumping protocol used in GWAS studies [49]. To prevent additional bias, we report a random variant from each locus, not necessarily the most strongly associated in discovery. This will effectively guarantee that each variant represents a single locus with only minimal linkage disequilibrium between variants, but is conservative in the sense that it discards any secondary signals present among the replicated variants. Furthermore, this approach may attenuate functional annotation burden testing if the strongest association in an LD block is preferentially causal. While certain papers specifically address the possibility of secondary signals by sequential conditional analysis of variants, the inconsistency of this analysis and absence of it in many papers led us to seek a uniform treatment of all papers in this study.

2.7.5 Definition of Replication

The concept of “internal replication” may be interpreted differently in different reports. We consider three definitions of replication for this study, to observe different characteristics of the data:

1. replication at nominal $\alpha = 0.05$ [“nominal”]

2. replication at $\alpha = \frac{0.05}{M_{\text{paper}}}$ [“Bonferroni”]

3. replication within deciles of variants [“deciles”]

We specifically only consider methods in which replication is determined from the replication study alone. The nominal and Bonferroni methods are commonly used. We use the decile method to investigate the behavior of the predicted power [50] to replicate according to the strength of an association signal. We compute decile $\chi^2$ goodness of fit, using average power to replicate in each bin across all variants. This permits a formal analysis of differential performance of replication at different levels of replication stringency.
Given a set of variants and their predicted power to replicate at a given $\alpha$ threshold, the number of observed replications is distributed as Poisson binomial with success probabilities equal to each individual variants’ power to replicate (see below). This is a generalization of a Binomial distribution in which each Bernoulli trial is allowed to have a known but variable success rate. We use the implementation of this distribution in R [51]. We further adapt the standard two-tailed Binomial test for use with the Poisson binomial CDF implemented in this package.

We note that under certain assumptions the number of replications will asymptotically be distributed normally. However, depending on the $\alpha$ considered, many variants analyzed here have power of effectively, or within machine precision, 0 or 1; with our limited sample size, the convergence properties of our dataset will be undesirable, and thus we use the exact distribution at the cost of computational efficiency. This process may be considered a fitting of the model according to which the WC-corrected discovery data correctly explain the observed replication data.

In several instances, we evaluate the effects of filtering certain subsets of papers based on various criteria, and the extent to which this causes fit criteria to return to null expectation. As this evaluation is potentially confounded by reduced statistical power, in all cases we test whether the change in p-value is significantly different from expectation under random subsampling of variants matched on total power to replicate amongst the observed variants.

### 2.7.6 Power to Replicate

Assuming the discovery and replication sample of a study are drawn from the same source population with shared expected effect at each variant, the power to replicate a discovered variant $v$ for a quantitative trait under the additive model is

$$\text{power}(\alpha, \text{ncp}_v) = 1 - \chi^2_1((\chi^2_1)^{-1}(1 - \alpha), \text{ncp}_v)$$

In brief, the power to detect a signal at an $\alpha$ threshold of $p$ is the probability of the variant exceeding the required test statistic from the null, but under the alternative distribution which is noncentral $\chi^2_1$ with per-variant noncentrality parameter.
\[ ncp_v = \frac{\text{variance explained}}{\text{residual variance}} \approx N \frac{2\beta^2_{\text{obs}} f(1 - f)}{\text{trait variance} - 2\beta^2_{\text{obs}} f(1 - f)} \]

where \( N \) is the replication sample size and \( f \) is the replication allele frequency of the variant. This \( f \) should be the actual allele frequency in the replication sample; however, studies sometimes report \( f_{\text{hapmap}} \) from the closest reference ancestry as a means of protecting patient anonymity. Overall, the predicted number of replications across all variants in a paper is the sum of the power to replicate, as a function of predicted effect size and replication sample size and frequency, across all variants analyzed.

### 2.7.7 Poisson Binomial Distribution

The standard binomial distribution represents the probability of observing \( k \) successes in \( n \) trials with a fixed success probability \( p \); in other words, the sum of \( n \) Bernoulli trials with a common success probability. The Binomial probability density function is thus

\[ B(k, n, p) = \binom{n}{k} p^k (1 - p)^{n-k} \]

The Poisson binomial distribution is the extension of the Binomial distribution to a series of Bernoulli trials with arbitrary success probabilities at each trial. Thus the probability density of this distribution is defined as follows. Consider each \( S \) of the \( n! \) permutations of the integers \( 1, 2, \ldots, n \). Then the pdf is

\[ PB(k, n, p_1, \ldots, p_n) = \sum_S \left[ \prod_{i=1}^{k} p_{S(i)} \prod_{i=k+1}^{n} (1 - p_{S(i)}) \right] \]

In other words, in the place of the standard Binomial coefficient, one must iterate through all possible sets of \( k \) integers on the sequence \([1, n]\). This probability is not closed and poses serious computational issues; this study uses a publicly available implementation of the density function in R[52, 51]. The standard two-tailed Binomial test in R was then manually modified to function with
2.7.8 Sample Ancestries

These papers demonstrate the coverage of population ancestries in the field of quantitative trait genetics. We report and analyze the ancestral coverage of these studies using the simplifying summary statistic of continent of ancestry (Europe, Africa, East Asia), tracing generally the ancestries of the original HapMap2 populations. We include a fourth category for African American samples, the largest admixed population nonnegligibly represented in the papers. This geographical partitioning matches the ancestry assumptions used in GWAS methods such as genotype imputation.

Ancestry group counts are computed from maximum reported sample size per cohort per paper. In studies where cohorts of different continental ancestry are meta-analyzed, sample sizes are appropriately partitioned to the contributing ancestries. No adjustment is applied for papers reporting on the same cohort. For comparison to what the field’s sample sizes would be under random global sampling, global population estimates are computed [37, 38].

2.7.9 Functional Annotation

We tested loci for nonrandom annotations. This test is usually conducted with access to the full set of variants tested in an individual study. As in this study design such information is masked, we restricted the analysis to papers using HapMap2 imputation in their discovery data; considered only SNPs present in HapMap2; and restricted the data further to European ancestry discovery data, which includes the majority of papers in the dataset.

We annotated all variants in the CEU subset of HapMap2 using ANNOVAR [53]. We computed the average rate of functional annotations in the true set of variants. To generate a null distribution, we matched true variants on allele frequency and, when appropriate, whether the variant was located in an exon. P-values are computed over 10000 simulated null sets.
Chapter 3

Bias Characterization in Probabilistic Genotype Data and Improved Signal Detection with Multiple Imputation

3.1 Abstract

Missing data are an unavoidable component of modern statistical genetics. Different array or sequencing technologies cover different single nucleotide polymorphisms (SNPs), leading to a complicated mosaic pattern of missingness where both individual genotypes and entire SNPs are sporadically absent. Such missing data patterns cannot be ignored without introducing bias, yet cannot be inferred exclusively from nonmissing data. In genome-wide association studies, the accepted solution to genotype missingness is to impute missing data using external reference haplotypes. The resulting probabilistic genotypes may be analyzed in the place of genotype calls. A general-purpose paradigm, called Multiple Imputation (MI), is known to model uncertainty in many contexts, yet it is not widely used in association studies. Here [54], we undertake a systematic evaluation of existing imputed data analysis methods and MI. We characterize biases related to uncertainty in association studies,
and find that bias is introduced both at the imputation level, when imputation algorithms generate inconsistent genotype probabilities, and at the association level, when analysis methods inadequately model genotype uncertainty. We find that MI performs at least as well as existing methods or in some cases much better, and provides a straightforward paradigm for adapting existing genotype association methods to uncertain data.

3.2 Summary

Genetic research has been focused on analysis of datapoints that are assumed to be deterministically known. However, the majority of current, high throughput data is only probabilistically known, and proper methods for directly handing such uncertain genotypes are limited. Here, we build on existing theory from the field of statistics to introduce a general framework for handling probabilistic genotype data obtained through genotype imputation. This framework, called Multiple Imputation, matches or improves upon existing methods for handling uncertainty in basic analysis of genetic association. As opposed to such existing methods, our work furthermore extends to more advanced analysis, such as mixed-effects models, with no additional complication. Importantly, it generates posterior probabilities of association that are intrinsically weighted by the certainty of the underlying data, a feature unmatched by other existing methods. Multiple Imputation is also fully compatible with multi-cohort meta-analysis. Finally, our analysis of probabilistic genotype data brings into focus the accuracy and unreliability of imputation’s estimated probabilities. Taken together, these results substantially increase the utility of imputed genotypes in statistical genetics, and may have strong implications for analysis of sequencing data moving forward.

3.3 Introduction

Genome-wide association studies (GWAS) have become a primary tool to elucidate the correlations between SNP genotypes and complex phenotypes in large cohorts. Association studies initially assumed the existence of genotype calls: for each sample at each assayed variant, either reference allele homozygote, heterozygote, or alternate allele homozygote. As such, the methods developed for analyzing GWAS also assumed the existence of such perfect-confidence genotype data. The association study design and related analysis methods have remained in force even as the field
has transitioned into the sequencing era and more complete data have become available. Yet in all situations, due to technical and financial limitations, association studies only partially assay the set of common variants in any organism. The variants included on a SNP array typically only include a small fraction of the total pool of variants present, and even sequenced variants are called incompletely and inconsistently. Furthermore, due to the low magnitude of effect of most trait-associated variants, studies prioritize sample size via multisite meta-analysis, involving genetic samples assayed on different technologies. This study design results in a complicated missingness pattern across the entire conceptual set of common variants in a sample. Yet due to shared linkage disequilibrium between different samples, this missingness can be overcome with the addition of external reference data.

Genotype imputation probabilistically estimates unknown genotypes for a study sample by leveraging external reference haplotypes ascertained at a superset of SNPs [21]. Genotype calls and genotype probabilities are fundamentally different. Genotype calls are considered certain data and, under the traditional additive model, may be represented as an integer count of a reference allele present for each study individual. Unfortunately, statistics developed for such a genotype model cannot be directly extended to probabilistic data, as they are incapable of representing the variance component introduced by uncertainty. This is critical to understanding the traditional challenge in analyzing imputed genotype data: existing methods are not directly compatible with uncertain data, so probabilistic genotypes must be projected to a lower-dimensional approximation resembling genotypes, with a concomitant loss of information and introduction of bias.

Three primary methods have been developed for the handling of uncertain data in genetic association studies. For the first method, and at one extreme, probabilities may be converted to call-like integral counts by choosing the genotype with the largest probability. This paradigm requires no additional modification of the analysis method for genotype calls, but almost all meaningful information about uncertainty is lost. For the second method, in some situations, genotype probabilities may be converted into expected counts of the reference allele: the “allelic dosage.” This strategy attempts to maintain some of the uncertainty of the genotype estimate by allowing non-integral values: for example, when the heterozygote and reference allele homozygote classes are equiprobable and the alternate allele homozygote has 0 probability, a sample is considered to have 1.5 alleles. Unfortunately, this strategy is only useful when the underlying algorithm extends to nonintegral data (for example, a generalized linear model with continuous predictor); furthermore, there is no
rigorous proof of the degree of bias or information loss incurred using this method. Finally, as a last method, the desired statistic may be modified to directly operate on probabilities [22]. This option attempts to include all uncertainty information at the cost of additional work creating a new algorithm. In practice, this type of custom algorithm design is limited to simple GLM methods; other studies in the field creating more complex statistical models for association studies do not undertake this additional work [55, 56, 57, 58, 59, 60].

In the statistics literature, Multiple Imputation (MI; distinct from genotype imputation; [23]) is the rigorous method of conducting analysis on probabilistic estimates of uncertain data. The details of MI are discussed in Methods. Briefly, a small number of complete datasets are randomly drawn according to and consistent with the probability data. These datasets are then analyzed using any standard analysis technique that asymptotically generates a normally-distributed effect estimate $\beta$ and standard error estimate $s$. Importantly, MI is not a method of estimating hidden data, but rather a method of handling existing estimates. The performance of MI, and indeed of all imputed data analysis methods, is reliant on the quality of the underlying genotype imputation. Imputation accuracy, the agreement between predicted and true genotype, tends to vary across both imputation “quality,” as estimated by most imputation software, and minor allele frequency. The extent of this variable performance has not, to our knowledge, been rigorously assessed.

In this study, we seek to rigorously evaluate this variable genotype imputation performance. We show a significant deviation between genotype probabilities generated by imputation, and empirical probabilities estimated at the same sites. This failure of probability consistency is an important confounding effect in imputed data analysis. We show that Multiple Imputation matches or improves upon performance of existing imputed data analysis regimes by better prioritizing true positive associations, while additionally being straightforwardly extensible to future analysis algorithms.

3.4 Results

3.4.1 Imputation Accuracy by Genotype, Frequency, and Quality

We compared the allelic dosage (Methods) to the “true” genotype count based on masked genotype data. Figure 3.1 shows the fraction error between allelic dosage from imputation and masked genotype, stratified by genotype class, allele frequency, and reported imputation quality metric (hereafter called “$r^2$”). We observe that a single quality metric $r^2$ masks significant deviations in mean
Figure 3.1: Relationship between quality of estimated allelic dosage from IMPUTE2 imputation and predictors of imputation quality. Data are estimated from 10% of the original chip (59808 SNPs) masked from imputation. Discordance of predicted allelic dosage (Y-axis) is the fraction difference between dosage computed from imputation probabilities and dosage based on masked genotype data: for example, if the true genotype is reference homozygote and the allelic dosage from imputation is 1.4, the discordance is |2 − 1.4| / 2 = 0.3. Left panel: imputation quality greater than 0.9; right panel: quality between 0.8 and 0.9. Clusters correspond to minor allele frequencies of 10%; individual bars represent quality stratified by masked genotype. “Predicted” bars correspond to expected concordance assuming independence of individual haplotypes. Error bars represent 95% confidence intervals of mean discordance estimate.

quality between different genotype classes and different allele frequencies. For imputed SNPs reported to be of high quality ($r^2 > 0.9$, left panel; top 40% of GWAS-used SNPs, Figure 3.1 and Figure 3.3), and with sufficiently high minor allele frequency, imputation is indeed well-behaved: variants with minor allele frequency above 0.3 have less than 3% discordance with similar performance across genotype classes. However, SNPs of lesser imputation quality (0.8 ≤ $r^2 ≤ 0.9$, right panel; approximately 18% of GWAS-used SNPs, Figure 3.1 and Figure 3.3) or low minor allele frequency are inconsistently imputed. Minor allele homozygotes and heterozygotes in particular are subject to highly inflated error rates. While differences across the SNP strata are observed with both IMPUTE2 and minimac3 imputations, the magnitudes observed are distinct (Figure 3.2 and Figure 3.4). Equivalent results are observed when evaluating performance by fraction of best-guess genotypes from imputation not matching masked genotypes (Figure 3.5).
Figure 3.2: **Relationship between quality of estimated allelic dosage from minimac3 imputation and predictors of imputation quality.** Data are estimated from 10% of the original chip masked from imputation. Discordance of predicted allelic dosage (Y-axis) is the fraction difference between dosage computed from imputation probabilities and dosage based on masked genotype data: for example, if the true genotype is reference homozygote and the allelic dosage from imputation is 1.4, the discordance is $\frac{|2-1.4|}{2} = 0.3$. Left panel: imputation quality greater than 0.9; right panel: quality between 0.8 and 0.9. Clusters correspond to minor allele frequencies of 10%; individual bars represent quality stratified by masked genotype. Error bars represent 95% confidence intervals of mean discordance estimate.

Figure 3.3: **Distribution of imputation quality from IMPUTE2 imputation.** X-axis: IMPUTE2 info (quality) metric; Y-axis: proportion of full set of variants within this quality bin. Distribution is left-truncated at common quality threshold.
Figure 3.4: **Distribution of imputation quality from minimac3 imputation.** X-axis: minimac3 $r^2$ metric; Y-axis: proportion of full set of variants within this quality bin. Distribution is left-truncated at common quality threshold. Final bin with quality greater than 1 indicates small percentage of variants where empirical variance exceeds that of the expected binomial distribution.

Figure 3.5: **Relationship between quality of best guess genotypes from IMPUTE2 imputation and predictors of imputation quality.** Data are estimated from 10% of the original chip masked from imputation. Discordance of predicted genotypes (Y-axis) is the fraction of best guess genotypes for a given bin that do not match the corresponding masked genotype. Left panel: imputation quality greater than 0.9; right panel: quality between 0.8 and 0.9. Clusters correspond to minor allele frequencies of 10%; individual bars represent quality stratified by masked genotype. Error bars represent 95% confidence intervals of mean discordance estimate.
Figure 3.6: **Evaluation of the consistency of probability scores from IMPUTE2 imputation.** Data are estimated from 10% of the original chip (59808 SNPs) masked from imputation. X-axis: 0.02-width bins of imputation probabilities; Y-axis: mean deviation between expected and observed accuracy. Data series correspond to results stratified by genotype class. Left panel: IMPUTE2 info metric greater than 0.9; right panel: info metric between 0.8 and 0.9. Error bars represent 95% confidence intervals around mean consistency estimate.

### 3.4.2 Imputation Probability Consistency

We next examined the imputation probabilities themselves, to evaluate whether probabilities generated by imputation software correspond to the empirical probability of observing a genotype at a particular site. Results for this comparison for IMPUTE2 probabilities are shown in Figure 3.6, across strata of reported quality and predicted call probability. The empirical accuracy significantly deviates from the predicted, and much more so with decreasing $r^2$; Figure 3.7 shows similar plots comparing the effect of decreasing minor allele frequency on this distortion, and show weaker but significant changes with decreasing frequency. Of note, the heterozygote genotype class behaves in a distinct but complementary fashion relative to the two homozygote classes.

These results are distinguishable between imputation programs: the effect is much stronger in the IMPUTE2 imputation. The most substantial difference between the two programs is IMPUTE2’s use of sequential imputation windows to improve performance through parallelization, with potential accuracy tradeoffs, yet we have observed no differences caused by modifying this parameter. We note that differences in imputation performance under different conditions have been observed extensively.
Figure 3.7: Evaluation of the consistency of probability scores from IMPUTE2 imputation, stratified by allele frequency instead of imputation quality. Data are estimated from 10% of the original chip (59808 SNPs) masked from imputation. X-axis: 0.02-width bins of imputation probabilities; Y-axis: mean deviation between expected and observed accuracy. Data series correspond to results stratified by genotype class. Left panel: minor allele frequency greater than 0.4; right panel: minor allele frequency less than 0.1. Error bars represent 95% confidence intervals around mean consistency estimate.
and unpredictably (see, among many, [61, 62, 63, 64, 65]). Our observations are consistent with intermittent observations of MACH-family algorithms nominally outperforming IMPUTE-family algorithms in some cases, yet the precise reason(s) for these differences between similar software has never, to our knowledge, been demonstrated.

### 3.4.3 Signal Enrichment with Multiple Imputation

We next sought to evaluate whether the ability to prioritize verified trait-associated SNPs in an association study ranking was detectably different using MI or other existing algorithms. We considered 73 replicated loci from a large (N=339224 individuals) GWAS for BMI [66]. As expected with our modest sample size of 2802, we have little power to detect the majority of these variants at genome-wide significance $\alpha = 5 \cdot 10^{-8}$. Nevertheless, if the variants are associated with the trait at all, one expects the variants to be relatively better ranked in the final list of variants than variants chosen at random from the study. We conducted a BMI genome-wide association study in Health ABC, as discussed in Methods. Using these SNP association results, we computed the rank percentile of each published variant, comparing these results for two existing imputed data analysis algorithms and MI.

We evaluated the area under the receiver operating characteristic curves for PLINK dosage (for allelic dosage), SNPTEST score (for parametric use of full probabilities), and MI (Figure 3.8). MI significantly outperforms all other methods (one-tailed DeLong test $p < 2 \cdot 10^{-16}$). We repeat this analysis with a height GWAS in NFBC66 (Methods), and find a similar significant improvement in signal detection by MI compared to other methods (one-tailed DeLong test $p < 0.000376$). We conclude from this analysis that significant information loss may occur when uncertainty is incompletely handled in imputed data analysis. There is no evidence in this comparison that Multiple Imputation is ever inferior to other methods.

### 3.4.4 MI Changes the Null Distribution of Variants

Improved signal detection by MI may be attributed to several causes. In observing the additional variance component in the multiple imputation model, $s_B^2$ (see Methods), we note that variability introduced by genotypic uncertainty should result in decreased rankings for variants regardless of trait being analyzed. Uncertainty in probabilities also differs with variant-wide $r^2$, implying that diffe-
Figure 3.8: **Receiver operating characteristic curves for Health ABC BMI association with IMPUTE2 imputation.** True positive associations are 73 established BMI variants from [66]. Tracks correspond to uncertainty handling methods: allelic dosage (“Dosage”), score test (“SNPTEST”), MI on imputed probabilities. Statistics on thresholded genotypes are not included due to poor performance similar to that of “Dosage.”

rent association tests may lead to different expected distributions of variants in a ranked association study across the quality spectrum, whether variants are trait-associated or not.

We sought to evaluate the null distribution of variant ranks across imputation quality. We calculated the percentile rank of each variant in the Health ABC BMI GWAS and sorted the variants into bins across imputation quality. Results from this investigation are shown in Figure 3.9. All tested methods have a statistically significant correlation between imputation quality and average rank (all Pearson correlation test $p < 2 \cdot 10^{-16}$; comparable results for nonparametric tests). PLINK and SNPTEST have indistinguishable magnitudes of effect (shared effect size $-0.017$, indistinguishable with $p = 0.44$). Multiple Imputation shows a significantly stronger correlation than the other tests (effect size $-0.351$, greater than other tests with $p = 3.4 \cdot 10^{-7}$). MI produces test statistics that much more significantly incorporate uncertainty, through in particular the between-draw variance component.

Published trait-associated variants tend to impute with higher quality than random variants from a dataset due to selection bias, leading to the possibility that this global incorporation of quality by MI is leading to better performance in a manner unrelated to the trait itself. To control for trait-
unrelated differences between published variants and the global SNP distribution, we computed empirical matched null sets for every published variant by drawing 1000 SNPs matched to this variant with parameters $r^2 \pm 0.01$ and frequency $\pm 0.01$ of the true variant. Here we are drawing variants from the association study itself; this null allows us to detect effects specifically on the significant tail of the distribution over the bulk of variants. For most traits, these variants will overwhelmingly be unassociated with the outcome, and thus this null will correspond to a true null of no genetic association; in the case that this condition is untrue, this null will lead to relatively conservative assumptions about the added performance of MI. We then regenerated ROC curves for each analysis, this time controlling for the null rank of SNPs matched to the published variant list based on these parameters. These results approximately correspond to trait-specific enrichment effects caused by the various analysis methods. We find that adjustment for trait-secular shifts in variant quality affects BMI and height differently. For BMI, MI continues to outperform PLINK and SNPTEST (one-tailed DeLong $p < 1.177 \cdot 10^{-10}$). For height, null adjustment reverses the previously-observed trends, such that MI tends to underperform PLINK and SNPTEST (one-tailed DeLong $p < 0.0025$). We note that for this trait in particular, the null derived from drawing from the association study itself may be conservative given the broad genetic basis for human height [67].
3.4.5 Extensibility

To underscore the benefits of MI not simply on regression with existing uncertainty handling methods, but additionally on more complex algorithms more difficult to directly adapt for use with probabilities, we applied MI to EMMAX [56], a commonly used linear mixed model software package. We ran EMMAX on both the Health ABC BMI GWAS and the NFBC66 height GWAS, with default parameter settings and the IBS-based kinship matrices. We compared EMMAX with thresholded genotype data and MI on imputed probabilities. The published version of EMMAX only accepts integral count genotypes, thus no other comparisons were included in this test.

We detect significant improvement in relative percentile increase for BMI-associated variants that is abrogated by null adjustment (one-sided DeLong test \( p = 6.663 \cdot 10^{-12} \) and \( p = 0.09933 \) respectively). We find weak evidence for improved overall efficiency for known height variants (one-sided DeLong test \( p = 0.01994 \)), which under null adjustment becomes a significant underperformance of MI relative to rounded genotype control (one-sided DeLong test \( p = 2.242 \cdot 10^{-7} \)); again, we note that this null may be overly conservative in this trait context [67]. Evaluation of the null distribution of variants for this experiment, analogous to the analysis in Figure 3.9, shows that the MI-mediated rank shift is much smaller across both the frequency and variant quality spectrum. We propose that the error model used in EMMAX is partially compensating for the additional uncertainty variance component redundantly modeled by MI.

3.5 Discussion

In this study, we analyze the application of Multiple Imputation in the particular context of genotype imputation. We find that existing analysis methods tend to increase noise in association testing by incompletely modeling the uncertainty of genotype estimates across the entire set of imputed variants. This results not just in lower quality variants, for which information is limited, receiving inappropriately high ranking; but also variants for which underpowered but significant trait association is present becoming lost in statistical noise introduced by imputation. We furthermore detect imputation program-specific inconsistencies in posterior genotype probabilities that differently affect the three genotype classes. Overall, we characterize the complexity of imputation probabilities and show that MI can improve association testing with uncertain data at the cost of increased post-imputation computational time.
This study particularly elaborates on the interpretation of a p-value from a probabilistic association study. Based on the results in Figure 3.9, we see that existing analysis tends to rank variants uniformly across $r^2$: the p-value weakly incorporates imputation quality, enough so that the tests are not correctly calibrated but insufficiently so to actually correct for differential uncertainty. This poor calibration is verifiable with a true null: observing the null distribution of variants in our study regressed against random standard normal traits, we observe a comparable correlation between variant imputation quality and SNP ranking for SNPTEST ($p = 0.001953$). Dosage analysis does not, in that context, recapitulate the correlation effect observed in our data, suggesting that the trend is specific to our trait-associated null. This poor calibration is potentially shared by many analysis methods in the field: following the simulation work of [68], we have found low magnitude but significant anticonservative miscalibration when operating on probabilistic genetic data and null traits for linear regression and a Kruskal-Wallis test on best-guess variants; dosage analysis; and their generalized Kruskal-Wallis test that directly handles uncertainty (all Wilcoxon rank sum $p < 2 \cdot 10^{-16}$). There is also a significant correlation between uncertainty for the Kruskal-Wallis based tests in this simulation context (ANOVA $p < 2 \cdot 10^{-16}$). Association p-values are intrinsically affected by genotype uncertainty, altering the expected null distribution from uniform random, but this is simply an uncharacterized bias introduced by incompletely handling uncertainty, not a formal characteristic with proven properties.

We suggest that this behavior is undesirable: if a SNP estimate is uncertain, then the behavior of an ideal inference should be to proportionally downweight the ranking of that SNP in that particular study in a predictable fashion. With Multiple Imputation, this uncertainty-induced variability is correctly apportioned to the between-draw variance component $s_B^2$, resulting in unmodified effect estimates that are still suitable for cross-cohort meta-analysis. In the case of our example BMI association study, the positive control variants used were all imputed with high confidence, yet MI still showed a trait-specific rank improvement in these SNPs, due to the selective downranking of low confidence variants with no or lesser phenotype association. In short, we see little reason to continue using dosage or score test methods for handling imputation probabilities, when a straightforward alternative with superior statistical properties exists.

The results in Figure 3.1 and Figure 3.6 broadly characterize imputation performance across variant quality, frequency, and genotype class. Although parts of these results are hinted at in various imputation papers (in particular, [69], but also [22, 70, 71, 63]), we have not before encountered a
presentation of the severity of quality degradation across the full spectrum of frequencies. We urge analysts in all contexts to avoid genotype thresholding, in particular using IMPUTE2 probabilities: with decreasing frequency and quality, the likelihood of rounding to an incorrect genotype grows extremely high. We also observe strong per-genotype class differences in performance, most severely affecting heterozygotes. This is not a surprising result, given the integral nature of computational prephasing in genotype imputation. In the modern genotype imputation protocol, prephasing is a one-time burdensome computation that is not repeated, whereas the imputation step itself is comparatively rapid and repeated many times as new reference panels become available. A useful future analysis would investigate the specific impact of differential phasing quality on imputation probability bias, and our results emphasize that prephasing should be reconducted as advancements in the field yield higher quality phasing solutions.

We have observed in the literature a tendency to ignore uncertainty in genotype data when developing new algorithms. This elision is understandable in the sense that complete-case analysis is typically substantially more straightforward to implement, and as a first approximation, uncertain data may be converted seemingly straightforwardly to genotype “calls” via techniques such as rounding. Yet in the case study of simple regression, we see that this approximation bears with it a cost in loss of statistical power to detect trait-associated variants. One of the great benefits of MI is its simple application to complete-case analytical tools, as each round of MI generates complete case data. In the case of standard regression, this is a straightforward benefit; in the case of a method such as EMMAX, one must balance the desirable avoidance of rounded probabilities with potential complications of the standard MI variance component model.

One of the strongest justifications for the widespread use of imputation is the facilitation of multisite multicohort meta-analysis, in which summary statistics from separate association studies are combined to increase statistical power. The benefits of integrating imputation quality into inference are magnified in this context. Under the current regime, either \( \{\beta, s^2\} \) or \( \{N, p\} \) pairs are the exclusive data provided to meta-analysis tools [47], leading to a downstream analysis that treats different estimates of variants with different imputation qualities as estimates identical to one another in expectation ascertained with perfect confidence. With nested model multiple imputation ([72], Methods), the improved model described in this study may be extended to meta-analysis. A nested model MI meta-analysis will explicitly compensate for variable imputation quality across contributing studies with inflated variance components corresponding to noise in the mean effect.
estimate or inflated standard error (i.e., imprecision in effect estimate introduced either by variable imputation quality, fluctuations in allele frequency, or differential effect due to LD changes in different studies). The cost of this extended meta-analysis regime is limited. In addition to the standard \( \{\beta, s^2\} \) pairs currently used to combine analysis, the within- and between-draw variance components must also be submitted. These data require additional disk space but are relatively trivial to provide. Crucially, the summary statistics from each individual draw are not required for this meta-analysis, such that the growth in memory requirements for meta-analysis does not scale with the number of draws conducted.

We note that by explicitly handling variable imputation quality in different studies, this meta-analysis regime introduces potential sources of heterogeneity in the final meta-analysis result. Studies with aberrantly high uncertainty may strongly influence the resulting meta-analyzed association statistic. Note that this source of heterogeneity already exists, but is not rigorously modeled and currently must be addressed by ad hoc filtering and quality control. We propose that heterogeneity from differential imputation quality may be quantified by a heterogeneity test analogous to the effect estimate heterogeneity \( I^2 \) metric in METAL [47]. This test would quantify heterogeneity specifically in the between-test variance component \( \hat{s}^2_B \) from contributing studies, which captures the noise between individual MI draws introduced by meaningfully uncertain data. This test would enable standardized detection of cases in which low imputation quality may require custom secondary genotyping for validation.

In the statistics literature, starting with [23] and moving forward, the recommended number of MI draws has varied widely. The original recommendation was that oftentimes 3–5 draws were sufficient to retain most of the accuracy while minimizing computational burden. More recent publications (i.e. [73]) have suggested that the original estimates of draws were insufficiently stringent. In this study, we have used an a priori setting of 10 draws after testing various draw numbers’ effects on quality of MI output, and balancing this impact with the added burden of multiple additional MI rounds. We find that this draw count surpasses the point at which MI effect estimates tend to converge (Figure 3.10), but may in the case of particularly poorly imputed variants lead to suboptimal estimates of between-draw variance (Figure 3.11). A useful target for future work would be the integration of dynamic computation of the number of draws for convergence for individual variants, which in this study is complicated by the manner in which our software interfaces externally with, rather than reimplements internally, existing analysis tools to maximize compatibility.
Figure 3.10: **Evaluation of the effect of Multiple Imputation draw count on estimated regression coefficients.** X-axis: number of MI draws; Y-axis: observed Pearson correlation coefficient between regression coefficient estimates from masked data and estimates from MI on imputed estimates over masked sites. Horizontal line corresponds to correlation between masked data and “best guess” genotypes using imputation probabilities.

Figure 3.11: **Evaluation of the effect of Multiple Imputation draw count on estimated between-test variance component.** X-axis: number of MI draws; Y-axis: absolute change in between-draw variance component from one MI round to the next. Note this is a traditional boxplot but the boxes are tightly clustered around y=0, leading to the boxes rendering as the small thick back lines for each x value at y=0. The emphasis of these results is on the distribution of outliers, corresponding to low-quality imputed variants with trait-correlated uncertainty.
The existence of software-specific differences in the consistency of imputation probabilities is an intriguing result that raises meaningful questions for the field. Though each program has its own algorithmic features and drawbacks, the practical choice between imputation software often reduces to the banal: ease of use, runtime and memory requirements. These features are important for imputation study designs that may take weeks to months to complete. Yet here we show a more substantive trend of probability bias that follows a distribution specific to particular programs. The precise cause of these distortions is not clear. Further investigation may be warranted to determine a method of adjusting the native probabilities generated from imputation using empirical distributions based on masked comparison data.

We interpret the potential performance improvement of MI over other methods as a call to reevaluate the use of thresholded genotype calls in other contexts. In particular, as the field continues transitioning to sequence-derived variant data, the impact of thresholded certainty in sequencing data analysis cannot be overlooked. We note that standard methods for rare variant analysis, including burden testing, adapt to the high uncertainty in low frequency variant calls by using strict quality control and relying on bulk information to resist the noise introduced by false positives. Yet interpreting the probabilistic output of sequencing technology as a mechanism for completing data in a NMAR setting (Methods), Multiple Imputation straightforwardly provides a consistent solution for even complex statistical analyses.

In the case of statistical tests that analyze linked variants together, per-site marginal information is no longer sufficient, as each MI draw must be taken from the joint distribution of all tested sites. Yet with sequencing read data, this kind of analysis is not impossible: joint information can be directly estimated from shared reads, creating a computationally challenging and yet feasible method of rigorously handling genetic uncertainty without even linkage assumptions. The creation of a standard module for interfacing with sequencing read data and dynamically generating such joint probability information would be a significant contribution to the field.

Similarly, the draws conducted in multiple imputation implicitly assume independence of samples. In the case of pedigrees or cryptic relatedness of samples, the sampling problem becomes much more challenging. The imputation software analyzed in this study (IMPUTE2, minimac3) is designed to handle unrelated population-based samples. The marginal probabilities generated by the algorithms are not reflective of explicit joint distribution between related samples and thus are not directly compatible with correlated genotype draws without additional modeling. One could synthesize
probabilities that scale with both the imputed probabilities and the Mendelian transmission rules, for example

\[
P(\text{child|data, parents}) = \frac{P(\text{child|data})P(\text{child|parents})}{\sum_{i=0}^{2} P(\text{child} = i|\text{data})P(\text{child} = i|\text{parents})}
\]

Unfortunately, such an approximation of the joint distribution would not be equivalent to actually modeling the combined structure of the data in the imputation software itself: for example, when the parents and children are modeled separately, certain genotypes will be given nonzero probabilities that would be rendered impossible by Mendelian transmission. In the case of BEAGLE [62], duos and trios are specially handled according to externally specified pedigrees. The resulting genotype probabilities then are reasonably considered to be conditional probabilities: for example \( P(\text{child|data, parents}) \). In this case, the appropriate probabilistic relationships should hold: for example,

\[
P(\text{child} = aa) = \left[ P(\text{mother} = aa) + \frac{P(\text{mother} = Aa)}{2} \right]
\cdot \left[ P(\text{father} = aa) + \frac{P(\text{father} = Aa)}{2} \right]
\]

and thus consistent drawing could be conducted in a trio by first drawing the child, then drawing the parent probabilities conditional on the result of the child’s draw. Ultimately, the suitability of MI to a given application is restricted to situations where the probability-generating method is itself appropriately suited.

Recent work ([74, 75, 76]) has produced algorithms (MIX, DISTMIX, ImpG-Summary) capable of imputing association statistics from summary data, without the need for individual-level genotype information. These methods offer substantial time savings relative to genotype imputation, at the cost of reduced overall quality of estimates relative to existing HMM methods. The imputed test statistic at a particular site is the mean of a multidimensional Gaussian distribution based on neighboring test statistics and linkage disequilibrium data: this intuitively corresponds to an MLE dosage estimate from genotype imputation probabilities. It is likely the method thus suffers from analogous disadvantages to those of incomplete uncertainty handling in genotype imputation. One
could imagine replacing a single mean imputed test statistic with instead a set of random variates drawn from the underlying Gaussian; these draws would act as drawn genotypes in the analysis in this study. Although multiple test statistics per variant would prove less useful for analysts of individual studies, they would be more completely reflective of the uncertain nature of these estimates. Furthermore, using sample size weighting, one could generate an MI regime in multi-cohort meta-analysis in, for example, METAL [47], in which each set of drawn test statistics is used to generate a separate meta-analysis dataset, and the resulting sets are combined using MI. Such a regime would require further investigation, and most likely would require a substantial number of iterations if many contributing cohorts used uncertain input data.

3.6 Conclusion

We characterize the behavior of modern genotype imputation tools in the context of GWAS, with a particular emphasis on the interaction of allele frequency and genotype class. We find systematic biases in imputation results, both in terms of concordance between estimated genotypes and masked data, and between genotype probabilities and their corresponding rates. There remains substantial room for improvement in genotype imputation tools, as these biases may manifest as unpredictable association errors in downstream analysis.

The use of MI is not completely foreign to the field of statistical genetics. We note, however, that its use is very limited, and has not been extensively compared to other, prevalent methods of handling genotype probabilities. We evaluate MI in comparison to existing methods and show MI performance is typically comparable to existing methods, and in certain contexts significantly outperforms other existing algorithms. Furthermore, we emphasize the ease with which MI is extended to, conceptually, all existing and future genotype call analysis methods with little additional effort. We foresee MI as a simple and effective component of all probabilistic genotype analysis.

3.7 Methods

3.7.1 Study Datasets

For this project, we applied for access to and downloaded two SNP array and phenotype datasets from the Database of Genotypes and Phenotypes (dbGaP) [77]. The Whole Genome Association
Study of Visceral Adiposity in the Health Aging and Body Composition (Health ABC) Study, dbGaP
accession phs000169.v1.p1, contains 2802 individuals genotyped on the Illumina Human1M-Duo SNP
array. STAMPEED: Northern Finland Birth Cohort 1966 (NFBC1966), phs000276.v2.p1, contains
5415 individuals assayed on the Illumina HumanCNV370 array. Both datasets were requested and
approved under project 7955; they are available from the dedicated General Research Use collection
and do not require IRB approval.

The phenotype data released under this collection are quite limited. For this study we limited
analysis to simple anthropometric traits. For Health ABC, we conducted a BMI association study
with BMI ($\mu = 27.4, \sigma = 4.77$) determined by age ($\mu = 73.6, \sigma = 2.87$) and sex (51.2% female).
For NFBC66, we conducted a height association study with height in centimeters stratified by sex
($\mu_W = 153.2, \sigma_W = 68.9; \mu_M = 286.1, \sigma_M = 66.4; 52\%$ female); age was not included for this
single-year birth cohort.

3.7.2 Sample Quality Control

Datasets on dbGaP have already been subjected to a round of cleaning by their depositors. Never-
theless, for thoroughness we cleaned the SNP array data using a standard QC protocol. Briefly,
variants with minor allele frequency less than 1%, Hardy-Weinberg Equilibrium p-value and either
per-individual or per-SNP missingness greater than 5% were removed. Cryptic relatedness was esti-
mated using genome-wide IBS estimation in PLINK [49]. A large cluster of approximate first-cousins
was detected (IBS $\hat{\pi} \approx 0.125$). For the purposes of this analysis, whether this is indicative of un-
reported pedigree structure or technical artifacts in genotype collection is irrelevant, as we are not
undertaking novel variant discovery, rather conducting comparisons relative to control data.

The remaining SNPs were pruned to an independent subset of SNPs using PLINK --indep with
default parameters, and these variants along with the maximal independent subset of individuals
were used for unrooted principal component analysis in EIGENSOFT [58]. Standard population
stratification along geographical axes is observed, confounding novel variant discovery but not ef-
fecting within-sample comparisons.

To prepare for genotype imputation, SNPs with complementary variant alleles were removed from
the dataset, and positions and SNP rsIDs were updated to those of the 1000 Genomes reference panel
we used (see below). Complementary allele variants are challenging to reconcile with datasets of
potentially different strand alignments. This is true when handling external datasets subject to
unknown prior manipulation, but in particular in the case of modern Illumina arrays, which are only annotated with “TOP/BOT” annotations instead of reference strand calls.

3.7.3 Prephasing and Genotype Imputation

Following modern genotype imputation guidelines [63], we first prephased our data using SHAPEITv2 [78] with the recommended parameters. Phased haplotype data were then probabilistically imputed to the Version 3 1000 Genomes Phase 1 Integrated global reference haplotypes [79] using IMPUTE2 [70]. For comparison purposes, to establish whether effects observed were specific to the software in use, in parallel we phased the genotype data using MACH [69] and imputed the resulting phased haplotypes to the same reference panel using minimac3 [71]. Using the global reference data, a large proportion of the approximately 40 million variants in the reference dataset are expected to not segregate in the study samples and impute very poorly; thus, before downstream analysis, variants with program-specific quality metric less than 0.4 were removed entirely. No per-genotype filtering was conducted, to avoid reintroducing NMAR bias (see below).

3.7.4 Standard Analysis Methods

For standard association models under generalized linear models (in this case linear regression), various methods currently exist for analyzing probabilistic genotypes. We selected two widely-used methods of analysis for the purposes of this study. The genetics software PLINK, as of version 1.07, has an “allelic dosage” method of genotype imputation, in which the additive predictor

\[ \text{dosage} = 2P(aa) + P(Aa) \]

is included in a generalized linear model (note that this value is not equivalent to a perfect confidence genotype, as it is permitted to be a decimal number between 0 and 2). This method projects the two parameter posterior probability into a single dimension, thus losing information in many cases: for example, this would consider the posterior probabilities \( \{0, 1, 0\} \) and \( \{ \frac{1}{3}, \frac{1}{3}, \frac{1}{3} \} \) to be equivalent.

The IMPUTE2 software used for imputation in this study has an accompanying analysis software package, called SNPTEST [22]. This software comes with a custom score test for explicitly handing genotype probabilities from imputation. The software has changed substantially since initial release; the best documentation available for the probability-handling methods is at https://mathgen.
For this investigation, we solely use SNPTEST’s frequentist methods, which can explicitly handle uncertainty.

### 3.7.5 Statistical Missingness Theory

Genotype imputation is a discipline-specific solution to a general statistical problem called “informative missingness” [23]. Consider a conceptual data matrix $Y$ containing all phenotype and covariate data for a study. Classical statistical analysis assumes $Y$ is complete, containing no null entries, and that all entries are known with perfect confidence. Study designs with null entries can be compelled into this format by removing all null datapoints before statistics are performed, resulting in a so-called “complete-case analysis.”

The effect of deviations from these assumptions vary depending on the characteristics of the missingness itself. Now assume that $Y$ is the conceptual matrix containing all true datapoints with perfect confidence. Missingness observed in realistic studies is encapsulated by a second matrix, $M$, where each entry $M_{ij}$ is 1 if $Y_{ij}$ is missing in the true study, and 0 otherwise.

Using this framework, missingness can be partitioned into three general classes. If the distribution of $M$ is independent of $Y$, the missingness is called “missing completely at random” (MCAR). In this situation, corresponding to the classical model, all missingness can be safely ignored in downstream analysis, with a potential loss in statistical power but no introduction of bias.

If instead the distribution of $M$ is dependent on the data matrix $Y$, missingness can be classified in two separate cases. If the dependency can be reduced to simply the observed subset of $Y_{obs}$, where $Y = Y_{obs} \cup Y_{mis}$, then the missingness is, somewhat misleadingly, termed “missing at random” (MAR). Data that are MAR cannot be ignored while safely avoiding the introduction of bias. However, due to the exclusive dependency on $Y_{obs}$, the missing values can be probabilistically imputed from $Y_{obs}$ and added into downstream statistical analysis.

In the worst case, the distribution of $M$ is irreducibly dependent on $Y$; such missingness is called “not missing at random” (NMAR). NMAR data cannot be removed without potentially introducing bias, and furthermore cannot be predicted solely from the observed data $Y_{obs}$. The case of genetic data collection is invariably NMAR, as missingness created by collection technologies such as SNP arrays and sequencers exhibit different performance at different underlying genotypes, and the selection of variants for SNP arrays is itself biased by numerous factors including but not limited to predominant ancestry in early variation projects such as HapMap, variant location in the genome.
and neighboring sequence content, and allele class at the site of interest.

Genotype imputation is an attempt to project the NMAR missingness of genetic data collection into an MAR condition. Due to linkage disequilibrium (LD), the nonrandom segregation of neighboring variants over a limited number of generations, one can add externally collected, ancestrally related haplotypes to the data (the \( Y \) matrix). The resulting partition of this matrix \( Y_{\text{obs}} \) now ideally contains sufficient information to probabilistically estimate missing genotype data in the original dataset at both typed and untyped variants.

### 3.7.6 Multiple Imputation

The framework for Multiple Imputation is established in [23]. Briefly, the method assumes that missing datapoints have been probabilistically estimated using some external method. From these probabilities, an arbitrary \( d \) complete datasets are drawn from the specified probability distributions. In the case of standard single-variant analysis, conducting these draws is straightforward as linkage disequilibrium can be ignored: thus for each sample at each SNP, the genotype is drawn independently of all others.

Each draw is independently subjected to the desired statistical test. This results in \( d \) sets of \( \{ \hat{\beta}_i, \hat{s}_i^2 \} \) effect and standard error at each variant. The Multiple Imputation consensus test statistic is computed from the following values:

\[
\hat{\beta}_{MI} = \frac{1}{d} \sum_{i=1}^{d} \hat{\beta}_i \\
\hat{s}_W^2 = \frac{1}{d} \sum_{i=1}^{d} \hat{s}_i \\
\hat{s}_B^2 = \frac{1}{d-1} \sum_{i=1}^{d} \left( \hat{\beta}_i - \hat{\beta}_{MI} \right)^2 \\
\hat{s}_{MI}^2 = \hat{s}_W^2 + \left( 1 + \frac{1}{d} \right) \hat{s}_B^2
\]

Here, \( \hat{\beta}_{MI} \) is the consensus effect estimate; \( \hat{s}_W^2 \) is the within-draw sample variance; \( \hat{s}_B^2 \) is the between-draw sample variance; and \( \hat{s}_{MI}^2 \) is the total sample variance. The test statistic is the ratio of the test statistic and sample standard error, \( \frac{\hat{\beta}_{MI}}{\sqrt{\hat{s}_{MI}^2}} \), which is distributed as \( T \) with \( (d - 1) \left( 1 + \frac{d \hat{s}_W^2}{(d+1)\hat{s}_B^2} \right)^2 \) degrees of freedom. The resulting probability may be interpreted as a posterior
probability incorporating both the evidence for association in the study and the actual reliability of
the genetic data.

Software implementing this Multiple Imputation regime (in beta) may be found at https://github.com/cpalmer718/statgen-mi. This package features modularized, extensible interfacing with existing analysis software, and bsub/qsub integration. In total, this implementation of MI requires $d$ times as long to run, and $d$ times as much disk space, as a single analysis run, though running in tranches per chromosome on a cluster can reduce the maximum memory and effective time use by removal of intermediate files and quasiparallelization.

### 3.7.7 Meta-Analysis of Multiple Imputation

With Multiple Imputation, a simple regime for seamlessly correcting for different proportions of uncertainty in the contributing analyses is available. Extending the logic used when combining $d$ individual MI draws, the following meta-analysis regime applies [72], for $M$ contributing cohorts to a multisite meta-analysis:

$$
\hat{\beta} = \frac{1}{M} \sum_{i=1}^{M} \hat{\beta}_{MI,i}
$$

$$
\hat{s}_W^2 = \frac{1}{M} \sum_{i=1}^{M} \hat{s}_{MI,W}^2
$$

$$
\hat{s}_B^2 = \frac{1}{M} \sum_{i=1}^{M} \hat{s}_{MI,B}^2
$$

$$
\hat{s}_{meta}^2 = \frac{1}{M-1} \sum_{i=1}^{M} \left( \hat{\beta}_{MI,i} - \hat{\beta} \right)^2
$$

The new variance component $\hat{s}_{meta}^2$ is the between-site variance of estimated test statistics. Assuming a balanced study design in which each site runs the same number of MI rounds, the total variance of this nested model multiple imputation is $\hat{s}^2 = \hat{s}_W^2 + (1 + \frac{1}{d}) \hat{s}_B^2 + (1 - \frac{1}{M}) \hat{s}_{meta}^2$. The ratio of $\hat{\beta}$ to $\sqrt{\hat{s}^2}$ is distributed approximately $T$ with degrees of freedom $\frac{1}{M(d-1)} \left( \frac{(1+\frac{1}{d})\hat{s}_W^2}{\hat{s}^2} \right)^2 + \frac{1}{M-1} \left( \frac{(1+\frac{1}{M})\hat{s}_{meta}^2}{\hat{s}^2} \right)^2$. In the case of imbalances in the number of draws conducted in each contributing cohort, more complex expressions might be derived, or alternatively a conservative estimate of $\min(d_m)$ may be used for the weighting factor in the total variance and degrees of freedom.
Chapter 4

Detecting Novel Variation in an Ashkenazi Jewish Schizophrenia Cohort

4.1 Abstract

Missing genetic heritability in complex human traits may be explained in part by rare variation [5]. Genome-wide association studies are not suited for analysis of such variants, as their primary benefit is cost savings through the assay of a predetermined set of common SNPs. Rare variants undetectable in large outbred populations may increase in frequency when captured in bottleneck populations [80], potentially reducing the number of samples required to detect a signal. We thus consider novel variation underlying risk of schizophrenia within a retrospectively ascertained sample of Ashkenazi Jews. Previous work has shown a significant enrichment of novel putatively-functional variants within Alzheimer’s patients in a study population of similar ancestry [31], as well as within a large population-based schizophrenia cohort of different ancestry [10]. We find no such enrichment in our study. There is strong evidence that the inclusion of relateds and duplicates leads to an inflated novel variant call rate when sites that would otherwise be singletons are presented as doubletons in the study. We further note nonrandom patterns of call rate and variant quality within the samples of this study, primarily affecting the unmatched controls. There is evidence suggesting
compromised selection and randomization of cases and controls during study design, leading to systematic differences between case and control sequencing due in large part to stratified identity by descent (IBD) between groups. These observations provide substantial support for stringent study design and randomization when evaluating novel variant burden.

4.2 Summary

The sequencing era has substantially shifted the focus of human genomics from common variation to rare and novel variation as possible contributors to trait variation. The detection of rare variation is complicated by high false positive rates and low counts of rare alleles, making standard association paradigms insufficient. To enrich a sample for interesting rare variation, we conduct a novel variant burden analysis in schizophrenia cases of Ashkenazi Jewish descent, whose bottlenecked ancestry is likely to increase the prevalence of high-impact variation. We find, in contrast to other studies of comparable size, substantial sequencing artifacts that preferentially affect novel variation, rendering standard “burden of rare allele” analysis unusable. Most notably, we have characterized the impact of including distantly-related samples in a sequencing study, which if not properly randomized at sample collection will create artifactual patterns of variant discovery that are correlated with phenotype. In sum, though we find little of biological interest, this study provides insight into the technical and study design concerns of novel variant testing that may be informative for future studies.

4.3 Introduction

Modern statistical genetics considers the genetic underpinnings of complex heritable traits, including both quantitative traits with continuous measurement (such as height \[8\], body mass index \[66\], blood marker levels \[81\], etc.), as well as binary traits with no apparent Mendelian basis (such as type II diabetes \[82\], Alzheimer’s disease \[83\], asthma \[84\], and many others). The traditional model for complex trait genetics involves the Common Disease/Common Variant (CD/CV) hypothesis \[1\], in which variants of high population frequency (typically greater than 1% or 5% minor allele frequency) contribute the majority of genetic heritability in traits exhibiting non-Mendelian behaviors. Borne out of both biological and practical justifications, CD/CV has found significant empirical support through hundreds of association studies (\[85\]; also Chapter 2 of this document).
Yet to many, the primary conclusion of the GWAS era is the incomplete picture of heritability presented by CD/CV: for most traits, detectable common variants do not explain the entirety of expected genetic effect, leading to the so-called “Missing Heritability” problem [12]. The selective restriction of common variation to small contributions to trait variance is coupled with the possibility of much larger magnitudes of effect in rare variation. In the most extreme case, individual susceptibility to a trait or disease may be concentrated in personal novel mutations of large effect, substantially complicating the process of describing the genetic underpinnings of heritable traits across an entire population.

Schizophrenia is a serious psychiatric disorder with a putative genetic basis. Likely causal variants have been detected in both the common [86] and rare [10] frequency spectrum. The significantly reduced fertility of schizophrenia patients [87] indicates the presence of strong negative selection on the trait, suggesting the disorder is a reasonable candidate for whole genome sequencing to identify additional rare causal variants of large effect.

In light of limited budgets and finite sample sizes, the question of optimal study design is particularly pressing for rare variant analysis. Mere detection of variants with minor allele frequency below 1%, or even 0.1%, requires genome sequences from many thousands of samples, which may be impractical in many cases. Statistical power to detect significant association with an outcome requires even more onerous sample sizes.

Most genetic analyses are formulated based on outbred populations, in which genetic diversity is subject to certain steady-state characteristics: Hardy-Weinberg equilibrium, approximately constant effective population size, fixation, etc. In the context of rare variants, recent bottleneck populations offer several possible advantages over outbred populations. Rare haplotypes from the source population that happen to be sampled in the founders of the new branch may be disproportionately overrepresented for many generations. In that time, over a limited number of generations and with restricted mating, otherwise deleterious variation may be sustained at unusually high frequencies.

A multistage sequencing study was conducted in 349 schizophrenia cases and 225 controls, matched by Ashkenazi Jewish ancestry (though without further matching between cases and controls). Limited sample size led to the application of rare variant burden testing, in the first attempt analogous to published work in a similarly restricted context ([31]; Methods). While there is little evidence for enrichment of functionally-annotated novel variation in schizophrenia cases, we detect substantial evidence for stratification effects and cryptic relatedness correlated with disease status,
<table>
<thead>
<tr>
<th>Functional Category</th>
<th>Sites</th>
<th>Novel alleles in cases</th>
<th>Novel alleles in controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initiation codon</td>
<td>37</td>
<td>26</td>
<td>16</td>
</tr>
<tr>
<td>Loss of function</td>
<td>803</td>
<td>854</td>
<td>413</td>
</tr>
<tr>
<td>Protein-altering</td>
<td>7457</td>
<td>6625</td>
<td>3616</td>
</tr>
<tr>
<td>Synonymous</td>
<td>2981</td>
<td>2906</td>
<td>1571</td>
</tr>
<tr>
<td>Splicing</td>
<td>478</td>
<td>576</td>
<td>291</td>
</tr>
<tr>
<td>Stopgain</td>
<td>289</td>
<td>253</td>
<td>106</td>
</tr>
</tbody>
</table>

Table 4.1: Distribution of novel variants, partitioned by functional class from Variant Effect Predictor (VEP) [90], out of $1.75 \cdot 10^7$ variants included in sequencing call dataset after full QC. Functional category definitions are taken from [31]; note that VEP was used for their confirmation analysis, which gave them results indistinguishable from their main results. Due to several indications of data quality failures, indels are excluded from this analysis.

which is a function of both study design and burden test model. Taken as a whole, we provide meaningful observations about the particular challenges of novel variant burden testing that are informative to future sequencing study designs.

### 4.4 Results

#### 4.4.1 Sample Characteristics

577 samples of Ashkenazi ancestry were sequenced using the Illumina X10 platform. This sample includes one case and 27 controls that were either (1) IBS $\hat{\pi} > 0.05$ with another sample in the study, or (2) included as duplicates of other sequencing samples, and were removed from downstream analysis. After joint calling (GATK [88], New York Genome Center), variant quality filtering, and additional stringency filters on segments annotated by the 1000 Genomes Project as problematic for sequencing (Methods), the study contained approximately $1.35 \cdot 10^7$ variants across all frequencies. After filtering for variants previously observed in 1000 Genomes Phase 3 [89], novel variants were partitioned into groups based on functional annotation (Table 4.1).

#### 4.4.2 Variant Quality Stratification

In the quality control process, affection status was repeatedly observed as a significant predictor of variant quality. Across all sequenced variants, without additional QC filtering beyond that performed by the sequencing center, lower variant quality was significantly associated with higher minor allele
frequency in controls (Figure 4.1; case status-permuted \( p < 0.006 \)). This additionally manifests as significant deviations in pseudo-identity by state (IBS) relatedness in cases and controls (see Figure 4.2), such that controls appear much more strongly related within-class than cases \( (p < 0.00001) \). This signal appears to be driven by the inclusion of poorly sequenced sites, creating a false IBS signal that is correlated with outcome. Note that this pool of variants is expected to contain many poorly-sequenced variants, and is not representative of the final callset; rather, we consider this callset first to observe meaningful batch effects that may be obscured by later quality control.

Filtering variants to “PASS” quality reduces but does not abrogate this association. While genome-wide variant quality is no longer associated with control frequency (Figure 4.3, \( p = 0.7 \)), variants with minor allele count less than 10 (corresponding to approximate minor allele frequency < 0.01) still demonstrate this association \( (p < 0.004) \).

Applying additional filtering in this study according to the 1000 Genomes low confidence regions,
Figure 4.2: **Without variant filtering, spurious differential relatedness in samples correlated with affectation status.** Shown are histograms of pairwise IBS distributions within-cases, within-controls, and between-status, including an LD-pruned subset of common (minor allele frequency > 0.05) variants with no additional quality filtering. X-axis: pairwise (apparent) IBS; Y-axis: number of sample pairs in IBS bin. Excluded from analysis are 27 controls and 1 case known to be related. All distributions are significantly different from each other (PLINK IBS test method, all \( p < 0.03 \), all but one less than \( 10^{-3} \)). Note that due to the inclusion of sequencing failures, these numbers do not represent true IBS signals but rather are predominantly sequencing errors nonrandomly correlated between samples.
Figure 4.3: **Removal of non-PASS variants removes some but not all of the association between VQSLOD and outcome.** As in Figure 4.1, the X-axis depicts minor allele count bins with the genome-wide signal represented in the left-most bar, while Y-axis is average VQSLOD(case-control). The genome-wide signal is removed, as are many of the frequency bins, in particular those at the highest minor allele counts that corresponded to strong Hardy-Weinberg equilibrium violations. However, the signal at the lowest minor allele count bin remains ($p < 0.004$).
which suffer from consistently poor sequencing quality, the noticeable case status/quality association is finally removed (Figure 4.4). The pseudo-IBS association detected genome-wide is apparently removed by the final tier of variant filtering (Figure 4.5); however, the between-group permutation p-value for nonrandom clustering by affectation status is $p = 0.061$ in the same direction as the original stratification effect, based as usual on independent common variants.

### 4.4.3 Novel Variant Stratification

Noting that uncommon variants selectively retained stratification signals through the QC process, we specifically examine substrata of novel variants considered within this study for stratification signals associating sequencing quality and disease status. Novel synonymous variants, used effectively as loading controls in the Fisher testing in [31], show a strong within-class association between variant quality and control frequency (Figure 4.6; $p < 0.00135$, upper 95% confidence bound on 100000 permutations). The subclass of novel variants annotated as “protein-altering,” a superset of loss-of-function variants along with missense and stop-loss variants, show a similar association ($p < 0.00045$, upper 95% confidence bound on 100000 permutations). No such association is detected in novel loss-

---

**Figure 4.4:** Additional removal of 1000 Genomes pilot region variants removes the remainder of unexpected association with schizophrenia outcome. X-axis: minor allele count bins with genome-wide signal as left-most bar; Y-axis: average VQSLOD(case - control) within corresponding bin.
Figure 4.5: With stringent variant filtering, differential relatedness in samples statistically uncorrelated with affection status. Shown are histograms of pairwise IBS distributions within-cases, within-controls, and between-status, after two stages of quality filtering. X-axis: pairwise IBS; Y-axis: number of samples in IBS bin. Excluded from analysis are 27 controls and 1 case contributing to pairwise IBS $\hat{\pi} > 0.05$. Distributions are not significantly different from each other (PLINK IBS test method, all $p > 0.061$).

We hypothesized that the stratification signal could be specific to individual contributing studies within the 14 sets of controls submitted for sequencing, as opposed to within the two sets of cases. To this end, we conducted within-outcome permutations of study labels, and tested for nonrandom distribution of novel singleton quality ranking (so a non-parametric test on VQSLOD).

For novel loss of function variants, we do not observe a significant association between variant quality rank and all group labels taken as a single categorical predictor (Kruskal-Wallis Rank Sum $p = 0.158$), nor do we see nominally significant associations between quality and individual group labels (all individual $p > 0.05$; though likely enriched for tests with $p < 0.1$, with such enrichment $p = 0.017$). For novel synonymous variants, the group label categorical predictor is significantly associated with variant quality rank (Kruskal-Wallis Rank Sum $p = 0.0045$), though once again no individual group is significant (all permutation $p > 0.05$).

As VQSLOD is not an absolute metric but rather a complicated and difficult to interpret approximation of variant quality, we also consider novel singleton call counts within contributing groups. Note that the relationship between group VQSLOD and group singleton count is not unique, and
Figure 4.6: **Within novel variant annotation categories, significant association between variant quality score and schizophrenia.** In this case as opposed to prior figures, the X-axis represents variant annotation category, while the Y-axis still represents average VQSLOD(case - control) amongst variants in each class. Significantly higher VQSLOD in cases is detected for both the protein-altering and synonymous categories (both $p < 0.002$).
may have a variety of interpretations given the context. Furthermore, group singleton count represents an abstraction that ignores missingness and uncertainty of underlying sequencing data by introducing a confidence threshold that is then hidden from downstream analysis.

Considering novel loss of function variants, sample group is significantly associated with variant count per sample (Kruskal-Wallis Rank Sum $p = 0.003$). Three individual groups are associated with aberrant mean rank amongst included individuals (Figure 4.7; from left to right, $p = 0.0194, 0.03194, 0.012$; probability of observing three such results at $\alpha < 0.05$ under null: $p = 0.043$). Considering instead novel synonymous variants (Figure 4.8), sample group is again associated with variant count per sample (Kruskal-Wallis Rank Sum $p = 0.0088$). Only one group is significantly associated in the permutation test (red bar, $p = 0.00072$). In both cases, subcohorts with significantly inflated counts correspond to groups that provided duplicates or related samples to be sequenced; although samples were pruned before analysis to contain a maximal set of samples unrelated at pairwise IBS $\hat{\pi} < 0.05$, the inclusion of these samples in sequencing has left a residual signal amongst novel variants that were presented to the sequencer and calling algorithm as doubletons.

### 4.4.4 Residual IBS Stratification at Doubletons

We observed unusual behavior amongst novel doubletons. In conducting standard GWAS cryptic relatedness sample filtering by IBS $\hat{\pi}$, we removed one case and 27 controls contributing to IBS relationships with $\hat{\pi} > 0.05$, leaving as many samples as possible while still maintaining an apparently pairwise-unrelated dataset. As expected, the 28 removed samples are substantially overrepresented in doubletons in the unfiltered dataset (expected to randomly participate in approximately 9.5% of doubleton pairs; observed 25% to 40%, depending on the functional subclass).

Amongst the 549 remaining samples, two cases and 24 controls were kept that were part of the filtered IBS relationships; this is standard protocol for GWAS to maximize remaining sample size. We observed that these individuals were overrepresented in novel singletons ascertained from the filtered dataset (all $p < 2 \cdot 10^{-16}$). This is consistent with these individuals enjoying an aberrantly high discovery rate when they are sequenced in duplicate but analyzed alone.

However, amongst doubletons discovered in the filtered dataset, we see large discrepancies in the samples contributing to doubleton pairs. While the overall counts of doubletons amongst functional classes are reasonably consistent with expected counts of case/case, case/control, and control/
Figure 4.7: **Novel loss of function singletons show association between count and cohort.** X-axis: contributing cohort for sequencing; Y-axis: number of novel loss of function singletons. Bars correspond to group means, while points correspond to individuals contributing to each cohort. Intervals represent approximate 95% confidence around expected group mean. Red bars are groups with non-parametric permutation test $p < 0.05$; due to rounding and handling of ties, these occasionally differ from the confidence intervals. Note that coordinates of points are jittered for visibility, but in fact represent integral Y-axis counts. Overall Kruskal-Wallis Rank Sum test for association between study subgroup and singleton count: $p = 0.003$; permutation test for association in individual groups (for red bars only, from left to right: $p = 0.0194, 0.03194, 0.012$).
Figure 4.8: **Novel synonymous singletons show association between count and cohort.**

X-axis: contributing cohort for sequencing; Y-axis: number of novel synonymous singletons. Bars correspond to group means, while points correspond to individuals contributing to each cohort. Confidence intervals are parametric 95% intervals around expected group mean. The red bar is the group with non-parametric permutation test $p < 0.05$; while similar to the parametric test denoted by the confidence intervals, they differ slightly and should not be confused. Note that the coordinates of points are jittered for visibility, but in fact represent integral Y-axis counts. Overall Kruskal-Wallis Rank Sum test for association between study subgroup and singleton count: $p = 0.0088$; permutation test for association for the significant control subgroup $p = 0.00072$. 
Table 4.2: Distribution of novel doubleton variants, partitioned by functional class from Variant Effect Predictor (VEP) [90]. Data are partitioned between doubletons involving two case samples, two control samples, or one of each status. Percentages represent percent of individuals of a particular disease status involved in at least one doubleton pair. All percentages are significantly lower than expected assuming independent observations of doubletons (all $p < 10^{-22}$).

<table>
<thead>
<tr>
<th></th>
<th>Case/Case Variants</th>
<th>% Samples</th>
<th>Case/Control Variants</th>
<th>% Samples</th>
<th>Control/Control Variants</th>
<th>% Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss of function</td>
<td>26</td>
<td>13%</td>
<td>37</td>
<td>10%</td>
<td>/ 18%</td>
<td></td>
</tr>
<tr>
<td>Protein-altering</td>
<td>192</td>
<td>66%</td>
<td>182</td>
<td>39%</td>
<td>/ 54%</td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td>72</td>
<td>32%</td>
<td>73</td>
<td>19%</td>
<td>/ 30%</td>
<td>25</td>
</tr>
</tbody>
</table>

control pairs based on the total number of called sites, each substratum excludes a large proportion of the total set of possible interacting samples (Table 4.2). In all cases, the number of samples entirely excluded from novel discovered doubletons is inconsistent with a discovery process randomly sampling available individuals (all $p < 10^{-22}$). This calculation assumes doubletons are discovered independently of one another, which is invalid if this process is affected by long shared IBD segments between individuals. In the presence of IBD, as expected in Ashkenazim, the effective number of independent draws (discovered doubletons) will be lower. In order to make the observed percentage coverage of individuals be consistent with random sampling, the sampling process must in fact be independently discovering less than 10% of the total number of doubletons in order to generate good fit of this model.

We considered whether individuals are equally likely to contribute to case/case or control/control doubletons, respectively. We downsampled individuals from the case/case pairs to match the number of unique individuals in the control/control subsets, and then observed the resulting number of doubletons that would be discovered from such a set of cases. For all three functional classes in Table 4.2, cases contribute significantly fewer doubleton variants per sample than controls (no downsampled case sets discovered as many doubletons as the true control set, in 100000 attempts for each functional class).

We finally investigated whether the observed pairs of individuals contributing to novel doubletons exhibit nonrandom IBS relationships. IBS is computed from independent common variant distributions; yet we hypothesized that it might be a reasonable proxy for the IBD relationships considered here. We thus took the observed distribution of IBS $\hat{\pi}$ amongst novel doubletons, and compared it to the distribution of IBS amongst the same doubleton set with permuted individual labels, stratified
<table>
<thead>
<tr>
<th>MAC</th>
<th>Loss of Function</th>
<th>Synonymous</th>
<th>Normalized</th>
<th>Fisher p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.18</td>
<td>1.88</td>
<td>1.16</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>1.36</td>
<td>1.99</td>
<td>0.69</td>
<td>0.08</td>
</tr>
<tr>
<td>3</td>
<td>1.52</td>
<td>1.87</td>
<td>0.81</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>3.40</td>
<td>1.62</td>
<td>2.09</td>
<td>0.08</td>
</tr>
<tr>
<td>5</td>
<td>1.86</td>
<td>1.68</td>
<td>1.10</td>
<td>0.87</td>
</tr>
<tr>
<td>6</td>
<td>2.53</td>
<td>1.05</td>
<td>2.42</td>
<td>0.02</td>
</tr>
<tr>
<td>7</td>
<td>3.38</td>
<td>1.80</td>
<td>1.88</td>
<td>0.20</td>
</tr>
<tr>
<td>8</td>
<td>2.00</td>
<td>2.39</td>
<td>0.84</td>
<td>0.71</td>
</tr>
<tr>
<td>9</td>
<td>2.86</td>
<td>1.42</td>
<td>2.01</td>
<td>0.23</td>
</tr>
<tr>
<td>All</td>
<td>2.08</td>
<td>1.84</td>
<td>1.13</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Table 4.3: Fisher and Cochran-Mantel-Haenzel summary data for loss of function versus synonymous novel variant discovery. Data represent relative enrichment in cases versus controls. The “normalized” column refers to the ratio of the preceding two columns; values greater than 1 indicate relatively more loss of function novel variants detected in cases than expected based on synonymous baseline. MAC: minor allele count strata for CMH test.

by case status. Due to the heavily skewed distribution of IBS in the final QC set, we compare these distributions using the non-parametric Wilcoxon rank sum test. Within the novel protein-altering and synonymous variant subsets of doubletons, the distribution of IBS amongst paired individuals is significantly higher than expected by chance \((p = 0.0152\) and \(p = 0.0449\), respectively; 100000 permutations of status-stratified sample ID). No significant deviation is detected for the loss of function doubletons \((p = 0.492)\).

4.4.5 Novel Variant Burden Testing: Fisher/CMH

The preceding results strongly suggest technical artifacts are interfering with novel variant discovery in this sample. With this in mind, we consider skeptically a novel variant burden test in this study. In addition to the Fisher test conducted in [31], we conduct a conceptually similar Cochran-Mantel-Haenzel test (Methods) for contingency data, with case status and functional versus synonymous alleles as the meaningful margins and minor allele count as confounding stratum. For novel loss-of-function variants, there is no significant association between affection status and allele burden relative to synonymous control \((p = 0.077;\) corresponding Fisher test ignoring allele count \(p = 0.103)\). Loading of functional alleles between cases and controls is shown in Table 4.3.
4.4.6 Novel Variant Burden Testing: GMMAT

In light of possible residual cryptic relatedness in this study, along with concerns about unusually large segments of IBD in samples from bottleneck populations, we have investigated other burden tests that specifically model intrasample correlations. A more extensive discussion of available methods and limitations may be found in the Discussion; we show here the results of GMMAT [91], which effectively fits a generalized linear mixed model (GLMM) testing per-individual case status as a function of novel functional allele count. Note that under this model, novel synonymous allele burden is not a significant predictor of case status ($p = 0.63$).

Considering loss-of-function variants, we find significantly increased burden of novel alleles in schizophrenia cases (affectation-permuted GMMAT p-values, 10000 simulations (see Methods): $p = 0.0059$. This result considers the sample subset consisting of 549 samples with maximum pairwise IBS of 0.05, and furthermore explicitly controls for the apparent genetic relatedness matrix amongst reported-unrelated individuals as computed from pairwise IBS among common variants; however, in practice, the results computed with GMMAT are not meaningfully different from those obtained from the corresponding generalized linear model.

The functional category “loss of function” refers to the combined set of initiation codon, splicing, and stopgain variants as annotated by VEP [90]. Among these subcategories, novel stopgain variants are significantly enriched in cases (GMMAT permuted $p = 0.0035$); splice variants are not significant but trending consistently ($p = 0.06625$); and there are insufficient novel initiation codon variants in the sample, considering their distribution across 549 individuals, to draw meaningful within-class conclusions (see Table 4.1).

4.4.7 Uncalibrated Type I Error Rates

Due to the limited publication history of the Fisher model for directly testing rare variant burden, we investigated the behavior of various implementations of this model (Fisher/CMH, GLM, GMMAT [91], EMMAX-CMC [92]) when testing the null burden of random rare variants. In particular, we matched sets of exonic variants to the novel “loss of function” and novel “synonymous” categories of rare variants by minor allele count, and used these subsets of variants to determine the empirical null distribution of these tests. Note that the majority of novel variants in this study are singletons (see Figure 4.9), and so to avoid complications involving mixtures of allele counts, we present results
Figure 4.9: Distribution of minor allele counts across classes of variants. X-axis: minor allele count bin; Y-axis: proportion of sample. Novel variants, in this case those specifically annotated as “loss of function,” demonstrate a significantly higher proportion of singleton and generally low-count variants relative to the genome-wide distribution. This is expected, as novel variants are a priori expected to have lower frequency, in particular if subject to possible selective pressure.

Exclusively for the singletons that constitute more than 80% of the matched variants.

Empirical null distributions are shown in Figure 4.10 for the Fisher test, as well as for the generalized linear model equivalent used in this study. Testing random sets of frequency-matched exonic variants, which should not be associated as a whole with schizophrenia, versus the true set of novel synonymous variants generates a strongly inflated empirical null distribution. This inflation is not specific to the novel synonymous control variants per se: replacing them with a fixed set of frequency-matched exonic variants generates, depending on the particular draw, distributions with a spectrum of inflation or deflation. Shown in Figure 4.10 is one particular distribution with an inflation factor closely matching that observed with the true novel synonymous set. The magnitude of inflation is specific to the Fisher model, but null inflation of a lesser magnitude is observed with the GLM model as well (genomic control $\lambda = 1.23$, $p_{null} = 0.0017$).

Correctly calibrated null behavior is observed when the synonymous control set is resampled with every simulation (Fisher: $p_{null} = 0.7243$; GLM: $p_{null} = 0.2204$). The precise reasoning behind this behavior will be covered in the Discussion; for the moment, we simply report that some sort
Figure 4.10: The null distribution of the rare variant testing model is inflated under partial resampling. Shown are quantile-quantile plots (negative log$_{10}$ scale) for three null models for each of the Fisher test and GLM. Under the “syn control” model, a set of random exonic singletons, matched in count to observed novel loss of function singletons, is tested for differential burden versus the count of novel synonymous alleles in the sample. For the “fixed control” model, both the loss of function and synonymous singleton sets are randomly matched in the exome, but a single random synonymous-matched set is used as null control for each of the random loss of function-matched sets. Finally, for “resampled control,” each functional-matched set is paired with a different synonymous-matched set. The “fixed control” set shown here is selected from 1000 simulated sets to present the inflation factor closest to the one observed in the “syn control” model; more extreme inflation factors were observed and are not shown. The “resampled control” tests for both Fisher and GLM are indistinguishable from random expectation under a true null model by inflation factor (Fisher: $p = 0.7243$; GLM: $p = 0.2204$).
of null resampling is required for desirable type I error behavior, and permutation of case/control status will not correct this problem.

4.5 Discussion

In this study, we have undertaken a simple version of rare variant burden testing to find differences in novel variant accumulation between schizophrenia cases and controls. The subset of novel variants with “loss of function” annotations does not show significant differential loading between cases and controls using a previously published method. We have detected and attempted to diagnose or address as best as possible sample stratification, particularly in control samples, but it is open to debate whether we can possibly control for such trait-associated technical effects within novel variants, which are highly susceptible to sequencing errors.

The results of this study emphasize the importance of rigorous and consistent study design. At its core and beyond any particular technical issues, the study intrinsically suffers from retrospective trait collection. In the presence of such a limitation, it is of critical importance to enforce a strict collection regime. Cases and controls are typically collected using some sort of matching regime, often based on PCA (see, for example, [31, 93]). No such efforts were undertaken for this study, and retrospective matching from the available samples will inevitably result in a serious loss of power. One may suppose that this study design is the result of a desire to maximize sample count in the context of realistic collection constraints. Yet in this situation, the lack of formal sample restrictions has led to an outcome confounded by challenging stratification effects.

As the fixed synonymous control in the Fisher model appears to subject the test to an undesirable Type I error rate under the null, we investigated other methods of representing the baseline differential loading of alleles between cases and controls in the absence of disease effect. As shown in Figure 4.11, the distribution of random exonic singleton alleles recapitulates the approximate imbalance of cases to controls (actual ratio 1.74:1 in 549 samples). However, the loading of loss of function alleles (shown as red line in Figure 4.11) is significantly higher than expected under this null distribution ($p < 0.01$; similar imbalance observed in synonymous alleles). In effect, this is a restatement of the result from the null Fisher test on this sample. However, with the knowledge that the Fisher test has invalid asymptotic behavior within a single study, the offset of allele loading from random exonic variants complicates the choice of a set of variants from which to resample a null set,
as one must find sets of variants that have equivalent error distributions in cases and controls.

We find evidence in Figure 4.7 and Figure 4.8 that inclusion of related and duplicate samples in sequencing is more complicated than we originally considered. The initial analysis plan for this study followed the plan for an analogous common variant association study, at least as regards cryptic relatedness. As a first approximation, we pruned the sequenced dataset to contain only samples with pairwise IBS $\hat{\pi} < 0.05$, corresponding to second cousins, and considered these remaining samples as approximately unrelated. However, we removed the minimum possible number of samples to maintain the maximum downstream power while maintaining the unrelatedness constraint. We thus left in several dozen samples that contributed to the original relatedness pairs. What we observe is that these leftover samples retain a significant inflation in call rate at novel singletons, as these most challenging variants are provided with double evidence during the sequencing process. As is the case with all stratification signals we detect in this study, it is difficult to quantify the precise effect this signal will have on downstream analysis. The inclusion of relateds in this study was not random with regards to outcome: the vast majority of related samples were controls. We therefore cannot
assume that any associated effects will “average out” across the association test. An appropriately constructed burden test may be able to control for these call rate differences, but will likely have to assume that the impact of relatedness on call rate impacts all novel exonic variation to an equal degree. Any deviation from this assumption will induce an artifactual association between outcome and variant burden that cannot be controlled without additional validation sequencing, to which we do not have access.

The above observations about sample-wide relatedness and its impact on sequencing raises additional questions about the impact of nonrandom local identity-by-descent (IBD) present in either cases or controls. In individual regions with IBD segments, one would expect recent mutations to have higher sequencing quality due to the inclusion of the mutation in multiple samples. While the sequencing technology, as deployed in this study, does not explicitly accommodate IBD into its calling algorithm, additional copies of a variant will still have some impact within the IBD region. In many contexts these local improvements in call rate will either be (1) undetectable due to the high sequencing quality on common variants in the region, or (2) unconditionally advantageous as the result is simply better sequences. However, in the specific context of small-sample rare variant burden testing, one could reasonably construct a model in which nonrandom partitioning of IBD segments between cases and controls, or within control subgroups, or between classes of functional variants, could lead to artifactual trait associations. We mention this concern as a possible point of consideration for future sequencing studies: relatedness should either be formally addressed or confirmed to be uniformly distributed a priori.

A surprising observation during this analysis is the extent to which this study design, beyond technical artifacts from non-randomization, is fundamentally incompatible with existing statistical models. The possible presence of trait-associated variant quality and cryptic relatedness introduces the problem of conducting regression on a binary trait in the presence of an arbitrary and informative correlation structure.

The Fisher and Cochran-Mantel-Haenzel tests interrogated in this study assume independence of observations in all bins. This assumption is maintained in all major alternative methods for analysis of contingency data. Deviations from independence can typically only be accommodated in cases of structured dependency, such as paired sampling [94]. Furthermore, as we have seen, the CMH and Fisher tests require a resampled null distribution of variants for each independent burden test conducted, which is challenging to accomplish in the specific context of novel variation with
stratified sequencing performance between cases and controls.

The most obvious alternative to contingency testing is the use of regression. The standard model for association between predictors and outcomes in the field is obviously the generalized linear model (GLM). As this *ad hoc* burden test does not sum across individuals, the requirement for a null sampling of variants representing the imbalance between cases and controls in the study is removed. The GLM is effective in the case of independence of samples, but has no means of modeling sample clustering: specifically, the likelihood being fit in this model is only formally guaranteed to exist in the case of independence of samples [95]. Depending on the particular sample structure, standard regression may still perform well, but this is not a reliable outcome. In the case of our sample, we have a dataset with several dozen IBS-related samples primarily in controls; even with these particular samples removed, we are left with a sample from a bottleneck population that will have higher inter-sample correlation than is usual amongst population studies. Alternatives that explicitly handle inter-sample correlation are desirable.

Generalized least squares (GLS; not to be confused with GLM) offers the opportunity to extend regression to a correlated dependent variable. An implementation of generalized least squares that handles arbitrary correlation structure exists in R [96]. Unfortunately, GLS (in spite of the name) assumes a normally-distributed outcome. Alternatively, in MATLAB, an implementation of GLS can handle a binary dependent variable, if the mean and variance functions of the fit are specified to fit the logit function [97]. However, these two options are not combined in any existing general implementation, as far as we can determine. The primary concern is simply that the introduction of the nonidentity link function introduces dependency between the mean and variance estimators in the model.

Taking a different approach, we might consider traditional models for handling clustered (or equivalently longitudinal) data analysis. The two primary methods of handling clustered binary data in regression are generalized estimating equations (GEE; [98]) and generalized linear mixed models (GLMM; [99]). The choice of model primarily depends on the desired interpretation of the estimated parameters, either marginalized over random effects or conditional to them. Regardless of this particularity, the introduction of correlation structures takes place through the random effects parameters in both models. These take the form of categorical clustering variables, which partition the samples into subunits. Correlation is modeled *within* these clusters, whereas samples *between* clusters are still assumed to be independent; the algorithms then typically have asymptotic guaran-
tees dependent on the number of clusters [100]. Under structured pedigrees, these clusters could be approximated by family unit; however, in the case of unstructured cryptic relatedness, there is no easy partitioning.

Moving beyond general-purpose models, the primary in-field method for handling cryptic relatedness in association studies is the EMMAX software package [56]. This software confronts the computational complexity of deploying a full GLMM on whole-genome variant data by isolating a single variance component based on pairwise IBS estimates. EMMAX offers the notable distinction of not requiring cluster data as full GLMM does; rather, it computes the variance component corresponding to the provided covariance structure, and then conducts ordinary least squares on the result. This method has numerous similarities to GLS.

Although the initial release of EMMAX was closed source, it has since been integrated into an open-source genomics pipeline EPACTS [92]. From that source, and from the EMMAX paper [56], we find that variant burden testing on binary outcomes is apparently possible. However, inspecting the source code, we find the following:

- EMMAX does not directly support binary traits, they are treated as normal traits; and

- EMMAX does not explicitly support covariates; rather covariates are combined with the kinship matrix and are then simultaneously regressed out of the outcome

It would be exceedingly convenient to ignore normality of response variable for this analysis. The EMMAX paper loosely cites Armitage [101] to justify this behavior; among the canonical citations for the need to formally model binary outcomes with appropriate linker functions is [95]. It would seem that using an unrestricted continuous linear combination to estimate a binary response with dependent variance is a rather serious offense, as argued in [91] in particular for EMMAX.

Addressing the particular concern of EMMAX-style association analysis on retrospectively ascertained case/control traits without violating distributional assumptions is a topic of active research, falling vaguely into three categories, according to the general class of solution employed.

The software CARAT [102], as well as its predecessor from the same group, ROADTRIPS [103], and its successor CERAMIC [104], use a custom specification of estimating equations (analogous to GEE) in order to incorporate a genetic relatedness matrix as a correlation structure. The primary difference between the packages involves covariate and missing data handling. While these options may well be an adequate substitute for EMMAX for single SNP association, the underlying test
statistic is a score test derived specifically for single SNP testing. The model particularly assumes Hardy-Weinberg distribution of alleles and biallelic SNPs. While the authors cite a generalization for multiallelic SNPs [105], sums of rare events strongly violate the model’s distribution assumptions; furthermore, stratification effects, such as the differential cryptic relatedness brought about by sequencing errors observed in this dataset, are not addressed [106].

As CARAT/ROADTRIPS are genomics-specific variants of GEE, so the software GMMAT [91] extends GLMM into the genomics problem of sample-wide correlation estimated from a relatedness matrix. While the underlying model is distinct, GMMAT suffers from the same limitations as CARAT: statistical tests run under this model strongly assume common biallelic SNP behavior, and do not address nonrandom relatedness correlated with affection status. We have deployed this method in this study as it is minimally feasible to run, a trait not shared by all of its competitors. We have adopted extensive permutations to generate empirical p-values; however, this process is extremely computationally intensive, and does not necessarily address all the structural problems present in the dataset. Furthermore, upon actually deploying this model, we find that the results are effectively indistinguishable from those of GLM. At least from the perspective of common variant IBS as a proxy for sample correlation, there is no significant impact of sample correlation on this study’s burden test association. There may well be more subtle methods of specifying correlation structure based on the effects specific to rare variant sequencing quality that we observe, but in the absence of signal and preponderance of evidence of technical artifacts, we have not undertaken this effort.

Finally, in order to provide a complete picture of the state of the field, the LTSoFT package [107, 108, 109, 110] purports to address retrospectively collected binary trait association in the presence of arbitrary cryptic relatedness and biased relatedness sampling. Using epidemiological parameters specific to each individual trait analyzed, this model computes an expected (normally distributed) liability distribution underlying the trait. The resulting distribution is used as a normal outcome in standard, relatedness-aware LMM. In theory, if such a distribution could be reliably computed, it would be straightforward to run a burden test under LMM with variant burden as a standard, arbitrarily distributed quantitative covariate.

Unfortunately, several limitations prevent this model from being deployed in the context of the current study. It is difficult to judge the quality of software released so recently, yet LTSoFT (in its implementation, called LTFAM, designed to address nonrandom relatedness sampling) seems to
suffer from serious issues. Even the example data provided as a simple demonstration of the software run with errors, on multiple platforms. More seriously, both the example run and a test run on the schizophrenia dataset from this study report a grievous issue: the posterior mean liability (Fam-PML, in the software’s parlance), which is the normally distributed liability distribution used as the outcome in the final LMM, is effectively still binary. In the case of the example file, the case/control data have been transformed from placeholder numbers \{1, 2\} to rescaled numbers on (for cases) [2.32, 2.34] or (for controls) [−0.04, −0.02]; similar results are observed for the schizophrenia data.

The non-normal distribution of the expected liability values for the dataset suggest a more fundamental issue at work: the liability distribution will only be informative and smooth when adequate numbers of clinical and genetic predictors are applied to the model. In the case of the current schizophrenia dataset, no clinical predictors are available; the only non-genetic parameter with discriminatory power is apparent SNP-sex, with 1.4X disease likelihood in men. Providing a more complex model involving a large number of genetic predictors may yield a more reasonable liability distribution. Yet even under this model, issues remain. The epidemiological data used by the software is specified to be disease prevalence within bins of epidemiological predictors: in other words, prevalence of disease in individuals of age 35–45, or individuals of high cigarette consumption, etc. Unfortunately, example files with the software, when compared to the citations given for them, provide conflicting evidence of whether the software should in fact use prevalence data; or rather incidence data, as is included in several instances; or instead some other form of data, as several of the citations did not include the reported data at all. That these files were cited as specifically those used by the developers in [107] is additionally troubling.

In brief, we see no existing method for formally correctly handling a study design such as has been retrospectively determined in this study. From a certain perspective, it is inappropriate to blame the statistical models available for the failure to handle this analysis. Study design is an established field and has significant theory and tools available. Ignoring them during sample collection is standard enough in statistical genetics, but introduces predictable problems further on. We may consider this dataset to simply have a fundamental flaw that prevents it from being used for case-control analysis of novel variants.

This analysis raises serious questions about the validity of the Fisher test results from [31]. The evaluation of the null behavior of the statistical tests in Figure 4.10 reveals a flawed assumption at the core of the model deployed in the previous study. Asymptotically, the novel synonymous variant
control provides an appropriate representation of the null distribution of novel variants in a study. However, the relevant asymptote is based on repeated sampling of novel synonymous variants across multiple studies, and furthermore this asymptote is obligatory and does not extend to reasonable behavior in a single cross-section. Within an individual study, the observed ratio of case novel synonymous variants to control is a fixed sample from an error distribution around the conceptual true mean ratio. The deviation of the observed value from the unknown underlying mean value (approximately \( \frac{\text{number cases}}{\text{number controls}} \)) corresponds to a consistent directional bias in every Fisher/CMH test using this model within a single study.

In practice, this means that for a given functional variant class, we should expect the results of many different studies to behave correctly under the null; yet within a single study, multiple functional variant class tests using a single null will exhibit systematic null inflation or deflation, depending on the precise parameters (sample size, case/control imbalance, sequencing quality stratification between cases and controls, direction of observed loading of particular functional class) in a particular study. We note that the novel synonymous allele ratio observed in [31] is indeed depleted in cases relative to the actual case count, meaning all of the results from the study are likely anticonservatively biased and correlated.

Furthermore, the study in [31] provides no formal analysis of cryptic relatedness. Cryptic relatedness is, to some degree, a feature of all genetic studies. However, the nature of bottleneck populations leads to a much more substantial degree of sample relatedness between otherwise distinct families. In other contexts, this relatedness is actually beneficial [111]; however, especially in the context of small sample sizes (45 cases, 53 controls for the study in question), specific controls must be implemented to prevent meaningful stratification gradients that correlate with disease outcome. The reference Alzheimer’s study makes no mention of quality control assessment of cryptic relatedness, and cases and controls were ascertained separately with matching on the first two principal components.

While none of these results directly contradict the results of [31], they raise substantial questions about the reliability of the study. It is possibly trivial to point out that these issues and many more are, in the end, why it is traditionally considered inappropriate to conduct biological tests on tiny sample sizes using nonstandard statistical models. Nevertheless, the fact that this schizophrenia study, five times larger than [31], is subject to numerous confirmed stratification and batch effects shows once more that the sample sizes required to wash away concerns about informal study design are quite large in the context of human genomics.
4.6 Conclusion

We conduct rare variant discovery in a sample of Ashkenazi schizophrenia cases and controls. Limited by sample size, we investigate the case burden of novel loss of function variation. In lieu of a strong biological signal, we find substantial evidence for stratification effects in the set of novel variants, impacted in part by both sequencing quality and identity by descent. As relatedness was not a priori accounted for during the design of this study, any novel variant discovery test is likely to be fundamentally confounded by differential variant discovery rates in regions of IBD, so long as any heterogeneity exists between cases and controls.

The impact of IBD on variant burden and discovery has certainly been considered previously, even particularly with Ashkenazim [112]. However, in prior contexts, this sharing has been considered a benefit, as it could potentially enable long-range imputation or inference. Here, we see the potential for the introduction of confounding effects, in particular in small samples, when insufficient matching is performed between cases and controls. Of particular note is the signature left on samples even after GWAS-style post hoc IBS filtering is performed. These results, beyond considerations for the present study, may inform proper sequencing study design moving forward.

4.7 Methods

4.7.1 Sample Characteristics and Sequencing

574 individuals of Ashkenazi Jewish ancestry were sequenced on the Illumina X10 platform. This corresponds to 349 cases and 225 unmatched controls. The samples were jointly called by the New York Genome Center using GATK [88]. Standard PASS quality control calls were generated for variants with variant quality score log-odds (VQSLOD) less than $-1.9403$ (for single nucleotide polymorphisms) or $-0.7365$ (for insertions and deletions). Additional quality control was applied by collaborators by removing variants in genomic regions with the following filters:

1. remove SNPs in masked and low complexity regions
2. remove SNPs with VQSLOD $<-2$ in 1000 Genomes “pilot” regions
3. remove SNPs with VQSLOD $<-2.5$ in 1000 Genomes “strict” regions
4. remove SNPs with greater than 10% missingness across all samples
Furthermore for the purposes of this analysis, all insertions and deletions were removed from consideration.

4.7.2 Cryptic Relatedness

Apparent sample relatedness was computed using pairwise IBS computed in PLINK [49]. For this analysis, rare variants (minor allele frequency less than 1%) were removed, and remaining common variants were pruned by pairwise LD to reach approximate independence. Permutation testing for IBS stratification by affection status was implemented manually due to discrepancies in the corresponding function in PLINK.

4.7.3 Variant Quality Stratification Testing

Testing for stratification of variant quality by affection status was conducted by permutation testing on case/control labels. Within substrata (uncleaned data, PASS only, full QC only, clean novel functional SNPs only, etc.) frequency differentials were computed in 50000 permutations; only 549 samples with IBS $\hat{\pi} < 0.05$ were included in this analysis. Reported p-values are worst-case upper 95% confidence intervals on the estimated binomial distribution parameter according to the “binconf” module in the R Hmisc package [113].

4.7.4 Sequencing Call Rate Stratification Testing

Testing for stratification of call rate (or count) by contributing cohort was conducted by permutation testing on cohort label within case/control status. Permutation two-tailed p-value indicates the probability of observing as extreme a count in the same number of individuals randomly sampled from same outcome. P-value is from 10000 permutations; only 549 samples with IBS $\hat{\pi} < 0.05$ were included in this analysis. Reported mean counts were computed from observed samples; 95% confidence intervals were based on observed counts from permutations. Confidence intervals and nominal p-values do not perfectly coincide due to permutation approximations and the assignment of ties.
4.7.5 Novel Variant Selection

Following the analysis plan of [31], we selected variants for burden testing as follows. Variants had to be observed within the post-QC Illumina dataset amongst 549 IBS-unrelated individuals. Novel variant determination was based on prior observation of variants in 1000 Genomes Phase 3 [89]. Sequencing QC removed multiallelic variants, and due to additional concerns about stratification effects, indels were excluded from analysis. The frequency filter applied in [31] was relevant for their analysis due to limited sample size in their centenarian cohort; in this larger dataset, novelty made the 2% threshold redundant, as all variants with more than 20 minor alleles were previously observed.

4.7.6 Functional Annotation

The Ensembl Variant Effect Predictor [90] was used to generate functional annotations. Loss of function was defined as the union of the annotations “splice_acceptor_variant,” “splice_donor_variant,” “splice_region_variant,” “start_lost,” and “stop_gained.” Synonymous controls were defined as variants of the same annotation. Protein-altering variants included all loss of function variants, along with “missense” and “stop_lost” annotations; however, due to apparent strong association between variant sequencing quality and affectation status within this category, association results were not considered reliable and are not presented.

4.7.7 Fisher and Cochran-Mantel-Haenzel Tests

We test independence between schizophrenia case status and functional variant burden over non-functional variant burden with a Cochran-Mantel-Haenzel test [114]. Specifically, we test for independence conditional on the number of nonreference alleles present within the dataset for each variant. We conduct the analogous marginal Fisher test, summed across minor allele counts on [1, 10], for comparison. To further address possible convergence issues with these tests, we permute case status $10^4$ times and recompute p-values from the resulting bootstrapped distribution.

The Cochran-Mantel-Haenzel is effectively a dimensional extension of the Fisher test from [31]; we chose this method when substantial minor allele count-specific effects were observed to be correlated with relatedness between samples. For consistency, the CMH test assumes both independence of observations and homogeneity of effect within frequency strata: this is also interpretable as no
three-factor (affectation/functional/frequency) interactions, or odds ratios indistinguishable at indistinguishable at different levels of minor allele count. Testing for validity of these assumptions was conducted with IBS analysis (for independence) and the Breslow-Day statistic (for homogeneity of effect) ([114], implementation from [115]).

4.7.8 Extended Models of Burden Testing

Due to concerns about sample cryptic relatedness and informative stratification, which may be of particular concern for founder populations [116], we consider more flexible models for burden analysis that consider sample correlation structure. A thorough evaluation of existing methods and their limitations may be found in the Discussion. We highlight two methods for which we present results here.

EMMAX/EPACTS

The EMMAX software package [56] is widely used for handling cryptic relatedness in population-based samples for quantitative traits. Though the base package does not provide options for burden testing and is closed source, a modified version of the software has been incorporated into the software pipeline EPACTS [92]. We use the “emmaxCMC” collapsing test with per-individual novel synonymous allele count as covariate. It is important to note that both EMMAX and emmaxCMC treat a binary trait as a normally distributed, continuous outcome. Also note that in the Fisher-style analysis, synonymous allele count was acting both as a control for unbalanced design and for nonuniform sequencing efficiency between samples, whereas in this regression model it serves as a control only for the latter.

GMMAT

Among the available tools for directly handling binary trait association in the presence of stratification, we have selected GMMAT [91] due to feasibility and ease of use. GMMAT is an implementation of generalized linear mixed models with external covariance matrices specified for arbitrary numbers of random effects. We conduct affectation status permutations to generate empirical p-values, using the same regression model used in EMMAX/EPACTS. Note that this process takes several days for a single association test with permutations, and is not suitable for deployment in larger studies of more samples or more tests.
Chapter 5

Conclusion

Genome-wide association studies have provided the foundation for the last decade of human statistical genetics. The rapid expansion of this study model has led to certain methodological simplifications that provide fertile ground for improvement. Here, we consider complications of traditional variant association, and examine ways in which GWAS conventions may be formalized or improved. First, we provide the most extensive proof available of the inherent reliability of the quantitative trait GWAS model, and elaborate precisely what confounding factors lead to apparent difficulties in replication of GWAS signals. Second, we consider the handling of missing and uncertain data in genotype imputation, which can be improved to increase statistical power to detect association with minimal effort and no additional sample collection. Finally, we provide formal statistical analysis of novel variant detection in a bottlenecked population, and elaborate on several confounding effects specific to novel variation that have not previously been appreciated in the literature. Overall, this work offers substantial improvements to GWAS methods across the entirety of the study model.

We begin by confronting the profound lack of formal evaluation of the performance of quantitative trait GWAS. We collect all replication data across the entire field, and apply a rigorous and standardized replication scheme, most notably correcting for the Winner’s Curse in variant discovery. While it is now standard for GWAS papers to provide internal replication of discovered variants, the majority of publications fail to publish sufficient data to allow independent verification of their claims. Amongst the minority of papers that do provide proof of replication, unmatched ancestry between discovery and replication, as well as incorrect reporting of per-variant sample sizes reflecting the true rate of missingness at each variant, is sufficient to explain all aberrant replication rates. In
contrast to much public perception, we find and prove the field to be fundamentally reliable, and provide a simple tool for applying existing Winner’s Curse correction methods to users’ GWAS data.

We proceed to consider the appropriate handling of uncertain genotype data in association studies. The field at large employs several techniques for incorporating probabilistic genotypes into statistical tools, yet these tools ignore validated and formally correct method for uncertain data handling, Multiple Imputation, used in the statistical literature for decades. In part, this oversight is likely due to the inconvenience of modifying existing statistical tools to use MI; as such, we create a simple tool that creates an interface between GWAS tools and MI, removing the majority of the labor involved from the user. We show that using MI can substantially improve detection of true signals even in the context of underpowered studies, by correctly and explicitly introducing a variant component describing the variability in association effect introduced by genotype uncertainty.

In light of the Missing Heritability problem, the field has dedicated vast resources to the collection and study of rare variants; thus we investigate rare variant association with schizophrenia in a bottleneck population. As is altogether too often the case, we find that the study design flaws of standard GWAS magnify when considering specifically novel variation. Cryptic relatedness and IBD sharing between samples may lead to aberrantly high call rates of novel singletons once the standard post hoc relatedness filter is applied to samples. If cases and controls are nonrandomly sampled relative to these stratification effects, as is the case in this study, the resulting incidence rate of novel variation cannot be compared between groups. Furthermore, simple methods of adjusting for baseline novel variant discovery rate in samples, such as the Fisher test, must be carefully applied: the use of a single set of reference variants introduces correlated errors into downstream burden tests. Overall, we find that subtle deviations from random sampling, which may often be ignored in common variant GWAS, lead to irreparable flaws in the specific context of rare and novel variant burden tests.

While these studies have provided a substantial improvement to GWAS methodology, numerous issues remain. The Winner’s Curse correction data only represent quantitative trait GWAS; there remains the unlikely possibility that the reliability of quantitative trait GWAS is specific to this subset of the field; the tools created for this study are immediately extensible to binary trait GWAS results, and merely await the appropriate data collection. Multiple Imputation can only provide so much benefit: it is ultimately restricted by the quality of the underlying probabilities generated by genotype imputation. Yet we have shown that such genotype probabilities are themselves incorrect.
and biased in a predictable fashion. This issue may be addressed either by improvement of the imputation software by the appropriate investigators, or by creating a formal post hoc correction model for this bias, which would require a large pool of GWAS datasets to encompass all relevant parameters. Finally, these developments have provided reasonable improvements to several aspects of the GWAS model, but have not directly addressed the absence of expected explained genetic variance, arguably the most serious issue confronting the field. Rather, the replication project in particular has addressed certain issues raised about the fundamental viability of the field, and other investigators have shown that there remains substantial causal genetic variation that may be unlocked by truly colossal sample sizes. Yet there remains missing heritability, and no matter how wedded the field is to single variant association, at some point the challenge of arbitrary genome-wide epistasis testing must be addressed.

For this project we have chosen several particular GWAS vignettes to which to apply rigorous study, yet the GWAS model is complex and ever-growing. So long as the drive to produce GWAS-style results continues to expand, the field will continue to suffer from simplifications and oversights: there simply are not sufficient statistical and computational resources to address the cornucopia of study designs and challenges in GWAS. More narrowly, there is limited utility in providing computational tools and statistical solutions to problems in any biological field. The most difficult step remains, as ever, unaddressed: the tools and expertise must spread to investigators actually conducting these studies. This requires a combination of wise and user-friendly software design, as well as successful dissemination of the tools and their benefits. In the context of GWAS, the latter concern typically leads to tools only being adopted if they are developed by researchers attached to labs running large-scale GWAS meta-analyses. This is not in any way a judgment of the value of such tools, but it nevertheless leads to difficulties in producing tools in detached computational labs. It remains the case that the most fruitful study will be conducted by biologists and informaticians in close collaboration, each informing the others’ research.
Bibliography


[69] Yun Li et al. “MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes”. In: Genet. Epidemiol. 34.8 (Dec. 2010), pp. 816–834.

[70] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. “A flexible and accurate genotype imputation method for the next generation of genome-wide association studies”. In: PLoS Genet. 5.6 (June 2009), e1000529.


Appendix

Full Citation List for Chapter 2


[63] Dong-Hai Xiong et al. “Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups.” In: Am. J. Hum. Genet. 84.3 (Mar. 2009), pp. 388–98.


[118] Christopher Newton-Cheh et al. “Genome-wide association study identifies eight loci associated with blood pressure.” In: *Nat. Genet.* 41.6 (June 2009), pp. 666–76.


[139] Sophie I Candille et al. “Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations.” In: PLoS ONE 7.10 (2012), e48294.


L M Butcher et al. “Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays.” In: Genes Brain Behav. 7.4 (June 2008), pp. 435–46.


127


[230] Margit Hørup Larsen et al. “Genome-Wide Association Study of Genetic Variants in LPS-Stimulated IL-6, IL-8, IL-10, IL-1ra and TNF-α Cytokine Response in a Danish Cohort.” In: PLoS ONE 8.6 (2013), e66262.


Patrick Sulem et al. “Genome-wide association study identifies sequence variants on 6q21 associated with age at menarche.” In: *Nat. Genet.* 41.6 (June 2009), pp. 734–8.

Ken K Ong et al. “Genetic variation in LIN28B is associated with the timing of puberty.” In: *Nat. Genet.* 41.6 (June 2009), pp. 729–33.


Nancy L Heard-Costa et al. “NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium.” In: *PLoS Genet.* 5.6 (June 2009), e1000539.


