IDENTIFICATION OF CANDIDATE CAUSAL VARIANTS AND ESTIMATION OF
GENETIC ASSOCIATIONS IN GWAS AND POST-GWAS STUDIES

by

Laura Louise Faye

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Dalla Lana School of Public Health
Biostatistics Division
University of Toronto

© Copyright by Laura Louise Faye 2013
ACCURATE IDENTIFICATION OF CANDIDATE CAUSAL VARIANTS AND ESTIMATION OF GENETIC EFFECTS IN GWAS AND POST-GWAS STUDIES

LAURA LOUISE FAYE

Doctor of Philosophy

Dalla Lana School of Public Health

University of Toronto

2013

Abstract

Genome-wide association studies (GWAS) and next generation sequencing (NGS) studies are powerful high-throughput methods of scanning the human genome that have dramatically increased our ability to identify disease-causing genetic variants and estimate the magnitude of their effects. Leveraging the power of these technologies requires statistical methods tailored to the real world complexities of the data from these studies. Statistical methods developed during the era of small candidate gene studies fail to account for the extended scope of genome-wide studies, which encompasses: (1) discovery of disease-associated regions; (2) localization of associations to individual risk variants; and (3) estimation of effect size. In addition, high-
throughput sequencing used for large samples differs from traditional Sanger sequencing in that genotyping error varies substantially over a region, which can distort evidence used to identify the disease-associated variant.

In this thesis, I model these factors in order to increase accuracy of genetic effect estimation and accuracy of identification of disease-causing variants within disease-associated regions. I address these factors in three related settings: (1) GWAS study used alone to both discover and estimate the size of genetic effect at disease-associated variants; (2) GWAS study followed with sequencing to both discover an associated region via GWAS SNPs and estimate the size of genetic effect using the sequencing data; and (3) GWAS study with sequencing or imputation used jointly to identify candidate causal variants and estimate the corresponding effect sizes within an associated region. I develop novel statistical methods to address the specific localization and estimation problems encountered in each setting. Extensive simulation studies are used to explore the nature of these problems and to compare the performance of the new methods with the standard methods. Application to the Welcome Trust Case Control Consortium Type 1 Diabetes dataset and National Cancer Institute BPC3 aggressive prostate cancer study demonstrates the difference the methods make in the interpretation of evidence in these high-throughput studies.
Acknowledgements

This thesis would not have been possible without the guidance and support of my supervisors Professors Lei Sun and Shelley Bull. I am grateful for the guidance and mentoring from Professors France Gagnon, Mary Corey and Andrew Paterson. I thank Professor Peter Kraft for his support and guidance during my summer internship with his group. I extend a special thanks to Esther Berzunza for her encouragement and practical support as a STAGE trainee. My research was supported by a CIHR Doctoral Research Award. I would like to thank the STAGE program for internship support that made my work with Dr. Peter Kraft feasible. Finally, I thank my family for their support and encouragement throughout my graduate program.
Contents

List of Abbreviations ............................................................................................................................................ xiii
Chapter 1. Introduction ........................................................................................................................................... 1
  1.1 Background ..................................................................................................................................................... 1
    1.1.1 Discovery and refinement of association ............................................................................................ 2
    1.1.2 Low power ............................................................................................................................................... 3
    1.1.3 Replication ............................................................................................................................................... 4
    1.1.4 Selection .................................................................................................................................................. 5
    1.1.5 Genotyping error .................................................................................................................................... 6
  1.2 Outline of thesis ............................................................................................................................................ 7
  1.3 References for Introduction .......................................................................................................................... 10

Chapter 2. A flexible genome-wide bootstrap method that accounts for ranking- and threshold-selection bias in GWAS interpretation and replication study design ........................................ 10
  2.1 Abstract ....................................................................................................................................................... 10
  2.2 Introduction .................................................................................................................................................. 11
  2.3 Methods ...................................................................................................................................................... 14
    2.3.1 Genome-wide bootstrap shrinkage estimation .................................................................................... 14
    2.3.2 Derivation of the genome-wide bootstrap shrinkage estimator $\hat{\beta}_{boot(k)}$ ......................... 16
    2.3.3 Genome-wide bootstrap confidence interval construction ............................................................. 18
    2.3.4 Conditional likelihood methods ......................................................................................................... 20
  2.4 Application to the Wellcome Trust Case-Control Consortium Data ....................................................... 21
    2.4.1 Application methods ............................................................................................................................ 22
    2.4.2 Application Results ............................................................................................................................. 24
  2.5 Genome-Wide Simulation Studies ................................................................................................................ 25
    2.5.1 Design .................................................................................................................................................... 25
    2.5.2 Results .................................................................................................................................................. 27
  2.6 Discussion .................................................................................................................................................... 29
  2.7 Conclusion .................................................................................................................................................. 34
  2.8 Tables ......................................................................................................................................................... 35
Chapter 3. Two-stage study designs combining genome-wide association studies, tag SNPs, and sequencing: accuracy of genetic effect estimates ................................................................. 69

3.1 Abstract .............................................................................................................................. 69

3.2 Background ....................................................................................................................... 70

3.3 Methods ............................................................................................................................ 72

3.3.1 Scenario 1: common tag SNP, common causal SNP .................................................... 73

3.3.2 Scenario 2: common tag SNP, multiple rare causal SNPs ........................................... 74

3.3.3 Scenario 3: tag SNP with MAF 1–5%, multiple rare causal SNPs ............................... 75

3.4 Results .............................................................................................................................. 75

3.4.1 Scenario 1 ..................................................................................................................... 75

3.4.2 Scenario 2 ..................................................................................................................... 77

3.4.3 Scenario 3 ..................................................................................................................... 77

3.5 Discussion ........................................................................................................................ 78

3.6 Conclusions ..................................................................................................................... 79

3.7 Tables ................................................................................................................................ 80

3.8 Figures ................................................................................................................................ 82

3.9 References ........................................................................................................................ 82

Chapter 4. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification ...................................................................................................................... 85

4.1 Abstract ............................................................................................................................ 85

4.2 Introduction ....................................................................................................................... 86

4.3 Materials and Methods ................................................................................................... 90

4.4 Results .............................................................................................................................. 94
Tables

Table 2.1 Application of bias-reduction methods to WTCCC T1D: point estimates and confidence intervals of the case-control odds ratio ................................................................................................................... 35

Table 2.10 Single-SNP simulation results: performance of bias-reduced point estimators .... 44

Table 2.11 Single-SNP simulation results: Proportion of estimates within 10% or 25% of true log odds ratio ................................................................................................................................................. 45

Table 2.2 Genome-wide simulation: performance of bias reduced point estimators for low power alternative case ............................................................................................................................................ 36

Table 2.3 Genome-wide simulation: performance of bias reduced 95% confidence intervals for low power alternative case ........................................................................................................... 37

Table 2.4 Genome-wide simulation: Bias and reduction provided by rank for Naïve, NMLE and GW Bootstrap with correction ................................................................................................................... 38

Table 2.5 Genome-wide simulation: performance of bias reduced point estimators for log OR . 39

Table 2.6 Genome-wide simulation: Proportion of estimates within 10% or 25% of true effect size ......................................................................................................................................................... 40

Table 2.7 Genome-wide simulation: Sample size estimated using bias-reduced methods for replication study with 80% power ........................................................................................................... 41

Table 2.8 Genome-wide simulation: Sample size estimated for false positive SNPs using bias-reduced methods for replication study with 80% power ........................................................................... 42

Table 2.9 Computational Time Required For Genome-wide Bootstrap Point Estimate & Confidence Interval .............................................................................................................................................. 43

Table 3.1 Bias in genetic effect estimates for a common tag SNP and multiple rare causal SNPs .................................................................................................................................................. 80

Table 3.2 Bias in genetic effect estimates for a rare tag SNP and multiple rare causal SNPs...... 81

Table 4.1. Notation ..................................................................................................................... 110

Table 4.2 Parameters and parameter values of the main simulation studies ...................... 111
Table 4.3 Localization success rates for simulation Scenarios 1, 2, 3 ......................... 112

Table 4.4 Localization success rates for simulation Scenarios 5a .................................. 113

Table 4.5 Trends in power and localization success rate due to tagging and genotyping accuracy effect ................................................................. 114

Table 4.6 Alternative Localization success rates for simulation Scenarios 2, 3, 4 ........... 116
Figures

Figure 2.1 Uncorrected, normalized maximum likelihood and genome-wide bootstrap with correction bias-reduced estimates for a sample of SNPs meeting the WTCCC moderate association criterion ................................................................................................................... 46

Figure 2.2 Genome-wide simulation results for effect size estimation in null and low power alternative case ........................................................................................................................................ 47

Figure 2.3 Genome-wide simulation results for replication sample size estimation in low power alternative case ........................................................................................................................................ 48

Figure 2.4 Genome-wide simulation with threshold-based selection (alternative case – high power) ................................................................................................................................................... 50

Figure 2.5 Genome-wide simulation results (alternative case - low power) .......................................................................................................................... 51

Figure 2.6 Single-SNP simulation results ..................................................................................... 53

Figure 2.7 Comparison of normalized maximum likelihood and single-SNP bootstrap estimates ................................................................................................................................................... 54

Figure 2.8 Relationship between Likelihood, power and conditional likelihood ......................... 55

Figure 3.1 Distribution of estimates for a common tag SNP and a common causal SNP ............ 82

Figure 4.1 Tagging effect decreases localization success rates with or without the selection effect ........................................................................................................................................ 117

Figure 4.2 Low genotyping accuracy further reduces localization success rates with or without the selection effect ........................................................................................................................................ 118

Figure 4.3 Well-tagged causal SNPs sequenced with low accuracy are unlikely to be correctly identified even as sample size increases ................................................................................. 119

Figure 4.4 Naïve test statistics and re-ranking statistics for regions surrounding rs78246868 in the 8q24.21 region for association with prostate cancer risk ......................................................... 120

Figure 4.5 Naïve test statistics and re-ranking statistics for regions surrounding rs8071558 in the 17q24.3 region for association with prostate cancer risk ......................................................... 121

Figure 4.6 Distribution of SNP-specific read depth .................................................................... 127
Figure 4.7 Tagging effect decreases localization success rates with or without the selection effect, rare SNP ................................................................. 128

Figure 4.8 Tagging effect decreases localization success rates with or without the selection effect, high frequency SNP. ................................................................. 129

Figure 4.9 Genotyping accuracy effect further reduces localization success rates with or without the selection effect, rare SNP ................................................................. 130

Figure 4.10 Genotyping accuracy effect further reduces localization success rates with or without the selection effect, high frequency SNP ................................................................. 131

Figure 4.11 Well-tagged causal SNPs sequenced with low accuracy are unlikely to be correctly identified even as sample size increases, rare SNP ................................................................. 132

Figure 4.12 Well-tagged causal SNPs sequenced with low accuracy are unlikely to be correctly identified even as sample size increases, high frequency SNP ................................................................. 133

Figure 4.13 Tagging effect decreases localization success rates with or without the selection effect, 3 SNPs: 1 tag, 1 causal, 1 non-causal sequencing SNP ................................................................. 134

Figure 4.14 Tagging effect decreases localization success rates with or without the selection effect 5 SNPs: 1 tag, 1 causal, 3 non-causal sequencing SNPs ................................................................. 135

Figure 4.15 Tagging effect decreases localization success rates with or without the selection effect 7 SNPs: 1 tag, 1 causal, 5 non-causal sequencing SNPs ................................................................. 136

Figure 4.16 Location of five statistically independent associated sub-regions at 8q24.21 locus for the BPC3 aggressive prostate cancer GWAS ................................................................. 137

Figure 4.17 Naïve test statistics and re-ranking statistics for regions surrounding rs1016343 on 8q24.21 and rs34255287 on 11q13.3 for association with prostate cancer risk ................................................................. 138

Figure 4.18 Naïve test statistics and re-ranking statistics for regions surrounding rs7816007 and rs6983267 on 8q24.21 for association with prostate cancer risk ................................................................. 139

Figure 4.19 Naïve test statistics and re-ranking statistics for regions surrounding rs382434 on 8q24.21 for association with prostate cancer risk ................................................................. 140
Appendices

2.9.1 Appendix 1 Derivation of the correlation correction for the genome-wide bootstrap estimator ................................................................. 56

2.9.2 Appendix 2 Derivation of the MAF correction for the genome-wide bootstrap estimator 59

2.9.3 Appendix 3 Single-SNP bootstrap shrinkage estimation ............................................. 61

4.8.1 Appendix 4 Derivation of Distribution of GWAS and Tag SNP Test Statistics .......... 122

4.8.2 Appendix 5 Tagging and Coverage of GWAS Platforms ......................................... 125

4.8.3 Appendix 6 Additional Figures .................................................................................. 127
List of Abbreviations

AML adjusted medium likelihood
CI  confidence interval
GW genome-wide
GWAS genome wide association study
LD linkage disequilibrium
MAF minor allele frequency
MLE maximum likelihood estimate
NGS next generation sequencing
NMLE normalized maximum likelihood
OR odds ratio
RMSE root mean squared likelihood
SD standard deviation
SNP single nucleotide polymorphism
SS single-snp
T1D type 1 diabetes
WTCCC Wellcome Trust Case Control Consortium
Chapter 1. Introduction

1.1 Background

Genome-wide association studies cast a wide net for disease-associated genetic variants by genotyping a subset of SNPs selected to capture the majority of common variation across the genome. We refer to the SNPs directly genotyped by the GWAS platform as “tag SNPs” and all others as “non-tag SNPs”. Many non-tag SNPs are indirectly captured through correlation with a tag SNP. Power to detect a genetic effect via GWAS depends on correlation between the tag and causal SNPs. In this thesis we use the term “causal SNP” to refer to a SNP with a variant that predisposes an individual to a disease and “genetic effect” to refer to the change in disease risk or quantitative trait value associated with a unit change in the number of copies of a variant at the causal SNP.

SNPs not directly genotyped on the GWAS platform can be estimated using imputation or measured by sequencing, and we refer to such SNPs as “imputed SNPs” and “sequencing SNPs” respectively. Imputation estimates genotypes at non-tag SNPs using the GWAS data together with external information on the LD structure between tag and non-tag SNPs. Association can be tested at imputed SNPs, and the power of the test will depend on how accurately the SNP is imputed. Sequencing provides information at the base pair level, and has the potential for better genotyping accuracy and more complete coverage than imputation. Accuracy of sequencing depends on read depth (how many times the same stretch of genome is read); power for tests of association at sequenced SNPs will depend on sequencing accuracy.
Imputation is often used to for meta-GWAS when each original GWAS includes a different set of tag SNPs.

1.1.1 Discovery and refinement of association

In this thesis we consider the following series of steps in the discovery and refinement of a disease-gene association:

(1) Discovery: GWAS provides a cost-effective preliminary search for disease associated loci. Association at a tag SNP is likely to be indirect, because only a small proportion of SNPs in the genome are included on a GWAS platform. The estimate of genetic effect in the discovery-GWAS is conditional on the test statistic reaching genome-wide significance, which induces upward bias in the magnitude of the estimate. Replication in an independent sample resolves this bias, however researchers often require a less-biased estimate of genetic effect at the discovery stage (e.g. for design of replication and refinement studies). In chapter 2 we explore the effect of selection bias and develop a method to reduce bias at the discovery stage.

(2) Refinement: Imputation and sequencing provide the base-pair level information required to refine the location of the causal SNP and to estimate the size of the genetic effect at that causal SNP. When the GWAS discovery sample is used for imputation or sequencing, selection bias causes SNPs in high correlation with the tag SNP to appear to be more strongly associated than they actually are. This can cause upward bias in effect estimates and can distort evidence for the location of the causal SNP. Differential genotyping error common to imputation and next generation sequencing can also distort evidence for the location of the causal SNP. In chapter 3, we examine the effect of
selection bias in the GWAS discovery stage on the estimate of genetic effect at the sequencing refinement stage. In chapter 4 we examine the effect of selection bias and differential genotyping error on the ability to correctly identify the causal SNP.

The effect of selection bias on genetic effect estimates, known as the Beavis effect [1] or the winner's curse [2], is well-documented. The winner's curse was first explored with respect to auctions in which the highest bidder tends, on average, to over-pay for the actual value [3, 4]. Similarly, when the genetic effect at a marker is only estimated if it crosses the threshold for significance or is the maximum over a number of markers, then the magnitude of the estimate will tend to be upward biased for the true effect size. Goring et al [5], Sun and Bull [6] and Garner [7] demonstrated the winner's curse effect in genome-wide case-control studies and linkage scans.

1.1.2 Low power

Low power to detect modest effect sizes common to GWAS exacerbates the consequences of selection bias. GWAS tend to suffer from low power due to the stringent p-values required to correct for multiplicity, tag SNPs that imperfectly capture variation at the causal SNP, and inadequate sample size. Low MAF SNPs (MAF < 0.05) can suffer from low power due to low allele counts in GWAS samples and/or due to poor tagging on GWAS platforms designed to capture common variation. Low MAF SNPs (MAF < 0.05) are important in the search for disease-gene associations because theories of quantitative variation predict that causal variants tend to have lower MAFs [8]. The magnitude of selection bias is evident when comparing genetic effect estimates from high profile GWAS to those from follow-up studies [9]. Todd et al [10] attempted to replicate six loci with type 1 diabetes associations reported by the
WTCCC [11]: two failed to replicate and for those that did replicate, replication genetic effect odds ratio estimates were up to 70% closer to 1 (i.e. closer to the null). Easton et al [12] conducted a three-stage breast cancer GWAS in which they attempted to replicate 30 SNPs that reached significance in the combined stage I & II analysis: 20 failed to reach statistical significance in the combined stage I, II & III analysis. Nair et al [13] conducted a psoriasis GWAS with follow up. Of 21 SNPs that reached significance for association, only 10 reached statistical significance in the replication study.

1.1.3 Replication

Replication studies designed using upward biased effect sizes or inaccurately identified causal SNPs may fail to replicate a genuine association, because sample sizes computed using upward-biased effect estimates can leave replication studies under-powered to detect the association discovered in the original study. In this case, truly associated variants that reached statistical significance in the original study but not the replication study may be mistakenly discarded as false positives. Lohmueller et al [9] conducted a meta-analysis of 301 published studies for 25 reported associations and found evidence that implies that failed replication is more commonly due to a lack of power in the replication study than due to a false positive association in the original study. Moreover, when replication samples are drawn from a different population than the original sample, the linkage disequilibrium (LD) pattern across the associated region can differ between the original and replication studies. SNPs that captured association indirectly in the original sample via LD may not capture the association in the replication sample, causing replication to fail despite a genuine association.
While GWAS provides a cost-effective platform to detect associations and genomic regions that may harbour variants that play a role in phenotypic variation, and fine-mapping arrays (e.g. Immunochip) can provide very dense coverage of specific regions, they do not provide the base-pair level information often required to localize potential disease-causing variants. Next generation sequencing (NGS) and imputation can be implemented in GWAS samples in order to better localize the GWAS-identified associations. In some studies, only the regions surrounding the significant tag SNPs are sequenced. In other studies, the entire genome is imputed or sequenced and significant sequenced or imputed SNPs are only reported if they are found in the region surrounding a significant or marginally significant GWAS SNP [11,14]. In GWAS meta-analysis, SNPs may be genotyped in one study and imputed in another. Prior to the work reported here, the effect of selection for significant GWAS tag SNP test statistics on the effect estimate at the sequenced or imputed causal SNP had been an open question.

1.1.4 Selection

Selection at a GWAS tag SNP may also complicate localization of the causal SNP in the imputation or sequencing stage. While not directly examining the problem, results of a simulation by Wiltshire et al [15] suggest that localization conditional on GWAS-based region selection changes the probability that the causal SNP is top-ranked. Whether GWAS-informed region selection affects downstream localization is a research question that has not been well addressed. Similar to GWAS, in both sequencing and imputation studies associated regions often contain multiple genome-wide significant SNPs. High correlation among SNPs can make the causal SNP difficult to distinguish [16]. Among GWAS SNPs, researchers tend to report the SNP with the most significant test statistic from each region as the most likely causal SNP [17]. Although intuitive, this strategy may be misleading because the top-ranked SNP tends to be the
SNP that is most highly correlated with the GWAS tag SNP (i.e. in strongest linkage disequilibrium, LD) [16, 18].

1.1.5 Genotyping error

Genotyping error introduces noise that can obscure evidence for the causal SNP as well. When genotyping error is high, and many highly correlated SNPs capture some of the association indirectly, it can be difficult to pinpoint which SNP is in fact the causal SNP. Both low coverage sequencing studies and imputation studies tend to suffer from genotyping error. Imputation allows researchers to test for association at a large number of SNPs without additional cost, improving the chance that the causal SNP is in the set tested for association. However, imputation accuracy can be low, especially for low MAF SNPs which tend to be less well tagged by common SNPs [19]. SNPs not represented in the reference panel cannot be imputed. Sequencing can provide base-pair level information at a larger number of SNPs, but the accuracy depends on high sequencing depth which can be costly. One of the advantages of sequencing over imputation is the ability to capture low and very low MAF SNPs that are not well tagged on GWAS platforms. However, low MAF SNPs can also be more difficult to call in low coverage sequencing studies due to the low allele counts. Therefore, the genotype error rates among the SNPs analyzed can vary substantially, depending on the genotyping methods used (GWAS, imputation, low or high-coverage sequencing). The differential error rates in turn adversely affect the localization accuracy. In this thesis I use the term genotyping error to refer to sequencing and imputation error as well as GWAS genotype calling error.
1.2 Outline of thesis

In this dissertation, I consider the effects of multiple testing and differential measurement error on inference in genetic association studies. This thesis is prepared as a paper thesis with five sections: introduction, three journal articles and overall discussion. Chapter 2 examines the effect of tag SNP selection bias on genetic effect estimates in genome wide association studies (GWAS). Chapter 3 explores the scenario in which GWAS associated-regions are fine-mapped by next generation sequencing. Chapter 4 examines the effect of selection bias and measurement error on the power to distinguish potential disease-causing variants from among many highly correlated non-causal variants. Each article includes a separate introduction and discussion specific to that material with the overall introduction and discussion designed to present the unifying theme. The three papers are

(1) A flexible genome-wide bootstrap method that accounts for ranking and threshold-selection bias in GWAS interpretation and replication study design, published in Statistics in Medicine in 2011 (30:1898-1912) with Shelley Bull, Lei Sun and Apostolos Dimitromanolakis

(2) Two-stage study designs combining genome-wide association studies, tag single-nucleotide polymorphisms, and exome sequencing: accuracy of genetic effect estimates, published in BMC Proceedings (5:S64) in 2011 with Shelley Bull

(3) Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification, currently under minor revision at PLoS Genetics in collaboration with Lei Sun, Shelley Bull, Peter Kraft and Mitchell Machiela.
As first author on each of these papers, I developed the methods and analytical expressions, designed and carried out the simulation studies and wrote the manuscripts.

In the second chapter, I discuss the problem of selection bias in GWAS. When the same sample is used to both detect and estimate the size of a genetic effect, then conditional on selecting for a large test statistic, the magnitude of the estimate of effect will be biased away from the null. This bias can be induced by selecting for the top-ranked test statistics or selecting for test statistics that cross the threshold for genome-wide significance. Several likelihood-based methods have been proposed that model and help reduce threshold-based selection [20, 21, 22, 23], however these methods do not account for ranking-based selection. The bootstrap-based selection method proposed by Sun and Bull [24] is designed to account for threshold- and ranking-based bias in the genome-wide linkage setting. As part of thesis research, I extend this bootstrap method to the genome-wide association study and introduced two adjustments to improve performance. The first adjusts for variance associated with minor allele frequency and the second adjusts for correlation between the set of samples included in each bootstrap and the set of samples not included in each bootstrap. The results of simulation studies demonstrate that the bootstrap method gives more accurate estimates of the true underlying genetic effect. In this work I also show that follow-up studies designed with upward biased effect estimates tend to be under-powered to replicate true positives; the bootstrap estimate leads to adequate sample size calculations and a greater probability of replicating the original finding.

In the third chapter, I explore the effect of selection bias in two-phase study designs: GWAS in stage 1 followed by next generation sequencing in stage 2. This work is a contribution to the 17th Genetic Analysis Workshop (held in Boston in 2010). While the term two-stage often
refers to a study with two separate samples, in the BMC paper, I use two-stage to refer to a study with a two-step analysis in a single sample. I demonstrate that selecting for a large value in one random variable changes the expected value of correlated variables as well. I use the simulated Genetic Analysis Workshop sequencing datasets to demonstrate that selection for a genome-wide significant tag SNP test statistic induces bias not only in the tag SNP test statistic but also in the sequencing SNP test statistic. Intuitively, one would expect that the bias transferred to the sequencing SNP is only large if correlation between the tag and sequencing SNP is high. I show that bias in the sequencing SNP can be high, even when correlation between the tag and sequencing SNPs is only moderate.

In the fourth chapter, I formally quantify the effect of selection for significant tag SNPs as well as differential genotyping error between SNPs on power to identify the causal variant in sequencing and imputation studies. Bias induced by selection at a tag SNP carries over to sequenced or imputed variants. Variants in strongest correlation with the tag SNP will experience the largest bias in their effect estimate. In this work, I demonstrate that this same bias also affects the relative rank of SNPs over a region so that the SNPs in highest LD with the tag are most likely to be top-ranked. Random genotype-calling error can be high for low coverage sequencing and imputation studies, and varies among SNPs. I demonstrate that genotyping error can reduce power to identify the causal SNP by reducing power at the causal relative to the other sequencing and tag SNPs. I show that these two factors can substantially reduce power to identify the correct causal SNP. I then develop an analytical method to correct for these factors that restores much of the lost power to identify the correct causal SNP.
Chapter 5 provides an overall discussion, a summary of the connections between the three articles, and directions for future research and applications.

1.3 References for Introduction


11. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature 447:661-678


Chapter 2. A flexible genome-wide bootstrap method that accounts for ranking- and threshold-selection bias in GWAS interpretation and replication study design

2.1 Abstract
The phenomenon known as the winner’s curse is a form of selection bias that affects estimates of genetic association. In genome-wide association studies (GWAS) the bias is exacerbated by the use of stringent selection thresholds and ranking over hundreds of thousands of single nucleotide polymorphisms (SNPs). We develop an improved multi-locus bootstrap point estimate and confidence interval, which accounts for both ranking- and threshold-selection bias in the presence of genome-wide SNP linkage disequilibrium structure. The bootstrap method easily adapts to various study designs and alternative test statistics as well as complex SNP selection criteria. The latter is demonstrated by our application to the Wellcome Trust Case Control Consortium findings, in which the selection criterion was the minimum of the p-values for the additive and genotypic genetic effect models. In contrast, existing likelihood-based bias-reduced estimators account for the selection criterion applied to a SNP as if it were the only one tested, and so are more simple computationally, but do not address ranking across SNPs. Our simulation studies show that the bootstrap bias-reduced estimates are usually closer to the true genetic effect than the likelihood estimates and are less variable with a narrower confidence interval. Replication study sample size requirements computed from the bootstrap bias-reduced estimates are adequate 75-90% of the time compared to 53-60% of the time for the likelihood method. The bootstrap methods are implemented in a user-friendly package able to provide point
and interval estimation for both binary and quantitative phenotypes in large-scale GWAS.

2.2 Introduction

Bias in genetic effect estimates such as the odds ratio (OR), a phenomenon also known as the winner’s curse or the Beavis effect, can occur in both genome-wide linkage and genome-wide association studies (GWAS) [1, 2]. Because the same sample is typically used for both gene discovery and effect estimation, and the genetic effect is estimated only when the test for linkage or association at a genetic marker is significant, the estimate is on average larger in magnitude than the true value [3, 4]. Under situations of low power, a dataset that produces a test statistic larger than the critical value for significance is an extreme dataset: the observed test statistic is from the tail of its distribution. This selection effect is exacerbated by the use of stringent significance criteria and the modest power to detect small effects typical in GWAS.

Anti-conservative bias in genetic effect estimates is of particular importance for replication study design. GWAS cast a wide net to identify disease-associated SNPs, false positives are common, and so further investigation is usually required to confirm associations. When an estimate from an original GWAS is used to calculate the sample size required for a replication study, the bias in effect estimation tends to yield under-estimates of sample size, leaving the investigators with a study underpowered to replicate a true association [5-7]. When the SNP is not replicated, the investigator may assume the original finding was a false positive, when in fact it is a true positive that was not replicated due to inadequate sample size. Accurate estimates of OR in the original GWAS are important in order to avoid discarding disease-associated SNPs at
the replication stage. In addition to replication study design, an accurate estimate of genetic
effect is important for clinical interpretation, estimation of the proportion of heritability
explained by significant SNPs, and for meta-analyses.

Comparison of estimates from gene discovery studies with those from replication studies
demonstrates the upward bias in the magnitude of the original estimates, (hereafter referred to as
upward bias) [5, 8]. For example, among the 6 associated regions reported in the original
Wellcome Trust Case Control Consortium (WTCCC) Type 1 Diabetes (T1D) study, two failed to
be replicated by Todd et al. [9]; and even among those replicated, the ORs estimated from the
replication sample were up to 70% closer to 1 than the corresponding estimate reported in the
original discovery sample.

In the GWAS setting, selection of extreme test statistics can arise from application of strict
significance criteria as well as from ranking over the whole genome. Threshold selection bias
arises when the genetic effect is estimated only for SNPs with p-values below a threshold \( \alpha \).
Ranking bias arises when the genetic effect is estimated for SNPs with p-values that are among
the \( K \) smallest in the experiment with or without the threshold requirement [10]. However, even
when selection is by threshold, the ranking effect is present as we demonstrate in the following.
Both threshold and ranking selection contribute to the upward bias, and recent work suggests that
ranking bias can be more severe in some cases [11]. Modeling ranking or threshold bias requires
joint consideration of all SNPs and this can be a difficult task because of the complex correlation
structure arising from linkage disequilibrium (LD) among SNPs.

Although the winner’s curse phenomenon has been recognized for some time, practical
methods to correct for the selection bias have received attention more recently. The proposed
remedies that do not require an additional independent sample fall into two categories: bootstrap resampling-based methods [7, 12-14], and likelihood-based approaches [15-18]. Although proposed for GWAS where both ranking and threshold selection contribute to the bias, all four likelihood approaches model threshold bias independently for each SNP, and thus do not directly account for the effect of competition among SNPs on the estimate for the particular SNP of interest.

The genome-wide bootstrap method developed by Sun and Bull [12] focused on bias-reduced point estimation in the genome-wide linkage setting. Within each bootstrap sample, the entire genome-wide scan was repeated in order to capture the effect of ranking and correlation structure on selection bias in addition to the effect of the significance threshold. Observations not included in a particular bootstrap sample were treated as a corresponding independent out-of-sample estimate. Wu et al. [13, 19] evaluated the bootstrap method for a quantitative trait locus (QTL) linkage scan and detailed estimation methods for the case of multiple significant markers. Based on comparison studies of alternative bootstrap estimators, these authors recommended a so-called shrinkage estimator in cases of low power, but cautioned that this estimator tends to overcorrect at moderate to high power [12, 13]. Yu et al. [7] applied the weighted bootstrap estimator proposed by Sun and Bull [12] to association studies but considered only the top-ranked significant SNP. Jeffries [14] proposed the use of quantile-based bootstrap confidence intervals (CI) for bias-reduction; this method considers ranking bias only.

In this report, we extend the bootstrap shrinkage estimator of Sun and Bull [12] to the GWAS setting. We introduce two adjustments crucial to genome wide analysis that also serve to reduce over-correction in the estimates. The first accounts for differences in variance associated with the
minor allele frequency (MAF) of a SNP, and the second adjusts for the negative correlation between the within- and out-of-sample estimates used in the shrinkage factor calculation. Simulation studies demonstrate the substantial improvement provided by these corrections. We also develop a method to construct bootstrap CIs that accounts for both threshold and ranking bias. We begin in Section 2 with the development of the genome-wide bootstrap methods for GWAS data. Application of the methods to the WTCCC T1D GWAS [9] in Section 3 demonstrates feasibility and reveals practical differences between the bootstrap and likelihood methods. To better understand the differences and similarities between the two approaches, we perform simulation studies in Section 4. We close with discussion and practical recommendations in Sections 5 and 6.

2.3 Methods

2.3.1 Genome-wide bootstrap shrinkage estimation

Let \( \hat{\beta}_{N(k)} \) be the naïve estimate of the genetic effect of the \( k^{th} \)-ranked SNP reported in a GWAS \((k=1,\ldots,K, K \geq 1)\), selected by either rank- and/or threshold-based criteria in a set of observations we refer to as the original sample. The genetic effect could be the log odds ratio from a logistic regression for case-control data or the regression coefficient from a linear regression for quantitative outcomes. In each bootstrap sample, we apply the exact same selection criteria as in the original sample. For the \( k^{th} \)-selected SNP in the \( i^{th} \) bootstrap sample (where \( i \) indexes the bootstrap samples with at least \( k \) significant SNPs), let \( \hat{\beta}_{Dn(k)} \) be the within-sample estimate (based on subjects included in bootstrap sample \( i \)) and \( \hat{\beta}_{Ei(k)} \) the out-of-sample estimate (subjects not included in bootstrap sample \( i \)). Let \( N_{(k)} \) be the total number of bootstrap samples with at
least \( k \) SNPs passing the selection criteria, where \( N_{(k)} \geq N \) and \( N \) is the required minimal number of bootstrap samples, e.g. \( N = 100 \). (Note that if the original sample produced 2 significant SNPs, each bootstrap sample could have 0, 1, 2 or more significant SNPs.) A genome-wide (GW) bootstrap estimator for the \( k^{th} \) SNP selected in the original sample is

\[
\hat{\beta}_{\text{boot}(k)} = \hat{\beta}_{N_{(k)}} - \frac{1}{N_{(k)}} \sum_{i=1}^{N_{(k)}} (\hat{\beta}_{D_{i(k)}} - \hat{\beta}_{E_{i(k)}})
\]

(2.1)

This estimator is an implementation of the method of Sun and Bull [12] proposed for linkage data, called a shrinkage estimator because it shrinks the naïve effect estimate by a bootstrap estimate of the bias in \( \hat{\beta}_{N_{(k)}} \).

The out-of-sample estimate \( \hat{\beta}_{E_{(i)}} \) is meant to mimic an estimate obtained in an independent sample. However, although \( \hat{\beta}_{E_{(i)}} \) and \( \hat{\beta}_{D_{(i)}} \) are estimated from mutually exclusive sets of observations, they are negatively correlated because the original sample is finite and observations excluded from one sample must be included in the other. This constraint is ignored in the original shrinkage estimator (2.1) developed for linkage data. In addition, unlike linkage analysis, the variance of the parameter estimate for a SNP depends on its allele frequency, which therefore also affects the bias of estimates drawn from the tail region defined by the selection criteria. This needs to be taken into account because the \( k^{th} \)-ranked SNP within each of the bootstrap samples may differ from the SNP detected in the original sample and have a different minor allele frequency.

Let \( p_{(k)} \) be the MAF of the \( k^{th} \)-selected SNP in the original sample, and \( p_{i(k)} \) be the MAF...
for the $k^{th}$-selected SNP in the $i^{th}$ bootstrap sample ($i$ indexes the bootstrap samples with at least $k$ significant SNPs). Let $\hat{\beta}_{N(k)}$ be the estimate from the original sample for the $k^{th}$ ranked SNP selected in the $i^{th}$ bootstrap sample. (Note that $\hat{\beta}_{N(k)}$ and $\hat{\beta}_{N(k)}*$ are different; the latter denotes the estimate from the original sample for the $k^{th}$ ranked SNP selected in the original sample.) Let $\hat{\sigma}^2_{Di(k)}$ be the variance for $\hat{\beta}_{Di(k)}$, $\hat{\sigma}^2_{Ei(k)}$ the variance for $\hat{\beta}_{Ei(k)}$, and $\hat{\sigma}_{DEi(k)}$ their covariance, all estimated empirically. A modified bootstrap shrinkage estimator for the GWAS setting with adjustment for MAF and correlation between $\hat{\beta}_{Di(k)}$ and $\hat{\beta}_{Ei(k)}$ is:

$$\hat{\beta}_{\text{boot}(k)}^* = \hat{\beta}_{N(k)} - \frac{1}{N(k)} \sum_{i=1}^{N(k)} \left( \hat{\beta}_{Di(k)} - \hat{\beta}_{Ei(k)}^* \right) \sqrt{2p_{i(k)}(1-p_{i(k)})} \right)$$

(2.2)

where $\hat{\beta}_{Ei(k)}^* = \hat{\beta}_{Ei(k)} - \frac{\hat{\sigma}_{DEi(k)}}{\hat{\sigma}_{Di(k)}} \left( \hat{\beta}_{Di(k)} - \hat{\beta}_{N(k)} \right)$

(2.3)

We empirically estimate the variance and covariance of $\hat{\beta}_{Di(k)}$ and $\hat{\beta}_{Ei(k)}$ for each SNP individually by taking a separate set of bootstrap samples and computing the sample variance and covariance. In practice, we also truncate the shrinkage estimate at the null, so that the direction of association in the bias-reduced estimate cannot contradict the original test.

2.2.2 Derivation of the genome-wide bootstrap shrinkage estimator $\hat{\beta}_{\text{boot}(k)}^*$

The bootstrap sample and the out-of-sample observations are mutually exclusive and drawn from the finite original sample, which induces negative correlation between the within- and out-of-sample estimates. Their approximate joint distribution is:
\[
\begin{pmatrix}
\hat{\beta}_{Di(k)} \\
\hat{\beta}_{Ei(k)}
\end{pmatrix}
\bigg| _{\hat{\beta}_{Ni(k)} = B} \sim N\left(B, \begin{pmatrix}
\sigma^2_{Di(k)} & \sigma_{DEi(k)} \\
\sigma_{DEi(k)} & \sigma^2_{Ei(k)}
\end{pmatrix}\right) \quad \sigma_{DEi(k)} < 0 \tag{2.4}
\]

It follows that conditional on the observed within-sample estimate \(\hat{\beta}_{Di(k)}\)

\[
E\left(\hat{\beta}_{Ei(k)} \bigg| \hat{\beta}_{Di(k)} = d, \hat{\beta}_{Ni(k)} = B\right) = B + \frac{\sigma_{DEi(k)}}{\sigma^2_{Di(k)}}(d - B). \tag{2.5}
\]

We correct \(\hat{\beta}_{Ei(k)}\) by \(\frac{\sigma_{DEi(k)}}{\sigma^2_{Di(k)}}(d - B)\) to remove the correlation (see Appendix 1 for further details of the derivation).

A heuristic explanation of equations (2.3) and (2.5) is as follows. Consider a SNP from the original sample selected in bootstrap sample \(i\) that was also significant in the original sample. The test statistic \(\hat{\beta}_{Di(k)}\big/\hat{\sigma}_{Di(k)}\) is approximately normally distributed and centered around \(\hat{\beta}_{Ni(k)}\big/\hat{\sigma}_{Ni(k)}\), which was also larger than the critical value. Then, the probability of a significant test statistic for this SNP is high in any bootstrap sample, and so the quantity

\(\left(\hat{\beta}_{Di(k)} - \hat{\beta}_{Ni(k)}\right) = (d - B)\) in (2.5) will be on average small and so the effect of selection for

\(\hat{\beta}_{Di(k)}\big/\hat{\sigma}_{Di(k)} > Z_{\alpha/2}\) will have minimal effect on the distribution of \(\hat{\beta}_{Ei(k)}\). Now consider a SNP selected in the bootstrap sample that was not significant in the original sample. When

\(\hat{\beta}_{Di(k)}\big/\hat{\sigma}_{Di(k)}\) is greater than the critical value, \(\hat{\beta}_{Di(k)}\) will tend to be much larger than \(\hat{\beta}_{Ni(k)}\).

Then the quantity \(\left(\hat{\beta}_{Di(k)} - \hat{\beta}_{Ni(k)}\right) = (d - B)\) in (2.5) will tend to be large and positive and \(\sigma_{DEi(k)}\) will be negative, so that the effect on \(\hat{\beta}_{Ei(k)}\) will be large and negative. In the GWAS setting with
500,000 or more SNPs, SNPs with null or weak association in the original sample may be selected by chance in many of the bootstrap samples. Therefore it is important to correct the mean of $\hat{\beta}_{Ei(k)}$ for each SNP by subtracting $\sigma^2_{DEi(k)} \left( \hat{\beta}_{Di(k)} - \hat{\beta}_{Ni(k)} \right)$ from each $\hat{\beta}_{Ei(k)}$.

Under the assumption of an additively coded SNP and Hardy-Weinberg equilibrium (HWE), the genotypic variance is $2p(1-p)$, where $p$ is the MAF of the SNP of interest. Within the class of generalized linear regression models, the variance of the regression coefficient $\hat{\beta}$ is inversely proportional to the genotypic variance (see Appendix 2 for details), and selection bias is related to the variance via the power [12]. In order to account for allele frequency differences across SNPs, we rescale each term in the shrinkage factor by $\sqrt{2p_{i(k)}(1-p_{i(k)})}$.

### 2.3.3 Genome-wide bootstrap confidence interval construction

We construct a symmetric $(1-\alpha)$% CI around the point estimate, $\hat{\beta}^*_{boot(k)}$ as in equation (2.2), using a bootstrap estimate of the standard deviation of $\hat{\beta}^*_{boot(k)}$. This requires a 2-level bootstrap sampling scheme [20]: we draw $M$ 1st level bootstrap samples from the original data, and for each one, compute the bootstrap shrinkage estimate via a 2nd bootstrap level. The CI width is computed with the standard deviation of the $M$ bootstrap estimates.

Stated more formally, let $\hat{\beta}_{Dj(k)}$ be the within-sample estimate for the $k^{th}$-selected SNP chosen in the $j^{th}$ 1st level bootstrap sample, which serves as the naïve estimate $\hat{\beta}_{Nj(k)}$ in the 2nd level computation. Let $N_{j(k)} (\geq N)$ be the total number of 2nd level bootstrap samples, within the $j^{th}$ 1st level bootstrap sample, with at least $k$ SNPs selected. Let $\hat{\beta}_{Dj_{-i}(k)}$ and $\hat{\beta}_{Ej_{-i}(k)}$ be the
corresponding within- and out-of-sample bootstrap estimates with $p_{j,i(k)}$ the corresponding MAF (estimated from the original sample) for the $i^{th}$ 2nd level bootstrap sample nested within the $j^{th}$ 1st level bootstrap sample. Let $\hat{\sigma}^2_{Dj_i(k)}$ be the empirically estimated variance of $\hat{\beta}_{Dj_i(k)}$ and let $\hat{\sigma}_{DEj_i(k)}$ be the estimated covariance between $\hat{\beta}_{Dj_i(k)}$ and $\hat{\beta}_{Ej_i(k)}$. Then for each 1st level bootstrap sample $j = 1$ to $M$, we obtain a bootstrap estimate

$$\hat{\beta}_{boot.j(k)}^* = \hat{\beta}_{Nj(k)} - \frac{1}{N_{j(k)}} \sum_{i=1}^{N_{j(k)}} \left( \hat{\beta}_{Dj_i(k)} - \hat{\beta}_{Ej_i(k)}^* \right) \sqrt{2p_{j,i(k)}(1-p_{j,i(k)})}$$

for $j = 1 \ldots M$. (2.6)

where $\hat{\beta}_{Ej_i(k)}^* = \hat{\beta}_{Ej_i(k)} - \frac{\hat{\sigma}_{DEj_i(k)}}{\hat{\sigma}_{Dj_i(k)}} \left( \hat{\beta}_{Dj_i(k)} - \hat{\beta}_{Nj(k)} \right)$

In order to construct an asymptotic normal confidence interval, the point estimate should be normally distributed and the 2nd level bootstrap estimates should mimic this distribution. To achieve this, we apply a shift and a natural log transformation. Although $\hat{\beta}_{boot(k)}^*$ and $\hat{\beta}_{boot.j(k)}^*$ have similar shaped distributions, they centre on different values: $\hat{\beta}_{boot(k)}^*$ estimates the population genetic effect parameter $\beta$; $\hat{\beta}_{boot.j(k)}^*$ treats the original sample as the “population”, and estimates the “parameter” which is $\hat{\beta}_{N(k)}$. We therefore shift the mean of the $\hat{\beta}_{boot.j(k)}^*$ estimates to coincide with $\hat{\beta}_{boot(k)}^*$. The point estimate, $\hat{\beta}_{boot(k)}^*$, is not normally distributed (Figure 2.2), so to approximate a normal distribution we apply the log transformation. As before, we also rescale the $\hat{\beta}_{boot.j(k)}^*$ estimates.
Defining $s(\cdot)$ as the sample standard deviation, the estimated standard deviation of the $M$
transformed estimates is therefore

$$
\hat{\sigma}^{**}_{\text{boot}(k)} = s(\hat{\beta}^{**}_{\text{boot}(j(k))}),
$$

which is used to construct an asymptotic $(1 - \alpha)\%$ CI for

$$
\log(\hat{\beta}^{*}_{\text{boot}(k)}) \pm Z_{1-\alpha/2} \hat{\sigma}^{**}_{\text{boot}(k)}
$$

where $Z_{1-\alpha/2}$ is the $(1 - \alpha / 2)^{th}$ percentile of the standard normal distribution. We back-
transform with the exponential function to derive the $(1 - \alpha)\%$ CI for the genome-wide bootstrap
estimate $\hat{\beta}^{*}_{\text{boot}(k)}$.

### 2.3.4 Conditional likelihood methods

Several existing bias-reduction methods are based on maximum likelihood, conditional on the
test statistic exceeding a critical value. Under a case-control design the likelihood of Zollner and
Pritchard [15] is formulated with allele frequency and penetrance parameters but requires
external population prevalence data for constrained estimation. A related maximum likelihood
estimator, due to Xiao and Boehnke [18], formulates likelihood in terms of risk allele frequency
differences between cases and controls. Zhong and Prentice [16, 21] apply a standard case-
control logistic regression likelihood and construct a weighted average of the naïve log OR estimate obtained from the logistic model and the estimate from a model in which the naïve estimate is the median of the conditional distribution. We will refer to this estimator as the Adjusted Median Likelihood (AML) estimator. In addition, Ghosh *et al.* [17] consider a more general class of normally distributed estimators with a Wald-like test of significance suitable for association studies of either case-control or quantitative outcomes. Working at the test statistic level, they take the average of two quantities. The first is the MLE of the mean of the distribution of the test statistic. The second is the mean of a random variable that follows the distribution of the likelihood function for the mean of the distribution of the test statistic, normalized to be a proper density. Multiplying by the standard error of the naïve estimate transforms this quantity from the test statistic level to the genetic effect estimate. We will call this estimator the Normalized Maximum Likelihood Estimator (NMLE). Simulation studies [16, 17, 22] demonstrate that the AML and NMLE both perform better than the standard MLE under the conditional likelihood. Our simulation studies (see Appendix 3) show the NMLE performs better than the AML, based on smaller root mean square error, and therefore we report comparisons of the bootstrap to the NMLE method.

### 2.4. Application to the Wellcome Trust Case-Control Consortium Data

We re-analyzed the significant SNPs from the WTCCC T1D [9] sub-study that were also assessed in the Todd *et al.* [8] replication study. The four estimators investigated here as well as in the simulation studies of section 4 are as follows.

*Uncorrected*
Naïve: original naïve estimator

Bias-reduced, single-SNP likelihood method

NMLE: Normalized Maximum Likelihood Estimator of Ghosh et al. [17]

Bias-reduced, genome-wide bootstrap methods

GW Bootstrap without correction: Genome-Wide Bootstrap shrinkage estimator without correction, equation (2.1) in Section 2.1

GW Bootstrap: Genome-Wide Bootstrap shrinkage estimator with correction, equation (2.2) and (2.9) in Section 2.2

2.4.1 Application methods

We obtained the individual-level genotypes and phenotypes for the WTCCC T1D sub-study which includes 1963 cases and 2938 controls genotyped at 356,946 SNPs. The WTCCC T1D study reported strong associations (p-value < 5x10^{-7}) at 5 SNPs and moderate associations (p-value < 10^{-5}) at 7 SNPs. The reported SNPs were the most significant in their regions, in total 601 SNPs met the criteria for moderate association and 472 SNPs met the criteria for strong association. The replication study of Todd et al. [8] unequivocally validated 4 WTCCC association findings (p < 1.35x10^{-9}), including one SNP meeting the WTCCC criteria for moderate association (rs2542151) and 3 SNPs meeting the WTCCC criteria for strong association (rs12708716, rs17696736, rs2292239). The replication study also provided moderate evidence (p = 0.0231) for rs17388568 from the WTCCC moderate significance table. Although the WTCCC reported rs11171739 as the most significant SNP in the 12q13 region, rs2292239
was more significant in the replication study of Todd et al. [8], so we report the latter. The replication study OR estimates are all smaller than the original study naïve estimates, demonstrating the effect of the winner’s curse (Table 2.1).

We applied bootstrap and likelihood methods to selected SNPs meeting at least the moderate association criteria. In each of the GW bootstrap applications, we used N = 500 samples for the point estimate and M = 100 samples for the variance estimate. We applied the WTCCC selection criteria: minimum p-value of the trend test (1 df) and the genotypic test (2 df) less than the threshold value of 5x10^{-7} for SNPs included in their strong association table (rs17696736, rs2292239, rs12708716), and less than 1x10^{-5} for SNPs in their moderate association table (rs2542151, rs17388568). Following the WTCCC, we estimated the log odds ratio for the additive genetic effect. For the GW bootstrap method we excluded SNPs with MAF <1%, Hardy-Weinberg equilibrium p-value <10^{-7}, and genotyping call rate <99%. As in Ghosh et al. [17], we applied the NMLE method using the Wald test (1 df) from the additive logistic regression model. The analysis of Todd et al. [8] stratified by geographical subregion, while the WTCCC T1D analysis did not. To facilitate our comparisons, we conducted unstratified analysis of the WTCCC dataset. For this reason the stratified naïve and NMLE estimates reported by Todd et al. [8] and Ghosh et al. [17] differ slightly from the estimates we present. In Table 2.1, we provide bias-reduced estimates for the five SNPs replicated by Todd et al. [8]. In Figure 2.1, we compare the bootstrap and likelihood bias-reduced estimates using the WTCCC moderate association threshold for a sample of SNPs with p-values between 1x10^{-5} and 1x10^{-8} that were randomly selected from the 601 significant SNPs.
2.4.2 Application Results

The bootstrap and likelihood methods shrunk the naïve estimate toward the null for 4 out of 5 replicated SNPs (Table 2.1). The amount of shrinkage for the 3 highly significant SNPs was small and similar among the methods. This reflects intuition: the small observed p-value implies high power to detect these SNPs at the chosen significance threshold and similar studies would also tend to produce significant findings. Therefore, the effect of threshold-selection is minimal. In contrast, the reduction in the least significant SNP was larger and varied among the methods. A p-value close to the threshold for selection implies lower power and more bias.

Analysis of all SNPs significant at the moderate association threshold shows the same trend (Figure 2.1). At highly significant p-values, the bias-reduced estimates are similar and close to the naïve estimate. The bootstrap corrects for ranking as well as threshold bias, and therefore corrects somewhat more than the likelihood. Both methods exhibit most shrinkage when the observed p-value is close to the threshold for selection ($1 \times 10^{-5}$). The likelihood tends to shrink estimates much more than the bootstrap for these marginally significant SNPs, and so we expect the difference between methods to be greatest for lower power SNPs.

Based on simulation studies described in the following section, which allow us to make general conclusions not possible from a single dataset even with replication, we note that the uncorrected GW bootstrap method tends to be overly conservative and is therefore not recommended. The corrected GW bootstrap estimates generally have smaller mean squared error and shorter confidence intervals than the likelihood estimates. Although the naïve estimates have similarly narrow CIs, coverage is expected to be unsatisfactory because they are based on a biased estimate. Selection bias can be severe for lower power SNPs; in section 4, we therefore
present simulations designed to evaluate the methods in the case of lower power.

2.5 Genome-Wide Simulation Studies

2.5.1 Design

We conducted genome-wide simulations to evaluate the performance of the GW bootstrap method and the NMLE when selection is threshold-based. In order to capture the effect of realistic correlation structure among SNPs, we fixed the actual WTCCC genotypes and simulated case-control status under a log additive multiple-SNP logistic model based on varying genetic effect sizes for 5 associated SNPs. In one configuration the associated SNPs had MAF values and corresponding odds ratios (ORs) as described in Table 2.2, which yielded a probability of selection for each SNP of 7%, 11%, 13%, 30% and 49% respectively. In a second configuration the same set of associated SNPs had the same MAF values but larger corresponding ORs so that the probability of selection for each SNP was 60%, 70%, 77%, 90% and 99% respectively (Table 2.5). The remaining 356,941 SNPs in the WTCCC genotype dataset had no genetic effect in our simulation model.

We simulated datasets until there were at least 500 replicates with significant estimates for each true positive SNP. We applied a threshold-based significance criterion: a trend test p-value less than the WTCCC strong significance level of $5 \times 10^{-7}$. For each dataset we computed the GW bootstrap bias-reduced estimates and confidence intervals with MAF and correlation corrections, the GW bootstrap without correction, and the NMLE of Ghosh et al. [17]. We applied the GW bootstrap using 100 level 1 bootstrap samples, and 100 level 2 bootstrap samples. For each simulation we computed the summary statistics and examined the distribution of the point
estimates. We compare the coverage and precision (i.e. average width) of confidence intervals for the naïve, NMLE and GW bootstrap with correction.

To assess the utility of bias-reduced estimates in planning replication studies, we calculated the sample size required for a replication study with 80% power for the naïve and bias-reduced estimates in turn and compared these values with the actual sample size required (using formulas presented by Slager and Schaid [23]). We assumed the replication sample would be drawn from the same population as the original sample, with an equal number of cases and controls. We examined each associated true positive SNP in our simulation model individually. For each SNP, we estimated the sample size required to replicate the association using the naïve, GW Bootstrap and NMLE estimates (as obtained above) in each of the simulated datasets in which that SNP was significant. For comparison, we computed the power to select the SNP of interest based on the generating OR and the estimated sample size. We report how often the estimated sample size is large enough to achieve 80% power.

At the replication stage, the type I error can be controlled by choice of an appropriate significance criterion, and most false positive SNPs would be eliminated by failure to replicate. However, in order to exclude a SNP, it would be desirable for a 95% CI in the replication sample to be sufficiently precise to exclude an OR that would be of interest, e.g. an OR of 1.15 or larger.

Therefore, for each false positive SNP in the simulated datasets we computed the sample size for replication, using the estimated log OR, and then assuming the computed sample size, calculated the corresponding standard error for the log OR under the null hypothesis of no association. Assuming a 95% CI half-width of log(1.15) = 0.140 would be sufficient to exclude an OR of 1.15, we determined how often this precision was achieved.
In order to evaluate the ability of each method to adapt to the ranking effect, we also compared the mean absolute bias of the naïve estimate (i.e. the amount of shrinkage that is required) to the amount of shrinkage provided by the NMLE and the GW bootstrap stratified by rank.

2.5.2 Results

The naïve estimate for true positive SNPs is upward biased which produces large RMSE, while the NMLE over-corrects slightly with a large RMSE due to high variance (Figure 2.2, Table 2.2). The GW bootstrap with correction has somewhat larger downward bias in comparison to the NMLE, but has smaller RMSE because the majority of bootstrap estimates form a mode just below the true value, and most estimates are within 25% of the true value (Table 2.6). For false positive SNPs, the NMLE method tends to have a higher proportion of estimates close to the true value, however bias and RMSE are larger than the GW bootstrap for MAF <0.4 (Figure 2.2, Table 2.5 and 2.6). The GW bootstrap without correction shrinks most estimates to a value close to the null. As a result, the method performs well for false positives where the true genetic effect is in fact at the null, but grossly over-corrects for true positive SNPs. The MAF and correlation adjustments reduce this over-correction by 65-100% in the low power case. Similar results were obtained for configuration 2 at high power (Table 2.5, Figure 2.4).

The naïve 95% confidence intervals for true positive SNPs with low power give poor coverage because they are centred around highly biased point estimates (Table 2.3). The likelihood bias-reduced confidence intervals have near-nominal coverage, and are wider than the naïve confidence intervals. A wide interval is required for nominal coverage due to the high variance of the likelihood point estimate. The genome-wide bootstrap confidence intervals have
higher than nominal coverage and are more precise than the likelihood intervals (i.e. nearly half the width).

Depending on the effect size and the MAF of the SNP, 75-90% of the time sample size calculations based on GW bootstrap estimates were at or above the size required for 80% power (Figure 3, Table 2.7 and Figure 2.5). On average, the sample size based on the GW bootstrap estimate tended to over-estimate the actual sample size required for replication by a factor of 1.3 to 2.4. In contrast, sample sizes calculated using the NMLE likelihood estimates under-estimated the sample size 38-52% of the time (depending on study design parameters), while the average sample size was larger than that for the GW bootstrap sample size calculations, due to the large variance of the likelihood estimate.

For the false positive SNPs, the sample size for the replication study given by the GW bootstrap was on average 1.7 to 2.6 times larger than that required and was sufficiently large in over 93% of datasets to exclude effect sizes of interest in the replication study (for MAF > 10%). In contrast, the sample sizes given by the NMLE method were slightly less likely to be adequate but were on average 5.4 to 6.9 times larger than required due to many very large estimates of sample size (Table 2.8).

Even when SNP selection is by threshold, there is also a ranking effect as SNPs compete for rank, and so the bias in the naïve estimate depends on the rank achieved by the SNP (Table 2.4). The OR tends to require more correction when a SNP achieves a higher versus a lower rank. The NMLE method, however, overcorrects when the SNP has a lower rank. The GW bootstrap corrects all ranks by about the same amount, which is closer to the correction that is required. As
a result, the GW bootstrap has smaller RMSE across ranks.

2.6 Discussion

In most GWAS, threshold and ranking selection are both present: estimates are reported only for significant SNPs leading to bias away from the null, and for SNPs with the same true effect size, top ranked SNPs are more highly biased than lower-ranked SNPs [10]. This is borne out in our simulations: bias is greater when rank is higher (Table 2.4). The likelihood methods address only the threshold effect, and tend to shrink lower-ranked SNPs the most. The GW bootstrap models both ranking and threshold bias: a higher rank among SNPs of similar significance implies greater ranking bias, even if highly significant p-values imply little threshold bias. On the other hand, a lower rank implies little ranking bias, but the less significant p-values among lower-ranked SNPs implies lower putative power and a larger threshold bias.

The bootstrap method of Sun and Bull [12] was developed for the multi-SNP linkage setting. When simply applied, as is, to the GWAS setting it seriously over-corrects, as demonstrated by our simulations. The MAF and correlation adjustments we develop in this report optimizes the bootstrap for genome-wide association analysis, providing significantly more accurate estimates for associated SNPs.

In Appendix 3 we present results for a bootstrap method that considers only a single SNP at a time, and is analogous to the single-SNP likelihood methods. The single-SNP bootstrap performed well only in cases of high power, where selection bias is less of a concern. In cases where ranking bias is more pronounced than selection bias or is the only bias present, as in the case of selection by rank, single-SNP methods may not reduce the selection bias sufficiently.
We recommend the GW bootstrap with correction in preference to both the single-SNP bootstrap and likelihood methods, especially when power is low to moderate.

Confidence intervals for bias-reduced estimates based on the GW bootstrap similarly account for both threshold and ranking bias, but they need to be interpreted carefully: while the SNP is deemed significant at a genome-wide criterion in the original analysis, the bias-reduced CI could nevertheless cover zero. This does not imply the SNP is no longer significant: the original p-value stands because the confidence interval – hypothesis testing duality does not hold for bias-reduced intervals. However, the bias-reduced CI width does indicate the level of uncertainty in the bias-reduced estimate and may be interpreted as the CI that would be obtained in an independent sample of the same sample size as the original.

In a genome-wide scan, the genetic model is often unknown, so multiple tests may be used in order to determine if a SNP shows significance under one of several genetic models. The WTCCC study, for example, used the minimum of the p-values from the trend and genotypic tests. To apply the likelihood method in this case would require a power function to be specified and maximized over the parameter of interest, which may not be straightforward. As long as the selection step can be automatically applied to the original dataset, it can be applied in the bootstrap. In addition, the bootstrap is flexible enough to be extended in several respects, for example, multiple-SNP tests, tests for models that include covariates and interactions, or selection criteria that incorporate external information such as gene ontologies or pathway information.

The genome-wide bootstrap tends to be conservative for true positive SNPs. In planning appropriately powered replication studies, a slightly conservative estimate is desirable. Under-
powered studies due to sample size calculations based on optimistic estimates has been cited as a major cause of failure to replicate [3]. As there is generally no way to distinguish between true and false positives, then when true and false positives have similar naïve estimates and MAF, they will have similar bias-reduced estimates. Although estimators that perform better for false positives can be specified [12], smaller bias for false positives comes at the expense of larger bias for true positives and vice versa. In planning replication studies, a good estimate for true positive SNPs is more useful than a good estimate for false positive SNPs. Consider a scenario in which the initial genome-wide scan identified a number of significant SNPs with some true positives and some false positives of similar naïve log OR. If the true positive genetic effect estimates are accurate then the replication study using the corresponding sample size estimates will be adequately powered to replicate truly associated SNPs. On the other hand, if true positives are severely under-estimated in order to achieve more accurate false positive estimates, the over-estimated sample size may be too large for the replication study to be feasible. The GW bootstrap performs well for true positives and our simulations show that sample size computed for replication is usually adequate.

The bootstrap method requires a sufficient number of resamples to accurately estimate the effect of selection. If too few bootstrap samples are used, the random variability of bootstrap sampling may produce a poor estimate. As variability depends on the dataset, it is advisable to run the bootstrap cumulatively with an increasing number of bootstrap samples until stability is achieved. We recommend a minimum starting point of 100 bootstrap samples. Bootstrap methods are computationally intensive, because the same analysis applied to the original dataset must be applied to each bootstrap sample. If 100 level 1 and 100 level 2 bootstraps (10,000
bootstraps in total) are used to construct a CI, this is computationally equivalent to repeating the original analysis 10,000 times. Computation time scales linearly with the number of bootstraps required, after time to read in the dataset is taken into account. Genome-wide bootstrap analysis of the data simulated under the low power alternative case described in Section 4.1, with selection via the trend test took 6.5 minutes to obtain a point estimate only and 3.5 hours to obtain confidence intervals using 100 bootstrap samples for both levels of bootstrapping on a moderately powered unix machine (2 x Quad Core Intel Xeon E5540 2.53GHz, 16 GB RAM) using parallel processing (Table 2.9). Computational time for the bootstrap depends on the complexity of the original analysis and number of significant SNPs in the original analysis. Closed form calculations (e.g. linear regression, trend test) are completed more quickly than calculations requiring numerical optimization (e.g. logistic regression). Inclusion of covariates may increase complexity and run time, in such cases applying the bootstrap method to the residuals can reduce computational requirements.

The GW bootstrap method would also be appropriate with multiple covariates, whether genetic or non-genetic, under the principle that analysis in each bootstrap replicates what was done in the original analysis. The BR-squared genome-wide bootstrap software can handle regression models with multiple covariates, but at present is limited to bias-reduction of one SNP at a time. Multiple SNP covariates can be included in the regression, but bootstrap shrinkage has to be done for each significant SNP, one parameter at a time, while the others are held fixed. For example, if instead of selecting SNPs with small p-values, one were to fit all possible gene-by-gene interaction models (i.e. 3 parameters: 2 additive main effects, 1 interaction) and select the models with the smallest p-values, the bootstrap method as described in this paper would not be
able to shrink all 3 of these parameters at once. The same limitation exists with the likelihood method.

While separate analysis of each SNP remains the norm for most GWAS, there is increasing interest in joint analysis of multiple SNPs via regression models. If multiple correlated SNPs are associated with a phenotype, the marginal effect estimates from single-SNP analysis may suffer from omitted variable bias. In this case, application of the genome-wide bootstrap we describe would reduce the winner’s curse, but not address the omitted variable bias. Although a multiple-SNP regression model that includes “significant” SNPs selected from the single-SNP analysis will yield joint estimates less subject to this type of mis-specification bias, the effect estimates for the SNPs that remain significant would still suffer from bias due to the winner’s curse. To reduce both types of bias that are present when associated SNPs are highly correlated, application of the bootstrap principle would require repeating the single SNP selection procedure and multi-SNP regression model fitting in each of the bootstrap samples. Further work to evaluate the nature and extent of the winner’s curse bias in the multiple SNP regression setting is warranted.

In whole genome sequencing studies, selection bias can be more severe than in GWAS because one is maximizing over a larger number of tests. The genome-wide bias reduction methods we describe are directly applicable to common variants from sequencing studies. For rare variants, the bootstrap approach applies in principle, but details of method evaluation and implementation remain open.
2.7 Conclusion

For bias-reduced estimation in GWAS, we recommend the GW bootstrap (with the GWAS-specific adjustments) particularly in cases where complex testing procedures are applicable. The GW bootstrap estimates are slightly conservative, which makes replication study sample sizes computed from the estimates adequate most of the time. Ranking can be a non-trivial source of bias, and among the methods we considered, only the genome-wide method accounts for the effect of all SNPs on the ranking of a SNP of interest. In some studies complex selection criteria make specification of the likelihood difficult, while the bootstrap is adaptable in that any well-defined criteria can be applied in each bootstrap sample. Efficient software implementing the bootstrap method is available [24, 25].
### 2.8 Tables

**Table 2.1** Application of bias-reduction methods to WTCCC T1D: point estimates and confidence intervals of the case-control odds ratio

<table>
<thead>
<tr>
<th>SNP</th>
<th>α</th>
<th>p-value</th>
<th>Naive</th>
<th>NMLE</th>
<th>GW Boot w/o correction</th>
<th>GW Boot w/ correction</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs17696736</td>
<td>5.00E-07</td>
<td>2.17E-15</td>
<td>1.39</td>
<td>1.39</td>
<td>1.35</td>
<td>1.37</td>
<td>1.16</td>
</tr>
<tr>
<td>rs2292239</td>
<td>5.00E-07</td>
<td>4.87E-10</td>
<td>1.31</td>
<td>1.27</td>
<td>1.23</td>
<td>1.28</td>
<td>1.28</td>
</tr>
<tr>
<td>rs12708716</td>
<td>5.00E-07</td>
<td>9.24E-08</td>
<td>0.79</td>
<td>0.89</td>
<td>0.89</td>
<td>0.83</td>
<td>0.83</td>
</tr>
<tr>
<td>rs17388568</td>
<td>1.00E-05</td>
<td>5.01E-07</td>
<td>1.26</td>
<td>1.17</td>
<td>1.11</td>
<td>1.18</td>
<td>1.08</td>
</tr>
<tr>
<td>rs2542151</td>
<td>1.00E-05</td>
<td>1.89E-06</td>
<td>1.29</td>
<td>1.13</td>
<td>1.11</td>
<td>1.19</td>
<td>1.29</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNP</th>
<th>α</th>
<th>p-value</th>
<th>Naive</th>
<th>NMLE</th>
<th>GW Boot w/o correction</th>
<th>GW Boot w/ correction</th>
<th>Replication</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs17696736</td>
<td>5.00E-07</td>
<td>2.17E-15</td>
<td>1.28-1.51</td>
<td>1.28-1.51</td>
<td>0.80-2.27</td>
<td>1.26-1.54</td>
<td>1.09-1.23</td>
</tr>
<tr>
<td>rs2292239</td>
<td>5.00E-07</td>
<td>4.87E-10</td>
<td>1.20-1.42</td>
<td>1.13-1.42</td>
<td>1.00-1.51</td>
<td>1.20-1.38</td>
<td>1.20-1.36</td>
</tr>
<tr>
<td>rs12708716</td>
<td>5.00E-07</td>
<td>9.24E-08</td>
<td>0.72-0.86</td>
<td>0.73-1.01</td>
<td>0.90-1.39</td>
<td>0.77-0.87</td>
<td>0.78-0.89</td>
</tr>
<tr>
<td>rs17388568</td>
<td>1.00E-05</td>
<td>5.01E-07</td>
<td>1.15-1.37</td>
<td>1.07-1.36</td>
<td>0.91-1.34</td>
<td>1.14-1.24</td>
<td>1.01-1.15</td>
</tr>
<tr>
<td>rs2542151</td>
<td>1.00E-05</td>
<td>1.89E-06</td>
<td>1.16-1.43</td>
<td>0.99-1.41</td>
<td>0.91-1.34</td>
<td>1.13-1.26</td>
<td>1.19-1.40</td>
</tr>
</tbody>
</table>

**Note:** Discovery samples are from the original WTCCC (2007) T1D GWAS; follow-up samples and results are from the replication study of Todd and others (2007). The association p-value is the minimum of the 1 df trend and the 2 df genotypic association tests observed in the WTCCC samples for SNPs meeting criteria for strong significance (p-value < 5E-7) and moderate significance (p-value < 1E-5). Point estimates and confidence intervals were calculated using the discovery WTCCC samples for the uncorrected (Naïve), normalized maximum likelihood estimate (NMLE), genome-wide bootstrap methods without correction (GW Boot w/o correction) and with correction (GW Boot w/ correction); Replication estimates use follow up samples only.
Table 2.2 Genome-wide simulation: performance of bias reduced point estimators for low power alternative case

<table>
<thead>
<tr>
<th>Pr(p&lt;α)</th>
<th>MAF</th>
<th>OR</th>
<th>10th</th>
<th>90th</th>
<th>10th</th>
<th>90th</th>
<th>10th</th>
<th>90th</th>
</tr>
</thead>
<tbody>
<tr>
<td>7%</td>
<td>0.33</td>
<td>1.16</td>
<td>1.29</td>
<td>1.16</td>
<td>1.05</td>
<td>1.17</td>
<td>1.25</td>
<td>1.34</td>
</tr>
<tr>
<td>11%</td>
<td>0.45</td>
<td>1.16</td>
<td>1.26</td>
<td>1.15</td>
<td>1.03</td>
<td>1.14</td>
<td>1.23</td>
<td>1.29</td>
</tr>
<tr>
<td>13%</td>
<td>0.46</td>
<td>1.16</td>
<td>1.26</td>
<td>1.16</td>
<td>1.03</td>
<td>1.15</td>
<td>1.24</td>
<td>1.30</td>
</tr>
<tr>
<td>30%</td>
<td>0.29</td>
<td>1.2</td>
<td>1.29</td>
<td>1.18</td>
<td>1.04</td>
<td>1.17</td>
<td>1.26</td>
<td>1.34</td>
</tr>
<tr>
<td>49%</td>
<td>0.13</td>
<td>1.34</td>
<td>1.42</td>
<td>1.28</td>
<td>1.06</td>
<td>1.23</td>
<td>1.37</td>
<td>1.50</td>
</tr>
</tbody>
</table>

**Mean of Estimates (OR)**

**Percentiles (OR)**

<table>
<thead>
<tr>
<th></th>
<th>Naive</th>
<th>NMLE</th>
<th>GW Boot w/o corr</th>
<th>GW Boot w/corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(p&lt;α)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7%</td>
<td>1.25</td>
<td>1.34</td>
<td>1.06</td>
<td>1.32</td>
</tr>
<tr>
<td>11%</td>
<td>1.23</td>
<td>1.29</td>
<td>1.06</td>
<td>1.27</td>
</tr>
<tr>
<td>13%</td>
<td>1.24</td>
<td>1.30</td>
<td>1.06</td>
<td>1.27</td>
</tr>
<tr>
<td>30%</td>
<td>1.26</td>
<td>1.34</td>
<td>1.06</td>
<td>1.31</td>
</tr>
<tr>
<td>49%</td>
<td>1.37</td>
<td>1.50</td>
<td>1.09</td>
<td>1.47</td>
</tr>
</tbody>
</table>

**Relative Bias (log OR)**

<table>
<thead>
<tr>
<th></th>
<th>Naive</th>
<th>NMLE</th>
<th>GW Boot w/o corr</th>
<th>GW Boot w/corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(p&lt;α)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7%</td>
<td>0.13</td>
<td>0.11</td>
<td>0.15</td>
<td>0.08</td>
</tr>
<tr>
<td>11%</td>
<td>0.10</td>
<td>0.08</td>
<td>0.14</td>
<td>0.04</td>
</tr>
<tr>
<td>13%</td>
<td>0.11</td>
<td>0.08</td>
<td>0.14</td>
<td>0.05</td>
</tr>
<tr>
<td>30%</td>
<td>0.10</td>
<td>0.10</td>
<td>0.16</td>
<td>0.06</td>
</tr>
<tr>
<td>49%</td>
<td>0.10</td>
<td>0.15</td>
<td>0.30</td>
<td>0.14</td>
</tr>
</tbody>
</table>

**RMSE (log OR)**

<table>
<thead>
<tr>
<th></th>
<th>Naive</th>
<th>NMLE</th>
<th>GW Boot w/o corr</th>
<th>GW Boot w/corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pr(p&lt;α)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7%</td>
<td>0.13</td>
<td>0.29</td>
<td>0.24</td>
<td>-0.11</td>
</tr>
<tr>
<td>11%</td>
<td>0.10</td>
<td>0.10</td>
<td>0.16</td>
<td>0.06</td>
</tr>
<tr>
<td>13%</td>
<td>0.08</td>
<td>0.08</td>
<td>0.14</td>
<td>0.05</td>
</tr>
<tr>
<td>30%</td>
<td>0.08</td>
<td>0.29</td>
<td>-0.14</td>
<td>0.83</td>
</tr>
<tr>
<td>49%</td>
<td>0.15</td>
<td>0.30</td>
<td>0.30</td>
<td>0.14</td>
</tr>
</tbody>
</table>

**Note:** 500 genome-wide case control datasets in which the SNP of interest was significant at p-value threshold α=5E-7 were simulated using the WTCCC T1D genotypes and OR as indicated. Mean of Estimates (average on OR scale), 10th and 90th percentiles (OR scale), Relative bias (average difference between estimated and true log OR divided by true log OR), and root mean squared error (RMSE of log OR estimates) were computed for the uncorrected (Naive), normalized maximum likelihood (NMLE), genome-wide bootstrap estimates without correction (GW Boot w/o correction) and genome-wide bootstrap with correction (GW Boot w/ correction) for true positive SNPs. Pr(p<α) is the probability of obtaining a p-value below threshold α for each SNP. Detailed descriptions of the methods are in Section 2.4.1.
### Table 2.3 Genome-wide simulation: performance of bias reduced 95% confidence intervals for low power alternative case

<table>
<thead>
<tr>
<th>Pr(p&lt;α)</th>
<th>MAF</th>
<th>OR</th>
<th>Coverage</th>
<th>GW Boot</th>
<th>Coverage</th>
<th>GW Boot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Naïve</td>
<td>NMLE</td>
<td>w/ corr</td>
<td>Naïve</td>
</tr>
<tr>
<td>7%</td>
<td>0.33</td>
<td>1.16</td>
<td>0.48</td>
<td>0.94</td>
<td>0.99</td>
<td>0.17</td>
</tr>
<tr>
<td>11%</td>
<td>0.45</td>
<td>1.16</td>
<td>0.60</td>
<td>0.96</td>
<td>0.99</td>
<td>0.16</td>
</tr>
<tr>
<td>13%</td>
<td>0.46</td>
<td>1.16</td>
<td>0.62</td>
<td>0.95</td>
<td>0.99</td>
<td>0.16</td>
</tr>
<tr>
<td>30%</td>
<td>0.29</td>
<td>1.20</td>
<td>0.79</td>
<td>0.93</td>
<td>0.99</td>
<td>0.18</td>
</tr>
<tr>
<td>49%</td>
<td>0.13</td>
<td>1.34</td>
<td>0.90</td>
<td>0.92</td>
<td>0.94</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**Note:** Datasets were simulated with study design parameters described in Table II for the low power alternative case. Coverage of 95% confidence interval and width (average width of confidence interval on log OR scale) were computed for the uncorrected (Naïve), normalized maximum likelihood (NMLE) and genome-wide bootstrap with correction (GW Boot w/ corr) estimates for true positive SNPs. Pr(p<α) is the probability of obtaining a p-value below threshold α for each SNP. Detailed descriptions of the methods are in Section 2.41..
<table>
<thead>
<tr>
<th>SNP</th>
<th>MAF</th>
<th>OR</th>
<th>log OR</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12708716</td>
<td>0.33</td>
<td>1.16</td>
<td>0.15</td>
<td>7%</td>
</tr>
<tr>
<td>rs11171739</td>
<td>0.45</td>
<td>1.16</td>
<td>0.15</td>
<td>11%</td>
</tr>
<tr>
<td>rs17696736</td>
<td>0.46</td>
<td>1.16</td>
<td>0.15</td>
<td>13%</td>
</tr>
<tr>
<td>rs9272346</td>
<td>0.29</td>
<td>1.20</td>
<td>0.18</td>
<td>30%</td>
</tr>
<tr>
<td>rs6679677</td>
<td>0.13</td>
<td>1.34</td>
<td>0.29</td>
<td>49%</td>
</tr>
</tbody>
</table>

**Table 2.4** Genome-wide simulation: Bias and reduction provided by rank for Naive, NMLE and GW Bootstrap with correction

<table>
<thead>
<tr>
<th>SNP</th>
<th>MAF</th>
<th>OR</th>
<th>log OR</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12708716</td>
<td>0.33</td>
<td>1.16</td>
<td>0.15</td>
<td>7%</td>
</tr>
<tr>
<td>rs11171739</td>
<td>0.45</td>
<td>1.16</td>
<td>0.15</td>
<td>11%</td>
</tr>
<tr>
<td>rs17696736</td>
<td>0.46</td>
<td>1.16</td>
<td>0.15</td>
<td>13%</td>
</tr>
<tr>
<td>rs9272346</td>
<td>0.29</td>
<td>1.20</td>
<td>0.18</td>
<td>30%</td>
</tr>
<tr>
<td>rs6679677</td>
<td>0.13</td>
<td>1.34</td>
<td>0.29</td>
<td>49%</td>
</tr>
</tbody>
</table>

Reduction Provided by Rank

<table>
<thead>
<tr>
<th>SNP</th>
<th>MAF</th>
<th>OR</th>
<th>log OR</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12708716</td>
<td>0.33</td>
<td>1.16</td>
<td>0.15</td>
<td>7%</td>
</tr>
<tr>
<td>rs11171739</td>
<td>0.45</td>
<td>1.16</td>
<td>0.15</td>
<td>11%</td>
</tr>
<tr>
<td>rs17696736</td>
<td>0.46</td>
<td>1.16</td>
<td>0.15</td>
<td>13%</td>
</tr>
<tr>
<td>rs9272346</td>
<td>0.29</td>
<td>1.20</td>
<td>0.18</td>
<td>30%</td>
</tr>
<tr>
<td>rs6679677</td>
<td>0.13</td>
<td>1.34</td>
<td>0.29</td>
<td>49%</td>
</tr>
</tbody>
</table>

Reduction Provided by GW Boot w/ corr by Rank

<table>
<thead>
<tr>
<th>SNP</th>
<th>MAF</th>
<th>OR</th>
<th>log OR</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12708716</td>
<td>0.33</td>
<td>1.16</td>
<td>0.15</td>
<td>7%</td>
</tr>
<tr>
<td>rs11171739</td>
<td>0.45</td>
<td>1.16</td>
<td>0.15</td>
<td>11%</td>
</tr>
<tr>
<td>rs17696736</td>
<td>0.46</td>
<td>1.16</td>
<td>0.15</td>
<td>13%</td>
</tr>
<tr>
<td>rs9272346</td>
<td>0.29</td>
<td>1.20</td>
<td>0.18</td>
<td>30%</td>
</tr>
<tr>
<td>rs6679677</td>
<td>0.13</td>
<td>1.34</td>
<td>0.29</td>
<td>49%</td>
</tr>
</tbody>
</table>

Note: Datasets were simulated with study design parameters described in Table II for the low power alternative case, with at least 50 datasets for each SNP at each rank. Reduction provided (average difference between corrected and uncorrected estimate) was computed for the genome-wide bootstrap with correction (GW Boot w/ corr) and Normalized Maximum Likelihood (NMLE). Absolute bias (average difference between the estimate of log OR and the true log OR) and RMSE (root mean square error) was computed for the uncorrected (Naive), NMLE and GW bootstrap with correction methods.
### Table 2.5 Genome-wide simulation: performance of bias reduced point estimators for log OR

<table>
<thead>
<tr>
<th>Alternative Case - High Power</th>
<th>Relative Bias (log OR)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAF OR Pr(p&lt;α)</td>
<td>Naïve</td>
</tr>
<tr>
<td>rs6679677</td>
<td>0.13 1.35 60%</td>
<td>0.21  -0.09  -0.62  -0.18</td>
</tr>
<tr>
<td>rs9272346</td>
<td>0.29 1.25 70%</td>
<td>0.23  -0.05  -0.50  -0.12</td>
</tr>
<tr>
<td>rs12708716</td>
<td>0.33 1.29 77%</td>
<td>0.08  -0.12  -0.48  -0.21</td>
</tr>
<tr>
<td>rs11171739</td>
<td>0.45 1.29 90%</td>
<td>0.06  -0.08  -0.26  -0.20</td>
</tr>
<tr>
<td>rs17696736</td>
<td>0.46 1.35 99%</td>
<td>0.04  0.01  -0.53  -0.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Null Case (OR = 1)</th>
<th>Absolute Bias (log OR)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF</td>
<td>Naïve</td>
<td>NMLE</td>
</tr>
<tr>
<td>0.05-0.1</td>
<td>0.41 0.21</td>
<td>0.03</td>
</tr>
<tr>
<td>0.1-0.2</td>
<td>0.30 0.16</td>
<td>0.02</td>
</tr>
<tr>
<td>0.2-0.3</td>
<td>0.25 0.14</td>
<td>0.03</td>
</tr>
<tr>
<td>0.3-0.4</td>
<td>0.23 0.14</td>
<td>0.04</td>
</tr>
<tr>
<td>0.4-0.5</td>
<td>0.23 0.13</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Note: 500 genome-wide case control datasets where the SNP of interest (in null case, SNP with MAF of interest) was significant at p-value threshold $\alpha = 5E-7$ were simulated using the WTCCC T1D genotypes and OR as indicated. Relative bias (average difference between estimated and true log OR divided by true log OR) and root mean squared error (RMSE) were computed for the uncorrected (Naïve), Normalized Maximum Likelihood (NMLE), genome-wide bootstrap without correction (GW Boot w/o corr) and genome-wide bootstrap with correction (GW Boot w/ corr) estimates for true positive SNPs. Absolute bias (average difference between estimated and true log OR) and RMSE was calculated for false positive SNPs. $Pr(p<\alpha)$ is the probability of obtaining a p-value below threshold $\alpha$ for each SNP. Detailed descriptions of the method are in Section 2.4.1.
Table 2.6 Genome-wide simulation: Proportion of estimates within 10% or 25% of true effect size

<table>
<thead>
<tr>
<th>Alternative Case (Low Power)</th>
<th>Proportion of estimates within 10% of true logOR</th>
<th>Proportion of estimates within 25% of true logOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP</td>
<td>MAF</td>
<td>log OR</td>
</tr>
<tr>
<td>rs12708716</td>
<td>0.33</td>
<td>0.15</td>
</tr>
<tr>
<td>rs11171739</td>
<td>0.45</td>
<td>0.15</td>
</tr>
<tr>
<td>rs17696736</td>
<td>0.46</td>
<td>0.15</td>
</tr>
<tr>
<td>rs9272346</td>
<td>0.29</td>
<td>0.18</td>
</tr>
<tr>
<td>rs6679677</td>
<td>0.13</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Null Case

<table>
<thead>
<tr>
<th>Proportion of estimates between 0 and log(1.1)</th>
<th>Proportion of estimates between 0 and log(1.2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAF</td>
<td>Naïve</td>
</tr>
<tr>
<td>0.05_to_0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.1_to_0.2</td>
<td>0</td>
</tr>
<tr>
<td>0.2_to_0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.3_to_0.4</td>
<td>0</td>
</tr>
<tr>
<td>0.4_to_0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: The proportion of estimates within 10% (25%) of the true log odds ratio is the proportion of uncorrected (Naïve) or corrected (AML, NMLE, SS Bootstrap) estimates that fall between 0.9β and 1.1β (0.75β and 1.25β) where β is the true log odds ratio for the SNP. Simulation study and estimators are described in Table 2.2.
### Table 2.7 Genome-wide simulation: Sample size estimated using bias-reduced methods for replication study with 80% power.

<table>
<thead>
<tr>
<th>SNP</th>
<th>Actual sample size required for 80% power</th>
<th>Naïve</th>
<th>NMLE</th>
<th>GW Boot w/ corr</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12708716</td>
<td>3096</td>
<td>1169</td>
<td>8283</td>
<td>4005</td>
</tr>
<tr>
<td>rs11171739</td>
<td>2944</td>
<td>1246</td>
<td>8485</td>
<td>4332</td>
</tr>
<tr>
<td>rs17696736</td>
<td>2895</td>
<td>1198</td>
<td>6975</td>
<td>3937</td>
</tr>
<tr>
<td>rs9272346</td>
<td>2177</td>
<td>1141</td>
<td>6817</td>
<td>3814</td>
</tr>
<tr>
<td>rs6679677</td>
<td>1916</td>
<td>1320</td>
<td>5982</td>
<td>4652</td>
</tr>
</tbody>
</table>

Note: Actual sample size is sample size required so that the replication study achieves 80% power for the particular SNP. Mean estimated sample size for 80% power is average replication study sample computed from uncorrected (Naïve), normalized maximum likelihood (NMLE) and genome-wide bootstrap with correction (GW Boot w/ corr) bias-reduced genetic effect estimates. Simulation study and estimators are described in Table 2.2.
Table 2.8 Genome-wide simulation: Sample size estimated for false positive SNPs using bias-reduced methods for replication study with 80% power

<table>
<thead>
<tr>
<th>MAF</th>
<th>Actual sample size required for CI narrow enough to exclude OR = 1.15</th>
<th>Mean estimated sample size for 80% power</th>
<th>Proportion of SNPs for which sample size is adequate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Naïve</td>
<td>NMLE</td>
</tr>
<tr>
<td>0.05 - 0.1</td>
<td>16165</td>
<td>2027</td>
<td>18187</td>
</tr>
<tr>
<td>0.1 - 0.2</td>
<td>3145</td>
<td>1545</td>
<td>17034</td>
</tr>
<tr>
<td>0.2 - 0.3</td>
<td>2098</td>
<td>1412</td>
<td>14530</td>
</tr>
<tr>
<td>0.3 - 0.4</td>
<td>1759</td>
<td>1355</td>
<td>11620</td>
</tr>
<tr>
<td>0.4 - 0.5</td>
<td>1605</td>
<td>1255</td>
<td>8378</td>
</tr>
</tbody>
</table>

Note: Actual sample size is sample size large enough so that the log OR CI width is twice log(1.15). Mean estimated sample size for 80% power is average replication study sample computed from uncorrected (Naïve), normalized maximum likelihood (NMLE) and genome-wide bootstrap with correction (GW Boot w/ corr) bias-reduced genetic effect estimates for false positive SNPs. Proportion of SNPs for which sample size is adequate is the proportion of SNPs for which the sample size computed from the naive or bias-reduced estimate is larger than the actual sample size required. Simulation study parameters are described in Table 2.2.
<table>
<thead>
<tr>
<th>Level 1 Bootstraps</th>
<th>Level 2 Bootstraps</th>
<th>Multi-CPU Option</th>
<th>Run Time</th>
<th>Point Estimate &amp; Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>Single CPU</td>
<td></td>
<td>26 min 32 hrs</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>16 threads</td>
<td></td>
<td>6.5 min 3.5 hrs</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>16 threads</td>
<td></td>
<td>15.5 min 16.5 hrs</td>
</tr>
</tbody>
</table>

Note: Genome-wide bootstrap point estimate and confidence intervals were obtained for a dataset simulated under low power alternative case with 2 significant SNPs using WTCCC genotypes (details in Table 2.2), threshold for significance was trend test p-value using br2 software [24] on moderately powered UNIX machine (2 x Quad Core Intel Xeon E5540 2.53GHz, 16 GB RAM) Run time is amount of time required using the indicated number of threads (with parallel processing option in br2 software).
Table 2.10 Single-SNP simulation results: performance of bias-reduced point estimators

<table>
<thead>
<tr>
<th>power</th>
<th>N</th>
<th>Naïve</th>
<th>AML</th>
<th>NMLE</th>
<th>SS</th>
<th>RMSE</th>
<th>Naïve</th>
<th>AML</th>
<th>NMLE</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>1650</td>
<td>0.63</td>
<td>-0.04</td>
<td>-0.16</td>
<td>0.28</td>
<td>0.17</td>
<td>0.17</td>
<td>0.13</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>10%</td>
<td>2050</td>
<td>0.47</td>
<td>-0.07</td>
<td>-0.18</td>
<td>0.17</td>
<td>0.13</td>
<td>0.16</td>
<td>0.12</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>30%</td>
<td>2950</td>
<td>0.26</td>
<td>-0.06</td>
<td>-0.19</td>
<td>0.04</td>
<td>0.08</td>
<td>0.13</td>
<td>0.11</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>3600</td>
<td>0.17</td>
<td>-0.05</td>
<td>-0.17</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
<td>0.10</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>90%</td>
<td>5720</td>
<td>0.03</td>
<td>0.00</td>
<td>-0.11</td>
<td>-0.05</td>
<td>0.03</td>
<td>0.04</td>
<td>0.08</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>99%</td>
<td>10000</td>
<td>0.01</td>
<td>0.00</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>

Note: The true odds ratio of the SNP of interest is 1.3 ($\log(1.3)=0.26$) with MAF of 0.25, and the number of cases equals the number of controls (=N/2). 1,000 datasets significant at the p-value threshold of 1E-6 were simulated for each combination of parameters. The relative bias is the average difference between estimated and true log OR divided by true log OR (i.e. 1.3) averaged over 1,000 replicates. RMSE is the square root of the mean squared error. Naïve is the uncorrected estimate; AML, NMLE, and SS Bootstrap are bias-reduced estimates. Detailed descriptions of the methods are in Appendix 3.
Table 2.11 Single-SNP simulation results: Proportion of estimates within 10% or 25% of true log odds ratio

<table>
<thead>
<tr>
<th>Alternative Case</th>
<th>Proportion of estimates between 0.262 +/- 0.0262</th>
<th>Proportion of estimates between 0.262 +/- 0.0655</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naïve</td>
<td>AML</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>0</td>
<td>0.09</td>
</tr>
<tr>
<td>10%</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>30%</td>
<td>0.01</td>
<td>0.13</td>
</tr>
<tr>
<td>50%</td>
<td>0.34</td>
<td>0.21</td>
</tr>
<tr>
<td>99%</td>
<td>0.58</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Null Case

<table>
<thead>
<tr>
<th></th>
<th>Proportion of estimates between 0 and 0.095</th>
<th>Proportion of estimates between 0 and 0.182</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Naïve</td>
<td>AML</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Note: The proportion of naive and bias reduced estimates of log OR for true positive SNPs that fall within 10% (left) or 25% (right) of the true log odds ratio (0.262) and the proportion of estimates for false positive SNPs that fall between 0 and 0.095 (left, corresponds to OR of 1 and 1.1) and 0 and 0.182 (right, corresponds to OR of 1 and 1.2) were computed for datasets simulated as described in Table 2.2.
2.9 Figures

Figure 2.1 Uncorrected, normalized maximum likelihood and genome-wide bootstrap with correction bias-reduced estimates for a sample of SNPs meeting the WTCCC moderate association criterion

Uncorrected (Naive), normalized maximum likelihood (NMLE) and genome-wide bootstrap with correction (GW Boot w/correction) bias-reduced estimates for a sample of SNPs meeting the WTCCC moderate association criterion (minimum p-value of the trend test (1 df) and genotypic test (2 df) less than 1E-5). NMLE and GW Boot estimates computed from WTCCC discovery samples using threshold of 1E-5. Vertical lines connect the WTCCC Naive estimate with corresponding bias-reduced estimates.
Figure 2.2 Genome-wide simulation results for effect size estimation in null and low power alternative case. Histograms of genetic effect estimates using the naïve, normalized maximum likelihood (NMLE) and genome-wide bootstrap without correction (GW Boot w/o corr) and with correction (GW Boot w/ corr) from 500 simulated case-control datasets. Open triangle is the mean. The far left plot is the null case for all false positive SNPs with MAF 25-50%. From left to right starting from the second plot, the probability of selection for each SNP (simulated under the alternative) is 13%, 30% and 50%. Dashed lines connect the 10th, 50th and 90th percentiles of the naïve estimates to their corresponding bias-reduced estimates. Simulation study parameters indicated in Table 2.2.
Figure 2.3 Genome–wide simulation results for replication sample size estimation in low power alternative case. Histograms of estimated sample size required for replication study with 80% power using the uncorrected (Naïve), normalized maximum likelihood (NMLE) and genome–wide bootstrap with correction (GW Boot w/ corr) from 500 simulated case–control datasets. Open triangle is the mean. The probability of selection for each simulated SNP is 13%, 30% and 50% from left to right. Simulation study parameters indicated in Table 2.2.
**Figure 2.4 Genome-wide simulation with threshold-based selection (alternative case – high power).** Histograms of estimates of genetic effect using the uncorrected (Naïve), normalized maximum likelihood (NMLE), genome-wide bootstrap without correction (GW Boot w/o corr) and with correction (GW Boot w/ corr) methods for 500 simulated significant case-control datasets. Open triangle is the mean. The probability of selection for each simulated SNP is 60%, 70%, 80%, 90%, 99% from top to bottom. Simulation study parameters are the same as for Table 2.5.
Figure 2.5 Genome-wide simulation results (alternative case - low power) Histogram of probability of selection in replication study when original genetic estimate is used to compute sample size required for replication with 80% power. Sample size computed using the uncorrected (Naive), normalized maximum likelihood (NMLE) and genome-wide bootstrap with correction (GW Boot w/correction) method for 500 simulated cases-control datasets. Open triangle is the mean. The probability of selection in the original study for each simulated SNP is 13%, 30% and 50% from left to right. Simulation study parameters are the same as for Table 2.2.
Figure 2.6 Single-SNP simulation results. Histograms of uncorrected (Naïve), adjusted median likelihood (AML), normalized maximum likelihood (NMLE) and single-SNP bootstrap (SS Boot) bias reduced estimates from 500 simulated significant case-control datasets. The power for each simulation is null case, 5%, 10%, (top row) 30%, 50%, 99% (bottom row) from left to right. Simulation study parameters are the same as for Table 2.9. Open triangle is the mean, dashed lines connect the deciles, minimum and maximum of the naïve estimates to their corresponding NMLE and SS bootstrap bias-reduced estimates.
Figure 2.7 Comparison of normalized maximum likelihood and single–SNP bootstrap estimates. Single–SNP simulation comparison of normalized maximum likelihood (NMLE) and single–SNP bootstrap (SS Boot) estimates of log odds ratio from 1000 simulated significant case–control datasets (left) and comparison of relationship of NMLE and SS bootstrap with trend test statistic used for threshold–based selection. Vertical and horizontal lines are the true log odds ratio. Power of simulation is 30%, odds ratio is 1.3 and minor allele frequency is 0.25, sample size is 2950 and the criterion for significance is trend test p–value less than 1E−6. The likelihood estimates generally had higher variance than either of the bootstrap methods. When the test statistic is large, the relationship between the likelihood estimate and test statistic is approximately linear, however there is an inflection–point below which estimates are drastically reduced. The consequence is that small changes in the data can result in large changes in the likelihood estimate when the test statistic is close to this threshold. For instance, in the single–SNP simulation where power was 0.05 (Figure 2.6), naïve log odds ratio estimates ranging from 0.43 to 0.44 corresponded to a moderate range of bootstrap estimates (0.29 to 0.38) but to a much larger range of likelihood estimates (0.14 to 0.32).
Figure 2.8 Relationship between Likelihood, power and conditional likelihood. Likelihood, power and conditional likelihood (likelihood conditional on two-sided Wald test significant at p-value threshold of $1E^{-7}$) for the mean parameter of the normal distribution where standard deviation is 1. Observed data for the likelihood is $Z_{obs}=5.5$. This corresponds to a test statistic for an estimated odds ratio of 1.3 and standard deviation for the estimate of 0.0475.
2.10 Appendices

2.10.1 Appendix 1 Derivation of the correlation correction for the genome-wide bootstrap estimator

The bootstrap is intended to repeatedly mimic a sampling experiment. In this case, the experiment is one in which a sample is drawn from an infinite population, and if the genetic effect is significant, it is estimated from this sample. Then a second independent sample is drawn from the population, and the true underlying genetic effect is estimated again.

In the absence of selection, the distribution of the first sample estimate $\hat{\beta}_1$ will center on the true underlying genetic effect parameter, $\beta$. Under selection, the distribution of the first sample estimate will center on a value larger than the genetic effect parameter $\beta$. The second sample estimate $\hat{\beta}_2$ will center on $\beta$ and is unaffected by selection in the first sample. The mean difference between the two estimates over a large number of replicates is the selection bias. In our hypothetical experiment, in the absence of selection in the first sample, where $\sigma_1^2$ and $\sigma_2^2$ are the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$, the joint distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$ is

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim N\left( \begin{pmatrix} \beta \\ \beta \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right)$$

However, with selection in the first sample $E(\hat{\beta}_1 | \hat{\beta}_1 / \hat{\sigma}_1 > \text{crit}) = \gamma$ where $\gamma > \beta$ and $E(\hat{\beta}_2 | \hat{\beta}_1 / \hat{\sigma}_1 > \text{crit}) = \beta$. The bias is $\gamma - \beta = E(\hat{\beta}_1 - \hat{\beta}_2 | \hat{\beta}_1 / \hat{\sigma}_1 > \text{crit})$. 
The bootstrap procedure treats the original finite sample as the population in which \( \hat{\beta}_N \) corresponds to the population parameter \( \beta \). The within-bootstrap sample mimics the first sample and \( \hat{\beta}_D \) corresponds to the parameter estimate \( \hat{\beta}_1 \). The out-of-bootstrap sample mimics the second independent sample and \( \hat{\beta}_E \) corresponds to the parameter estimate \( \hat{\beta}_2 \). In the absence of selection, the distribution of \( \hat{\beta}_D \) will center on \( \hat{\beta}_N \), whereas with selection the distribution of \( \hat{\beta}_D \) will center on a value greater than \( \hat{\beta}_N \). In this way, the within sample from the bootstrap scheme mimics the first sample in our experiment.

In the absence of selection in the first sample, the distribution of \( \hat{\beta}_E \) will also center on \( \hat{\beta}_N \), as intended. But the within and out-of samples are mutually exclusive and drawn from a finite sample, so when by chance \( \hat{\beta}_D > \hat{\beta}_N \), \( \hat{\beta}_E \) will be < \( \hat{\beta}_N \) and vice versa, leading to negative correlation. Therefore even in the absence of selection in the within sample, the joint distribution of \( \hat{\beta}_D \) and \( \hat{\beta}_E \) is

\[
\begin{pmatrix}
\hat{\beta}_D \\
\hat{\beta}_E
\end{pmatrix}
| \hat{\beta}_N = B 
\sim N \left( B, \begin{pmatrix}
\sigma_D^2 & \sigma_{DE} \\
\sigma_{DE} & \sigma_E^2
\end{pmatrix} \right),
\] where \( \sigma_{DE} \) is negative.

Selection for large \( \hat{\beta}_D \) selects for small \( \hat{\beta}_E \) and therefore with selection in the first sample:

\[
E(\hat{\beta}_D \mid \hat{\beta}_D > \text{crit}) = \gamma_D \ \text{where} \ \gamma_D > \hat{\beta}_N = B
\]

\[
E(\hat{\beta}_E \mid \hat{\beta}_D > \text{crit}) = \gamma_E \ \text{where} \ \gamma_E < \hat{\beta}_N = B.
\]

The expectation of \( \hat{\beta}_E \) conditional on \( \hat{\beta}_D = d \) is
\[
E(\hat{\beta}_E | \hat{\beta}_D = d, \hat{\beta}_N = B) = B + \frac{\sigma_{DE}}{\sigma_D^2} (d - B)
\]

In order to better mimic an independent estimate, we need to remove the correlation between \(\hat{\beta}_D\) and \(\hat{\beta}_E\). If we define \(\hat{\beta}_E^* = \hat{\beta}_E - \frac{\sigma_{DE}}{\sigma_D^2} (d - B)\) then

\[
\text{cov}(\hat{\beta}_D, \hat{\beta}_E^*) = \text{cov}(\hat{\beta}_D, \hat{\beta}_E - \frac{\sigma_{DE}}{\sigma_D^2} (\hat{\beta}_D - B))
\]

\[
= \text{cov}(\hat{\beta}_D, \hat{\beta}_E - \frac{\sigma_{DE}}{\sigma_D^2} \hat{\beta}_D)
\]

\[
= \text{cov}(\hat{\beta}_D, \hat{\beta}_E) - \frac{\sigma_{DE}}{\sigma_D^2} \text{var}(\hat{\beta}_D)
\]

\[
= \sigma_{DE} - \frac{\sigma_{DE}}{\sigma_D^2} \sigma_D^2
\]

\[
= 0
\]

\[
E(\hat{\beta}_E^*) = E\left( \hat{\beta}_E - \frac{\sigma_{DE}}{\sigma_D^2} (\hat{\beta}_D - B) \right) = B
\]

Letting \(\rho_{DE}\) be the correlation between \(\hat{\beta}_D\) and \(\hat{\beta}_E\), we have
\[
\text{var}(\hat{\beta}_E^*) = \text{var}\left(\hat{\beta}_E - \frac{\sigma_{DE}}{\sigma_D^2}(\hat{\beta}_D - B)\right)
\]
\[
= \text{var}\left(\hat{\beta}_E - \frac{\sigma_{DE}}{\sigma_D^2}\hat{\beta}_D\right)
\]
\[
= \text{var}(\hat{\beta}_E) + \left(\frac{\sigma_{DE}}{\sigma_D^2}\right)^2 \var(\hat{\beta}_D) - 2 \frac{\sigma_{DE}}{\sigma_D^2} \text{cov}(\hat{\beta}_E, \hat{\beta}_D)
\]
\[
= \sigma_E^2 + \left(\frac{\sigma_{DE}}{\sigma_D^2}\right)^2 \sigma_D^2 - 2 \frac{\sigma_{DE}}{\sigma_D^2} \sigma_{DE}
\]
\[
= \sigma_E^2 - \frac{(\sigma_{DE})^2}{\sigma_D^2}
\]
\[
= \sigma_E^2 (1 - \rho_{DE}^2)
\]

In practice, \( \rho_{DE} \) is around 0.7, and so \( \text{var}(\hat{\beta}_E^*) = 0.5\sigma_E^2 \).

This gives us a joint distribution that mimics the estimates in our hypothetical experiment.

\[
\begin{pmatrix}
\hat{\beta}_D \\
\hat{\beta}_E^* \end{pmatrix} \sim N\left(\begin{pmatrix} B \\ B \end{pmatrix}, \begin{pmatrix} \sigma_D^2 & 0 \\
0 & 0.5\sigma_E^2 \end{pmatrix}\right).
\]

2.10.2 Appendix 2 Derivation of the MAF correction for the genome-wide bootstrap estimator

In the class of regression models with response formulated as a function of a linear predictor, \( X^T\beta \), the design matrix \( X \) contributes to the variance of the regression coefficients, \( \beta \), through the quantity \( (X^T X) \). For example, in linear regression: \( E(Y) = X^T \beta \) and in logistic regression \( \logit\{\Pr(Y=1)|X\} = X^T \beta \). In general, \( \hat{\beta} \) and \( \text{Var}(\hat{\beta}) \) will be inversely related to \( \text{Var}[X] \). For example in the QTL linear regression model, \( Y = \mu + X\beta + e \) where \( Y \) is the vector of
quantitative trait values, $\mu$ is the intercept term, $X$ is the vector of centered genotypes for the associated SNP and $e \sim N(0, \sigma_e^2 I)$ is the vector of error terms:

$$\hat{\beta} = (XX)^{-1}X'Y$$
$$= \beta + (XX)^{-1}X'e$$
$$Var(\hat{\beta}) = (XX)^{-1} \sigma_e^2 / N$$
$$\approx \frac{\sigma_e^2 / N}{\text{var}(X)}$$

In the case of a single additively-coded SNP covariate, $x$ is the number of copies of the ‘risk’ allele, where $p$ is the frequency of the ‘risk’ allele, assuming Hardy Weinberg Equilibrium, we have $E[x = 2] = p^2, E[x = 1] = 2p(1 - p), E[x = 0] = (1 - p)^2$

It follows that:

$$E[x] = 2p^2 + 2p(1 - p) + 0(1 - p)^2 = 2p$$

and

$$Var[x] = p^2 (2 - x)^2 + 2p(1 - p)(1 - x)^2 + (1 - p)^2 x^2 = 2p(1 - p)$$

Therefore, in order to estimate the bias across many SNPs we rescale by $\sqrt{2p(1 - p)}$. 
2.10.3 Appendix 3  Single-SNP bootstrap shrinkage estimation

2.10.3.1 Single SNP Bootstrap Estimator

The single-SNP bootstrap shrinkage estimator, which examines data of a single-SNP alone, is a special case of the genome-wide estimator for which ranking adjustment does not apply (i.e. no index \((k)\) nor MAF correction). In addition, the effect of selection on \(\hat{\beta}_{Ei}\) due to correlation between \(\hat{\beta}_{Di}\) and \(\hat{\beta}_{Ei}\) is not large enough to require correction, because in the single-SNP case, the SNP selected in the bootstrap sample is always the SNP selected in the original sample. (As explained in Section 2.2) Therefore, the genome-wide estimator in equation (2.2) reduces to:

\[
\hat{\beta}_{boot} = \hat{\beta}_N - \frac{1}{N} \sum_{i=1}^{N} \left( \hat{\beta}_{Di} - \hat{\beta}_{Ei} \right)
\]

\((C.1)\)

2.10.3.2 Single SNP Simulation Studies

We evaluated the performance of the single-SNP methods (AML, NMLE and SS Bootstrap) when only threshold selection bias is present, considering accuracy of point estimation and CI coverage. For subsequent genome-wide simulations we focused on the NMLE, a choice based on smaller root mean squared error (RMSE) in comparison to AML. The four estimators investigated here are as follows.

\textit{Uncorrected}

\textbf{Naïve}: original naïve estimator
Bias-corrected, single-SNP methods

**AML**: Adjusted Median Likelihood estimator of Zhong and Prentice [16,21]

**NMLE**: Normalized Maximum Likelihood Estimator of Ghosh *and others* [17]

**SS Bootstrap**: Single SNP Bootstrap shrinkage estimator, equation (C.1) above

### 2.10.3.3 Single-SNP simulation study: design

We simulated case-control datasets with a single causal SNP under an additive genetic model with a range of odds ratio, sample size and power. For each dataset, we tested for association between the SNP genotype and case-control phenotype using the trend test. Given statistical significance of the association test at a prespecified level, we first obtained the naïve estimate of the log OR via logistic regression. We then computed the bias-reduced AML, NMLE and SS Bootstrap point estimates (using 100 bootstrap samples). For each simulation scenario we calculated summary statistics (bias, variance, RMSE, proportion of estimates in close proximity to the true value) over 1,000 significant replicates. We also plotted the empirical distribution of the different estimates to capture other features.

We chose study design parameters (Table 2.10) relevant to a genome-wide association study for a range of power levels: 5%, 10%, 30%, 50%. We also considered 99% power to assess the asymptotic behaviour of the methods. Because power is the main factor influencing the bias, and reflects the combination of parameters, we fixed the effect size and varied the sample size.
2.10.3.4 Single-SNP simulation study: results

Results were similar for simulations with ORs of 1.1, 1.3, and 1.5, so we focus on results for an OR of 1.3 (Figure 2.6 and Table 2.10). The sample sizes (corresponding to the power levels considered) are 1650 (5%), 2050 (10%), 2950 (30%), 3600 (50%), 5720 (90%), 10,000 (99%). At low power, the distribution of the naïve estimate is approximately a truncated normal, which is highly biased with a small variance (Figure 2.6). As power increases, the bias in the naïve estimate decreases and the variance increases slightly. In contrast, the AML, NMLE and SS bootstrap estimates are less biased with a somewhat larger variance.

At low power, as a result of the long right tail, the SS bootstrap estimate tends, on average, to under-correct for the bias (Figure 2.6 and Table 2.10). At 50% power, the bootstrap estimator is close to being unbiased, at high power it over-corrects slightly, but as power approaches 1, the estimate becomes unbiased. The mass of the bootstrap estimates forms a mode close to the true value giving better accuracy than the naïve estimate, and the majority of the bootstrap estimates are within 25% of the true log OR (Table 2.11). As power increases, MSE decreases as the bootstrap estimates form a tighter local mode around the true value (Table 2.10).

In contrast, the distribution of the NMLE estimates is wide and flat with a pool of estimates at low values, yielding a conditional MLE with downward bias (Figure 2.6 and Table 2.10). The conditional MLE agrees generally with the bootstrap estimates when the observed test statistic is large. However, as the observed test statistic approaches the critical value from above, the conditional MLE drops rapidly toward the null (Figures 2.6 and 2.7). Compared to the bootstrap, fewer conditional MLE estimates are within 25% of the true value of the log OR (Table 2.11). In general, a consequence of conditioning is a non-quadratic likelihood that tends to be quite flat at
parameter values close to the null when power is low (Figure 2.8). Since there is no sharp peak at
low power, the maximum of the conditional likelihood can be sensitive to small changes in the
data, resulting in an MLE with high variance. The AML estimates are less biased but have higher
RMSE than the NMLE estimates (for power <90%), which is due to the higher variance of the
AML estimates.
2.11 References


www.ashg.org/genetics/abstracts/abs04/f227.htm


9. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678


22. Faye L, Sun L, Dimitromanolakis A and Bull SB (2009) A comprehensive look at the likelihood and bootstrap approaches to overcome the winner’s curse in GWAS. (Abstract 126) Genetic Epidemiology 33:788-789


Chapter 3. Two-stage study designs combining genome-wide association studies, tag SNPs, and sequencing: accuracy of genetic effect estimates

3.1 Abstract

Genome-wide association studies (GWAS) test for disease-trait associations and estimate effect sizes at tag single-nucleotide polymorphisms (SNPs), which imperfectly capture variation at causal SNPs. Sequencing studies can examine potential causal SNPs directly; however, sequencing the whole genome or exome can be prohibitively expensive. Costs can be limited by using a GWAS to detect the associated region(s) at tag SNPs followed by targeted sequencing to identify and estimate the effect size of the causal variant. Genetic effect estimates obtained from association studies can be too large because of a form of selection bias known as the winner’s curse. Conversely, estimates at tag SNPs can be attenuated compared to the causal SNP because of incomplete linkage disequilibrium. These two effects oppose each other. Analysis of rare SNPs further complicates our understanding of the winner’s curse because rare SNPs are difficult to tag and analysis can involve collapsing over multiple rare variants. In two-stage analysis of Genetic Analysis Workshop 17 simulated data sets, we find that selection at the tag SNP produces upward bias in the estimate of effect at the causal SNP, even when the tag and causal SNPs are not well correlated. The bias similarly carries through to effect estimates for rare variant summary measures. Replication studies designed with sample sizes computed using biased estimates will be under-powered to detect a disease-causing variant. Accounting for bias
in the original study is critical to avoid discarding disease-associated SNPs at follow up.

### 3.2 Background

Selection bias in genetic association studies arises when the same sample is used for both gene discovery and effect estimation. Under the low power that is common in a genome-wide association study (GWAS), selection causes upward bias in the magnitude of genetic effect estimates because the effect size is estimated only when the test statistic exceeds the threshold for significance. This phenomenon is also known as the winner’s curse, and its effect on linkage analyses and on case-control association was demonstrated by Goring et al. [1] and Garner [2], respectively. In a two-stage design, the correlation between the tag single-nucleotide polymorphism (SNP) used for selection and one or more causal variants adds complexities to the understanding of the winner’s curse.

Linkage disequilibrium causes the test statistics and effect size estimates at the tag and causal SNPs to be correlated. Linkage disequilibrium can be quantified by $r^2$ or other measures, but in this case we focus on the Pearson correlation coefficient $r$ as an estimate of correlation $\rho$. Low correlation between the tag SNP and causal SNPs decreases the power to detect the effect at the tag SNP, which induces upward selection bias in both the tag and causal SNP estimates. On the other hand, as correlation decreases, the tag effect attenuates and selection at the tag SNP exerts less influence on the estimate at the causal SNP [1, 2, 3] The balance between these two trends determines the degree of bias in the estimates.
Sequencing studies can uncover rare SNPs, for which conventional tests, such as single-SNP linear or logistic regression, are not powerful. Recently, methods to test for and estimate the effect size of multiple rare variants have been proposed. Several investigators have proposed a region-specific regression method that collapses genotypes at all rare variants in a specified region [4, 5]. Within the region, \( n_i \) is the number of rare SNPs genotyped for individual \( i \), \( r_i \) is the number of rare SNPs at which individual \( i \) has a rare allele, and the independent variable is \( r_i/n_i \). Morris and Zeggini define the regression parameter \( \lambda \) as the genetic effect size for an individual carrying the rare allele at each of these \( n_i \) rare SNPs [4].

In a two-stage design the relationship between the tag SNP and a rare causal SNP is complicated by the minor allele frequency (MAF) and the contribution of multiple causal SNPs to \( r_i/n_i \). The correlation coefficient \( r \) between two SNPs has an upper bound that depends on the difference between the MAFs. A tag SNP with MAF > 5% will capture little of the variation at a rare SNP with MAF < 0.1%, especially in a small sample. A tag SNP with a low MAF tends to do better at capturing variation at a rare causal SNP. When multiple rare SNPs contribute to the genetic score for an individual, a useful tag SNP would be correlated with multiple SNPs.

In this paper, we study the consequences of a two-stage design for estimating the genetic effect at both the GWAS and sequencing stages. Using the Genetic Analysis Workshop 17 (GAW17) unrelated mini-exome data set [6] and the corresponding tag SNP genotypes drawn from the publicly available HapMap data set [7], we estimate the magnitude of the winner’s curse and the attenuation resulting from incomplete correlation between the tag SNP and causal SNPs. We demonstrate that selection bias occurs in both stages, even when the tag SNP is poorly correlated with the rare or the common causal SNP(s).
3.3 Methods

We examine three different two-stage scenarios described below in which first a genetic effect is detected at a tag SNP in a GWAS, then the gene is sequenced to find the true causal SNP(s), and finally the genetic effect is estimated at the true causal SNP(s). We compare the distribution of genetic effect estimates over varying correlations between the tag SNP and the causal SNP and over varying MAFs at both SNPs. The examination of the design matrix for the additive model shows that the relevant quantity for comparing the tag SNP and causal SNP effects is

\[
\gamma_{TC} = \frac{[2p_c(1-p_c)]^{1/2}}{[2p_T(1-p_T)]^{1/2}} \rho_{TC},
\]

where \(p_c\) is the causal SNP MAF, \(p_T\) is the tag SNP MAF, and \(\rho_{TC}\) is the correlation between the tag SNP and causal SNPs. Without loss of generality, we use \(\rho_{TC}\) to refer to both the correlation parameter and the Pearson correlation coefficient in the finite sample.

For analysis, we classify the GAW17 data into the following subpopulations: CEPH (European-descent Utah residents), Chinese, Japanese, Tuscan, Luhya, Yoruba, Europeans (CEPH + Tuscan), Asians (Chinese + Japanese), and Africans (Luhya + Yoruba). We define a common SNP as one with MAF > 5% and a rare SNP as one with MAF < 5% within a subpopulation. We compute \(r_i/n_i\) over all rare SNPs in a gene. We use a linear model with covariates Age and Smoking for trait Q1 and no covariates for Q2, as specified in the simulation model. To avoid
sparse data when stratifying by ethnicity, we use a logistic model with no covariates for disease status.

3.3.1 Scenario 1: common tag SNP, common causal SNP

In the first scenario, the tag SNP detected in the stage 1 GWAS is common and the causal SNP found in the stage 2 sequencing study is also common. We construct a two-stage study combining the GAW17 data set with the HapMap data set as follows. For the stage 1 GWAS, we obtain genotypes at tag SNPs for the individuals in the GAW17 data set (matched by individual ID) from the HapMap data set (build 36, available at http://hapmap.ncbi.nlm.nih.gov) [7]. For the stage 2 sequencing study, we use the GAW17 mini-exome data set for unrelated individuals. We included all 616 subjects for which there was genotype data in both the HapMap build 36 and GAW17 data sets.

For each stage 1 tag SNP, we test for additive genetic effect with trait Q2 using a linear model and select all data sets that are significant at a $p$-value threshold of 0.05. For each data set in which the tag SNP is significant, we estimate the effect at the tag SNP. In stage 2, we estimate the additive genetic effect size for the common causal SNP or rare SNP collapsing statistic $r_i/n_i$ for all data sets in which the tag SNP is significant. We compare the estimates obtained from data sets in which the tag SNP is significant with the estimates obtained for all 200 data sets. PLINK was used for all analyses [8, 9].

We describe the distribution of estimates using a boxplot (Figure 3.1). We present results for quantitative trait Q2 with causal SNP C6S5380 and tag SNPs from the HapMap data set that fall
within the VNN1 gene, as defined by the gene information file provided with the GAW17 data set. In Figure 3.1, we present boxplots for the distribution of the genetic effect estimates over 200 replications for six HapMap tag SNPs that are correlated with the causal SNP and one tag SNP that is not correlated with the causal SNP. Because we observed large effect sizes for this uncorrelated tag SNP on both sides of the null value, we used a one-sided test ($\alpha = 0.025$) to ensure that the boxplot would reflect the magnitude of the effect. Using the causal SNP as its own tag, we also include the case in which the tag SNP is perfectly correlated with the causal SNP. In Figure 3.1 we present 4 boxplots for each tag SNP: (1) the causal SNP genetic effect estimates over all 200 replicates (light blue), (2) the tag SNP estimates over all 200 replicates (orange), (3) the tag SNP estimates for all replicates in which the tag SNP is significant (red), and (4) the causal SNP estimates for all replicates in which the tag SNP is significant (blue).

### 3.3.2 Scenario 2: common tag SNP, multiple rare causal SNPs

In the second scenario, the tag SNP is common and there are multiple rare causal SNPs in the same gene at the sequencing stage. We construct a two-stage study from the GAW17 data set using all 697 individuals: The common SNPs are used as tag SNPs and the rare SNPs are used as sequencing SNPs. For each stage 1 tag SNP, we test for additive genetic effect with traits Q1, Q2, and disease status and select all data sets for which the tag SNP $p$-value is smaller than the significance threshold (the threshold was chosen so that estimated power was less than 20%) indicated in Table 3.1. We estimate the additive genetic effects of the parameter $\beta$ at the tag SNP and the parameter $\lambda$ for the $r/n_i$ rare SNP collapsing statistic.
3.3.3 Scenario 3: tag SNP with MAF 1–5%, multiple rare causal SNPs

In the third scenario, we consider the case in which the tag SNP has a low MAF and there are multiple rare causal SNPs in the same gene in the sequencing stage. We construct a two-stage study from the GAW17 data set using all 697 individuals: Each SNP in a causal gene with MAF between 1% and 5% is in turn used as the GWAS tag SNP; the rest are used as sequencing SNPs (i.e., the tag SNP is excluded from the collapsing statistic $r_i/n_i$). For each tag SNP, we test for additive genetic effect with traits Q1, Q2, and disease status. We select data sets and estimate effect sizes as in scenario 2.

3.4 Results

We use boxplots or summary measures to describe the effect of selection on the distribution of estimates at the tag SNP and causal SNPs. For scenario 1, we test at a fixed threshold for significance (Figure 3.1). For scenarios 2 and 3, we select significance thresholds required for low power. Power is estimated as the proportion of data sets for which the tag SNP test is significant. The mean effect estimate is computed over all 200 data sets and is also computed over all data sets for which the tag SNP additive test is significant. We compute the summary measure relative bias as the difference between the mean estimate over data sets for which the tag SNP is significant and the mean estimate over all 200 data sets, divided by the mean estimate over all 200 data sets (Table 3.1 and 3.2).

3.4.1 Scenario 1

The expected pattern of attenuation resulting from imperfect correlation is evident in the
distributions of the genetic effect estimates at the tag SNP and causal SNP. The tag SNP estimates (Figure 3.1, orange boxplots) increase with the $\gamma_{TC}$ quantity from equation 1. Selection for significant genetic effect at the tag SNP induces upward bias into the estimate. As $\gamma_{TC}$ increases, the power increases and the effect of selection decreases. The difference between the tag SNP estimates with and without selection is much smaller when $\gamma_{TC}$ is larger (Figure 3.1, orange and red box plots). When $\gamma_{TC}$ is 0, as for the null case, only the most extreme data sets are selected, and the bias away from the null value is large. When $\gamma_{TC}$ is 1, power is high and the bias is smaller.

Because of correlation between the tag SNP and the causal SNP, selection bias in the tag SNP estimate carries through to the causal SNP estimate. Two factors influence the bias at the causal SNP: (1) the correlation between the tag SNP and the causal SNP and (2) the power to detect the effect at the tag SNP. When $\gamma_{TC}$ is small, power is low and so the upward bias resulting from selection at the tag SNP is large. As $\gamma_{TC}$ increases, power increases and selection bias decreases. On the other hand, when $\gamma_{TC}$ is small, the correlation between the tag SNP and causal SNP estimates is small, and so the effect of selection at the tag SNP, although large, carries through to the causal SNP only to a small degree. As $\gamma_{TC}$ increases, the correlation increases and the effect of selection, although small, is carried through to a greater degree. These two effects tend to balance each other out so that the bias is similar over the small range of $\gamma_{TC}$ in the tag SNPs that we examined. The exceptions are at the extremes. If the tag SNP and the causal SNP are completely uncorrelated, then none of the selection at the tag SNP carries through to the causal SNP (Figure 3.1, boxplots at far left). If the tag and causal SNPs are perfectly correlated (in our case the tag SNP is in fact the causal SNP), then power to detect the SNP is so high that selection
bias is minimal (Figure 3.1, boxplots at far right).

3.4.2 Scenario 2

Rare causal SNPs can be difficult to tag with common SNPs because of the upper bound for correlation between two SNPs with different MAFs. In causal genes with multiple rare SNPs, we found 68 SNPs with MAF > 5%. Of these, there were only five SNPs that tagged the effect of the rare SNPs (correlation with \( r/ni \) was \( \rho > 0.4 \) for power to detect effect above significance level \( \alpha \) for at least one of \( \alpha = 0.001, 0.01, \) or \( 0.05 \)). The correlation was low in all five cases (Table 3.1).

The mean tag SNP genetic effect estimate was inflated by selection. Estimates of the rare SNP parameter were biased but to a lesser degree. When there is imperfect correlation between the tag SNP and causal SNPs, some but not all of the selection bias at the tag SNP is transferred to the causal SNP estimate.

3.4.3 Scenario 3

Rare causal SNPs can be more easily tagged with low-MAF SNPs, because two SNPs with similar MAFs have a higher upper bound for correlation. In causal genes with multiple rare SNPs, we found 162 SNPs with MAF between 1% and 5%. Of these, 12 tagged the effect of the rare SNPs (as defined earlier). The correlation was much higher for these rare tag SNPs than for the common tag SNPs in scenario 2. Bias in the estimate at the tag SNP was, on average, more severe when power was low (Table 3.2), and relative selection bias at \( r/ni \) was also more severe when power was low, although that bias was usually smaller at the causal SNP than at the tag SNP. When we examined cases for which the tag SNP bias was similar (152–186%), bias for the \( r/ni \) estimate tended to be higher when correlation was higher. This demonstrates how with higher correlation, more of the tag SNP bias is transferred to the causal SNP.
3.5 Discussion

Our mini-exome two-stage study design was different from a genome-wide two-stage study design in several ways. The sample size was small (hundreds instead of thousands of SNPs), the \( p \)-value threshold for selection was large (\( \alpha = 0.05 \) instead of \( 10^{-6} \)), and the correlation was less than ideal in most cases. The low correlation resulting from linkage disequilibrium between the causal and tag SNPs attenuated the effect size at the tag SNP, causing tag SNP estimates to fall well below the true causal SNP effect size. With a better set of tag SNPs, the attenuation would be less severe. To obtain an adequate number of significant data sets for each tag SNP, we used a liberal threshold for selection. A genome-wide significance level (\( 10^{-6} \)) would have decreased power, causing selection bias to be more severe.

Dickson et al. [10] demonstrated how multiple rare causal SNPs can be correlated with a single common tag SNP and can produce an apparent association at that tag SNP. Their simulation studies also showed that this synthetic association can occur over long ranges, often longer than would be covered by targeted resequencing around an associated GWAS SNP. In our study, we searched for tag SNPs that were correlated with \( r_i/n_i \) within the same gene and did not find more than two. In practice, expanding the examined regions may detect additional tag SNPs correlated with the \( r_i/n_i \) statistic for the region, but at the cost of additional sequencing.

Recently, methods have been proposed to correct for upward bias in genetic effect estimates at a GWAS tag SNP caused by selection at that same tag SNP [3, 11]. Extensions of the methods are needed for two-stage designs in which upward bias at one or more stage 2 sequencing SNPs is caused by selection at an imperfectly correlated stage 1 GWAS tag SNP. Selection bias is of
practical importance when designing a replication study. If the sample size is estimated from an upward-biased estimate of effect size at the causal SNP, then the replication study may be underpowered to detect the true association. Reliable estimates of genetic effect are also important for clinical interpretation and estimation of the proportion of heritability explained.

3.6 Conclusions

Targeted resequencing following a GWAS is becoming a cost-effective way to uncover causal variants not included in commercial genotyping chips. Because of low power, many association studies will be of low to moderate significance and follow-up studies will be required to confirm the associations. If the causal SNP genetic effect estimate in the original study is biased, then power to detect this SNP in a follow-up study will also be overestimated, and true associations not replicated because of low power may be misinterpreted as null associations. One might expect that when tag and causal SNPs are not well correlated, the effect of selection on the causal SNP estimate will be negligible. Our work indicates that this is not the case. To avoid discarding disease-associated SNPs in the follow-up stage, it is critical that investigators account for selection bias in the original resequencing study.
### 3.7 Tables

**Table 3.1 Bias in genetic effect estimates for a common tag SNP and multiple rare causal SNPs**

| Trait | Gene | Population | SNP     | Population | SNP     | Significance level for additive test | Estimated power for additive test | Over all data sets | Over data sets with significant tag effect | Relative bias (%) | Correlation between tag SNP and \( r/n \) | Mean effect estimate | Over all data sets | Over data sets with significant tag effect | Relative bias (%) |
|-------|------|------------|---------|------------|---------|--------------------------------------|----------------------------------|-------------------|---------------------------------------------|-------------------|---------------------------------------------|-------------------|-------------------|---------------------------------------------|-------------------|---------------------------------------------|-------------------|
| Q1    | KDR  | CEPH       | C4S1878 | 0.001      | 0.15    | 0.63                                 | 0.93                             | 49                | 0.41                                        | 5.51              | 6.19                                        | 12                | 6.69              | 7.08                                        | 6                 |
| DS    | PIK3C2B | Tuscan     | C1S9170 | 0.05       | 0.10    | 0.87                                 | 2.26                             | 160               | 0.40                                        | 6.69              | 7.08                                        | 6                 | 6.69              | 7.08                                        | 6                 |
| DS    | PIK3C2B | Tuscan     | C1S9171 | 0.05       | 0.10    | 0.87                                 | 2.26                             | 160               | 0.40                                        | 6.69              | 7.08                                        | 6                 | 6.69              | 7.08                                        | 6                 |
| DS    | PTK2B | CEPH       | C8S911  | 0.05       | 0.10    | 0.67                                 | 1.49                             | 122               | 0.51                                        | 1.99              | 2.91                                        | 46                | 1.99              | 2.91                                        | 46                |
| DS    | PTK2B | Chinese    | C8S925  | 0.05       | 0.06    | 0.21                                 | 1.23                             | 476               | 0.47                                        | 0.34              | 1.53                                        | 350               | 0.34              | 1.53                                        | 350               |

*Note: Results for scenario 2. Values computed as described in the Results section. DS is disease status.*
Table 3.2 Bias in genetic effect estimates for a rare tag SNP and multiple rare causal SNPs

<table>
<thead>
<tr>
<th>Trait</th>
<th>Gene</th>
<th>Population</th>
<th>SNP</th>
<th>Significance level for additive test</th>
<th>Estimated power for additive test</th>
<th>Over all data sets</th>
<th>Over data sets with significant tag effect</th>
<th>Relative bias (%)</th>
<th>Correlation between tag SNP and $r/n$</th>
<th>Mean effect estimate Over all data sets</th>
<th>Mean effect estimate Over data sets with significant tag effect</th>
<th>Relative bias (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>FLT4</td>
<td>Chinese</td>
<td>C5S141</td>
<td>0.01</td>
<td>0.075</td>
<td>0.52</td>
<td>1.00</td>
<td>94</td>
<td>0.46</td>
<td>3.26</td>
<td>4.05</td>
<td>24</td>
</tr>
<tr>
<td>Q2</td>
<td>PLAT</td>
<td>Luhya</td>
<td>C8S1797</td>
<td>0.05</td>
<td>0.105</td>
<td>0.29</td>
<td>0.98</td>
<td>241</td>
<td>0.46</td>
<td>0.37</td>
<td>1.71</td>
<td>356</td>
</tr>
<tr>
<td>Q2</td>
<td>SREBF1</td>
<td>CEPH</td>
<td>C17S1023</td>
<td>0.05</td>
<td>0.135</td>
<td>0.59</td>
<td>1.63</td>
<td>177</td>
<td>0.81</td>
<td>2.00</td>
<td>3.97</td>
<td>98</td>
</tr>
<tr>
<td>Q2</td>
<td>VLDLR</td>
<td>Tuscan</td>
<td>C9S404</td>
<td>0.01</td>
<td>0.055</td>
<td>0.84</td>
<td>2.28</td>
<td>173</td>
<td>0.51</td>
<td>0.99</td>
<td>2.53</td>
<td>155</td>
</tr>
<tr>
<td>Q2</td>
<td>VLDLR</td>
<td>Tuscan</td>
<td>C9S471</td>
<td>0.01</td>
<td>0.055</td>
<td>0.84</td>
<td>2.28</td>
<td>173</td>
<td>0.51</td>
<td>0.99</td>
<td>2.53</td>
<td>155</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>CEPH</td>
<td>C12S193</td>
<td>0.01</td>
<td>0.070</td>
<td>0.34</td>
<td>0.93</td>
<td>171</td>
<td>0.65</td>
<td>0.61</td>
<td>1.12</td>
<td>85</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>CEPH</td>
<td>C12S200</td>
<td>0.01</td>
<td>0.080</td>
<td>0.41</td>
<td>1.15</td>
<td>182</td>
<td>0.61</td>
<td>0.57</td>
<td>1.04</td>
<td>84</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>CEPH</td>
<td>C12S203</td>
<td>0.05</td>
<td>0.075</td>
<td>0.21</td>
<td>0.53</td>
<td>157</td>
<td>0.48</td>
<td>0.63</td>
<td>0.93</td>
<td>47</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>CEPH</td>
<td>C12S211</td>
<td>0.01</td>
<td>0.070</td>
<td>0.34</td>
<td>0.87</td>
<td>152</td>
<td>0.78</td>
<td>0.66</td>
<td>1.39</td>
<td>113</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>Japanese</td>
<td>C12S193</td>
<td>0.05</td>
<td>0.070</td>
<td>0.21</td>
<td>0.72</td>
<td>244</td>
<td>0.56</td>
<td>-0.12</td>
<td>0.36</td>
<td>403</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>European</td>
<td>C12S193</td>
<td>0.05</td>
<td>0.200</td>
<td>0.31</td>
<td>0.79</td>
<td>153</td>
<td>0.62</td>
<td>0.55</td>
<td>1.03</td>
<td>87</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>European</td>
<td>C12S200</td>
<td>0.01</td>
<td>0.070</td>
<td>0.30</td>
<td>0.92</td>
<td>203</td>
<td>0.45</td>
<td>0.49</td>
<td>0.95</td>
<td>93</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>European</td>
<td>C12S203</td>
<td>0.05</td>
<td>0.070</td>
<td>0.18</td>
<td>0.50</td>
<td>180</td>
<td>0.46</td>
<td>0.58</td>
<td>0.85</td>
<td>47</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>European</td>
<td>C12S211</td>
<td>0.01</td>
<td>0.055</td>
<td>0.32</td>
<td>0.90</td>
<td>186</td>
<td>0.74</td>
<td>0.59</td>
<td>1.22</td>
<td>107</td>
</tr>
<tr>
<td>Q2</td>
<td>VWF</td>
<td>All</td>
<td>C12S203</td>
<td>0.05</td>
<td>0.140</td>
<td>0.16</td>
<td>0.41</td>
<td>160</td>
<td>0.48</td>
<td>0.69</td>
<td>1.06</td>
<td>54</td>
</tr>
<tr>
<td>DS</td>
<td>PTK2B</td>
<td>CEPH</td>
<td>C8S883</td>
<td>0.05</td>
<td>0.075</td>
<td>0.95</td>
<td>2.40</td>
<td>153</td>
<td>0.53</td>
<td>2.35</td>
<td>13.83</td>
<td>489</td>
</tr>
<tr>
<td>DS</td>
<td>PTK2B</td>
<td>Chinese</td>
<td>C8S885</td>
<td>0.05</td>
<td>0.050</td>
<td>0.16</td>
<td>1.24</td>
<td>692</td>
<td>0.87</td>
<td>0.38</td>
<td>2.68</td>
<td>610</td>
</tr>
<tr>
<td>DS</td>
<td>PTK2B</td>
<td>European</td>
<td>C8S911</td>
<td>0.05</td>
<td>0.130</td>
<td>0.63</td>
<td>1.27</td>
<td>102</td>
<td>0.50</td>
<td>2.04</td>
<td>2.86</td>
<td>40</td>
</tr>
</tbody>
</table>

Note: Results for scenario 3. Values computed as described in the Results section. DS is disease status.
3.8 Figures

Figure 3.1 Distribution of estimates for a common tag SNP and a common causal SNP

Boxplot of estimates of genetic effect at a tag SNP (GWAS stage 1) and causal SNP C6S5380 (sequencing stage 2) on quantitative trait Q2 over 200 replicates with and without selection at a stage 1 tag SNP for additive genetic effect using a selection threshold $p < 0.05$. $p_c$ is the MAF of the causal SNP, $p_T$ is the MAF of the tag SNP, $\rho_{TC}$ is the correlation between the tag SNP and the causal SNP. Horizontal lines are the null effect size (zero) and the mean of causal SNP genetic effect estimates without selection. Because of sampling variation, the mean is different from the median (band in middle of boxplots).
3.9 References


9. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81:559–575

Chapter 4. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification

4.1 Abstract
Next generation sequencing has dramatically increased our ability to localize disease-causing variants by providing base-pair level information at costs increasingly feasible for the large sample sizes required to detect complex-trait associations. Yet, identification of causal variants within a candidate associated region remains a challenge. Counter-intuitively, certain factors that increase power to detect an associated region can decrease power to localize the causal variant. First, combining GWAS with imputation and/or low coverage sequencing to achieve the large sample sizes required for high power can have the unintended effect of producing differential genotyping error among SNPs. This tends to bias the relative evidence for association toward better genotyped SNPs. Second, re-use of GWAS data for fine mapping exploits previous findings to ensure genome-wide significance in GWAS-associated regions. However, using GWAS findings to inform fine-mapping analysis can bias evidence away from the causal SNP toward the tag SNP and SNPs in high LD with the tag. Together these factors can reduce power to localize the causal SNP by more than half. Other strategies commonly employed to increase power, namely increasing sample size and using higher density genotyping arrays, can, in certain common scenarios, actually exacerbate these effects and further decrease power to localize causal variants. We develop a re-ranking procedure that accounts for these adverse effects and substantially improves the accuracy of causal SNP identification, often doubling the probability that the causal SNP is top-ranked. Application to the NCI BPC3 aggressive prostate cancer
GWAS with imputation meta-analysis identified a new top SNP at 2 of 3 associated loci and identified several additional possible causal SNPs at these loci that may have otherwise been overlooked. This method is simple to implement using R scripts provided on the author’s website.

4.2 Introduction

The challenges of precise identification of disease-causing variants underlying GWAS signals have recently received much attention [1,2,3]. For post-GWAS statistical analysis that aims to accurately identify potentially causal variants, a major hurdle is the development of methods to distinguish disease-causing variants from their highly-correlated proxies. While GWAS-era statistical methods focused on identifying associated regions via tag SNPs at the coarse scale of GWAS arrays, next generation sequencing (NGS) technology offers the capability to not only detect associated regions, but to distinguish the causal SNPs within these associated regions. Here we make a distinction between ranking SNPs across the genome to identify an associated region, and ranking to pinpoint the potential causal variant within an associated region.

Identifying an associated region requires that trait-associated SNPs be ranked above null SNPs, while identifying the causal variant requires that among associated SNPs, associations due to causality are ranked above indirect associations due to other factors, e.g. linkage disequilibrium (LD). Zaitlen et al [4] proposed a measure of performance for sequencing and fine mapping analysis: their localization success rate metric is the probability that the causal SNP has the top-ranked test statistic within an associated region. When multiple SNPs are in high LD, the localization success rate drops dramatically as expected [5]. Udler et al (2010) investigated the
difficulty in overcoming the stochastic effect of high LD among causal and non-causal SNPs [5]. The sample size required to distinguish the causal SNP can be 1 to 4 times the size required to detect the effect at genome-wide significance. Zaitlen et al [4] showed that this problem could be overcome through joint analysis of samples from carefully selected populations with differing LD structure. Although candidate causal SNPs will require further bioinformatic or functional study to ultimately delineate potential causal mechanisms, optimized study design and analysis can point to the best possible candidate causal SNP(s) and help develop testable hypotheses about biological mechanisms.

Studies of complex traits now underway are leveraging the cost efficiency of integrating GWAS, low- and high-coverage sequencing, and imputation to achieve sample sizes in the tens of thousands [6,7]. For example, the Genetics of Type 2 Diabetes (GoT2D) study is combining low and high-coverage sequencing with 2.5M-SNP GWAS genotyping and imputation to achieve a total sample size of over 28,000 [8]. Sequencing the GWAS sample exploits the GWAS findings to ensure that an association signal is present at the genome-wide level and eliminates the cost of recruiting new individuals. Analysis of sequenced and imputed SNPs (which we refer to as post-GWAS in this thesis) can thus be informed by previous GWAS results, allowing a prioritized use of post-GWAS data in fine-mapping regions surrounding significant GWAS tag SNPs [9-11]. Selection of associated regions for further studies can also be based on combined GWAS and post-GWAS criteria [12,13]. For example, the WTCCC [13] required a marginally significant (p-value < 10^{-4}) GWAS SNP to support the evidence at a genome-wide significant imputed SNP. However, these strategies lead to two important issues that have received little attention in the context of causal SNP identification: (1) the effect of the
use of successful GWAS data for fine mapping (2) the effect of genotyping error rates that differ between sequenced or imputed SNPs.

The use of GWAS data that had contributed to the identification of an associated region for fine-mapping can adversely affect accurate causal SNP identification. For example, the simulation study of Wiltshire et al [14] showed that when a significant GWAS tag SNP is followed up by sequencing in the same sample, the tag SNP is in fact ranked higher than the true causal SNP 30% to 63% of the time, depending on the genetic model and effect size. When a GWAS tag SNP is selected based on small p-value, the magnitude of the association at the tag tends to be over-estimated; this form of selection bias is also known as the winner’s curse [15-19]. To a variable extent, depending on the LD pattern, this selection bias is carried over from the GWAS tag to fine-mapping sequenced or imputed SNPs [20]. While this earlier work empirically demonstrated the effect of selection for a significant GWAS tag SNP on the causal SNP, no work to date explores whether it also affects the rank of the causal SNP among all neighboring SNPs within an associated region, and if so how to correct for the bias.

High error rates and differences in error rates, due to differences in coverage, read length and depth, minor allele frequency (MAF), GC content, local sequence structure, and other sequence-specific factors, are common to NGS SNPs are well-recognized obstacles to analysis [21-29]. Error rates for low-read-depth sequencing studies are estimated to be 1%-3% [22,30,31], and as little as 1% error can produce a large loss in power [27]. The strategy of low-coverage sequencing in a portion of GWAS samples has been used to discover sequence variants and build a reference panel to drive imputation in the remaining samples, but the imputation accuracy can be worse than if all samples were sequenced [25]. The choice of lower-coverage
design is also motivated by reports that low-coverage sequencing in a large sample, alone or combined with GWAS and imputation data, can achieve superior power to detect associations compared to high-coverage sequencing in a small sample with similar cost [25,29,32,33]. However, whether the localization success rate of the causal variants responsible for these associations is similarly high has not yet been examined. High error rates that differ among SNPs also occur in high-coverage sequencing; for example, within targeted high-coverage regions, highly repetitive elements can be difficult to capture resulting in low accuracy for some SNPs [34].

Differential genotyping accuracy between studies has been shown to reduce the power of meta-analysis in the imputation setting [35], and differential accuracy between cases and controls has been shown to cause confounding and elevated type I errors [36, 37]. Accounting for differential genotyping accuracy in the association test can recover some of the lost power and reduce type I error [35,36]. However, whether it affects the ability to distinguish causal SNPs from correlated SNPs, and how best to account for the effect of differential genotyping accuracy jointly for all SNPs (GWAS tagged, imputed or sequenced) is an open question.

In this report, we first demonstrate that:

(1) Localization success rate decreases as the correlation between the tag and causal SNP increases.

(2) Selection at the tag SNP exacerbates this problem by increasing the magnitude of the association evidence at the tag SNP itself and at other neighboring SNPs in higher LD with the tag relative to the causal SNP.
(3) Differential genotyping or imputation error between SNPs further decreases localization success rate, with or without the tag selection.

(4) This problem can be exacerbated by increasing sample size, if genotyping accuracy at the causal SNP is lower than at neighboring SNPs.

We develop an analytic description of how these factors influence the probability of localization success and evaluate this probability for a range of plausible values. We then show how to properly adjust for the effects of these factors with a re-ranking procedure. We evaluate the performance of the method with extensive simulation studies under a wide range of realistic scenarios, and we demonstrate the practical effect of re-ranking with an application to the NCBI BPC3 aggressive prostate cancer GWAS with imputation [47].

4.3 Materials and Methods

Suppose that $M$ sequenced (or imputed) SNPs, $S_i$, $i = 1, ..., M$, in the region surrounding a significant GWAS tag SNP $G$ are ranked by the magnitude of their association statistics in order to identify the causal SNP $C$. Table 4.1 provides the notation for the various parameters and statistics used throughout the report. Briefly, $T_{Si}$ is the Wald test statistic at a sequenced SNP $S_i$; $\hat{r}_{GS_i}$ is the sample Pearson correlation coefficient between the imputed/sequenced genotypes (most likely or fractional allele dosage) for SNPs $G$ and $S_i$ ($r^2$ is the well-known pair-wise correlation measure of LD between two SNPs); $\hat{\rho}_{Sl}$ is the estimated correlation between the true genotype and the called genotype for a sequenced SNP $S_i$ (we use correlation as a measure of genotyping accuracy because of its simple interpretation in terms of power and genotyping quality; this quantity is provided by both MACH [24] and BEAGLE [45] software); $\delta_G$ and $\delta_{Sl}$
are proportions of samples with non-missing genotypes (termed call rates) at SNPs $G$ and $S_i$, respectively, and $\delta_{GS_i}$ is the joint call rate, the proportion of samples with non-missing genotypes at both SNPs, and $\delta_C$ is the call rate at the causal SNP.

Let $\hat{\Delta}_G$ be an estimate of the selection bias in genetic effect estimation at the tag SNP $G$ (described further below), that is the excess in the expected value of the test statistic $T_G$ at the tag SNP $G$ induced by selection based on its small p-value (or high rank). We call this phenomenon the selection effect ($\Delta_G$ is zero if the region was not selected via a tag SNP that achieved the given significance or ranking criterion in the same sample). Our proposed re-ranking statistic for a sequenced SNP $S_i$ is

$$T_{S_i}^* = \frac{T_{S_i} - \frac{\delta_{GS_i}}{\delta_G} \hat{\Delta}_G}{\hat{\rho}_{S_i} \sqrt{\delta_{S_i}}} \text{ for } i = 1,...,M. \tag{1}$$

Equation (1) depends on the selection effect $\hat{\Delta}_G$, the tagging effect $\hat{\rho}_{GS_i} \hat{\Delta}_G$, the genotyping accuracy effect $\hat{\rho}_{S_i}$ and scaling factors that depend on the call rates. Justification for equation (1) is as follows (Full details are provided in Appendix 4).

Without loss of generality, let $\beta > 0$ be the genetic effect (e.g. the log odds ratio or the regression coefficient in the model relating the phenotype and genotype) at the causal SNP $C$ which could be: one of the sequenced or imputed SNP $S_i$, $i = 1,...,M$; the GWAS tag SNP $G$ although this is unlikely; or neither if the genomic coverage was incomplete. Let the tag SNP $G$ be coded such that the coded allele is positively correlated with the causal allele. Let $\hat{\beta}$ be the genetic effect estimate and $\hat{\sigma}(\hat{\beta})$ be the estimated standard deviation (SD) of the estimate from $n$ observations. We assume that the distribution of the Wald test statistic at the causal SNP,
\( T_c = \frac{\beta}{\sigma(\beta)} \) is approximately normal \( T_c \sim N(\sqrt{n} \mu_c, 1) \), where \( \mu_c = \frac{\beta}{\sigma(\beta)} \frac{1}{\sqrt{n}} \). The following also applies to test statistics that are asymptotically equivalent to the Wald test statistic.

We first note that the selection effect in the observed test statistic \( T_{G_{obs}} \) at the GWAS tag SNP \( G \) is

\[
\Delta_G = T_{G_{obs}} - \sqrt{n} r_{CG} \mu_c. \quad (2)
\]

Here \( r_{CG} \) is the correlation between the genotypes of the causal \( C \) and the tag SNP \( G \). (We assume that the tag is coded so that it is positively correlated with the risk allele of the causal SNP.) The value of \( r_{CG} \) is unobserved and needed only in the theoretical formation of the problem not in the practical implementation, which we discuss later. The selection effect is most pronounced when there is low power at the tag SNP. For example, consider the case where we examine the region only if the tag SNP test statistic is significant at level \( \alpha \) (i.e. \( T_G \sim N(\sqrt{n} r_{CG} \mu_C, 1) \) and \( T_G > Z_\alpha \)). Then, using the Inverse Mills Ratio [38], the expected bias at the tag SNP is

\[
E[T_{G_{obs}} | T_{G_{obs}} > Z_\alpha] - E[T_{G_{obs}}] = \frac{\phi(Z_\alpha - \sqrt{n} r_{CG} \mu_C)}{1 - \Phi(Z_\alpha - \sqrt{n} r_{CG} \mu_C)}, \quad (3)
\]

where \( \phi \) is the standard normal density function and \( \Phi \) is the standard normal cumulative distribution function. Here we assume that \( E[T_{G_{obs}} | T_{G_{obs}} > Z_\alpha] \approx E[T_{G_{obs}} | T_{G_{obs}} > Z_\alpha] \), because the tag and causal SNPs are positively correlated and \( Pr(T_{G_{obs}} < -Z_\alpha) \) will be very small at genome-wide significance. This bias increases when the critical value \( Z_\alpha \) is extreme.
compared to \( \sqrt{n}r_{CG}\mu_C \), the expected value of \( T_G \) at the tag SNP, that is when there is low power to detect the tag SNP. In a similar manner, the expected value of \( T_G \) conditional on the tag SNP \( G \) achieving top-rank is inflated.

Second, the conditional distribution of the test statistic \( T_{S_i} \) at the sequenced SNP \( S_i \), conditional on the value of the observed test statistic \( T_{Gobs} \) at the tag SNP \( G \), is

\[
\{T_{S_i} \big| (T_G = T_{Gobs}) \} \sim N\left( \sqrt{n}r_{CSi}\mu_C + r_{GSi}\Delta_G, 1 - r_{GSi}^2 \right).
\]

The first term, \( \sqrt{n}r_{CSi}\mu_C \) is the association signal at the sequencing SNP, the second term \( r_{GSi}\Delta_G \) is the distortion due to the tag SNP. Therefore, \( \Delta_G \), the selection effect at the GWAS tag SNP \( G \) carries through to each sequenced SNP \( S_i \) in proportion to the correlation \( r_{GSi} \) between \( G \) and \( S_i \).

The combination of attenuation due to LD and upward selection bias at the tag, \( r_{GSi}\Delta_G \), distorts the association evidence so that SNPs in high LD with the tag are more likely to be top-ranked. We call this phenomenon the *tagging effect*, and use an estimate to remove bias from the conditional expected value of \( T_{S_i} \) in (4).

Third, differential call rates among SNPs (\( \delta_G \), \( \delta_{S_i} \) and \( \delta_{GSi} \)) and estimated genotyping accuracy (\( \hat{\rho}_{S_i} \) is the estimated and \( \rho_{S_i} \) is the actual correlation between the called genotype and true genotype) of sequenced or imputed SNP \( S_i \) appear in both the numerator and denominator of equation (1). In the numerator, the tagging bias, \( \hat{r}_{GSi}\Delta_G \) is scaled by a factor of \( \frac{\delta_{GSi}}{\sqrt{\delta_{S_i}\delta_G}} \) because correlation between the test statistics depends on the individual and joint call rates at the two SNPs. The bias-corrected statistic in the numerator is scaled by \( 1/\sqrt{\delta_{S_i}\hat{\rho}_{S_i}} \) because
\[ T_{S_i} \sim N\left( \sqrt{\bar{n}} \sqrt{\delta_{S_i} r_{CS_i}^* \mu_c}, 1 \right), \quad (5) \]

where \( r_{CS_i}^* \) is the correlation between the genotype of the causal SNP and the called or estimated genotype of the sequenced SNP (in contrast to \( r_{CS_i} \), for the true genotype of the sequenced SNP). Assuming the probability of genotyping error is independent of the actual genotype, then \( r_{CS_i}^* \approx \hat{\rho}_{S_i} r_{CS_i} \). It is clear that, without correction, smaller \( \rho_{S_i} \) (higher genotyping error) and smaller \( \delta_{S_i} \) (higher missing data rate) tend to lower the probability that SNP \( S_i \) would be top-ranked. We call this phenomenon the genotyping accuracy effect.

4.4 Results

4.4.1 Analytical studies of the adverse effects of selection, tagging and genotyping accuracy on the localization success rate

To conceptually demonstrate the joint effects of selection, tagging and genotyping accuracy on the localization success rate (the probability that the causal SNP is topped ranked within an associated region), we first consider the simplified case of 2 SNPs, one causal C (from sequencing or imputation) and one tag G (from GWAS) with correlation between the two SNPs ranging from \( r = 0.2 \) to 1 (from almost no LD to perfect LD). GWAS coverage is often lower than expected. The coverage of GWAS platforms tends to be overestimated for both sequenced and imputed SNPs (see Appendix 5 for further discussion of this point). We assume the MAFs of both SNPs are 0.12, the causal SNP has a log additive odds ratio (OR) of 1.25, and selection at the tag SNP, if present, is based on its association test p-value < 0.05 in a sample of 1000 cases.
and 1000 controls. Localization success rates for all figures were computed based on the analytical results in equations (2)-(4) and by numerically integrating over the following bivariate normal density function,

$$\left( \begin{array}{c} T_c \\ T_g \end{array} \right) \sim N \left( \begin{array}{cc} \sqrt{\eta \mu_c} & \sqrt{\delta_c \rho_c} \\ \sqrt{\delta_c \rho_c r_{cg}} & 1 \end{array} \right) \left( \begin{array}{c} 1 \\ \sqrt{\delta_c \rho_c r_{cg}} \end{array} \right).$$  \hspace{1cm} (6)

Analytical evaluations of equation (6) for this simple scenario give insight into the relative influence of the tagging and genotyping accuracy effects outlined in the introduction (Figures 4.1-4.3). We find similar patterns of influence for a rare SNP (MAF 0.02) and a more common SNP (MAF 0.25), and when the number of non-causal SNPs increases (Figures 4.7-4.12).

(1) **Tight linkage disequilibrium between SNPs can obscure the causal SNP (Figure 4.1).**

Figure 4.1 left panel shows that as the correlation between tag and causal SNPs increases (X-axis), the expected association evidence at the tag, $E[T_{tag}]$, approaches $E[T_{causal}]$ (Figure 4.1 panel A), resulting in a lower localization success rate (Figure 4.1 panel B). As expected, increasing the number of non-causal SNPs in strong correlation with the causal SNP increases competition for the top rank and decreases the localization success rate (Figures 4.13-4.15).

Increasing the number of non-causal SNPs in competition with the causal from 1 to 6 decreases the localization success rate from over 50% to less than 35% (bottom left panels of Figures 4.1 and 4.15).

(2) **Selection at the tag SNP inflates the association evidence at the tag, increasing the probability that it out-ranks the causal SNP (Figure 4.1).**

Figure 4.1 right panel shows that tag selection reduces the difference between the
expected association evidence at the tag and the causal, that is $E[T_{tag}]$ and $E[T_{causal}]$, compared with no selection, regardless of the LD between the two SNPs (Figure 4.1 panel C vs. 4.1 panel A). Consequently, the localization success rate conditional on selection can be reduced by 25% as compared to the unconditional localization success rate (Figure 4.1 panel D vs. Figure 4.1 panel B). Results are similar for the rare SNP and more common SNP cases (Figures 4.7 and 4.8).

(3) **Sequencing or imputation error decreases the localization success rate, with or without tag selection (Figure 4.2).**

Low genotyping accuracy at the causal SNP reduces the expected value of the causal SNP test statistic, leading to decreased localization success rate (Figure 4.2). For example, if the tag SNP was genotyped with perfect accuracy ($\rho_c = 1$), while the causal SNP was not, and if the genotype error at the causal SNP resulted in $\rho_c = 0.80$ (the blue dash-dotted curve), then the localization success rate would be reduced by an additional 10%-30% as compared to perfect genotyping accuracy (the black solid curve). Results are similar for the rare SNP and more common SNP cases (Figures 4.9 and 4.10).

(4) **Counter-intuitively, sample size can reduce localization success rate (Figure 4.3).**

When the causal SNP is less accurately genotyped than one of its highly correlated proxies (i.e. $\rho_c < \rho_G$ and $r_{CG}$ is large), the proxy SNP may capture the association better than the causal SNP. As a result, this proxy SNP will out-rank the causal SNP more than 50% of the time. In this case, the localization success rate would be less than 50%, and would decrease further as sample size increases (Figure 4.3). For example, if $\rho_c = 0.95$, $\rho_G = 1$ and $r_{CG} = 0.98$ (red dashed line), the localization success rate drops from 47% to 26% as sample size increases from
100 to 10,000. Lower $\rho_c$ would lead to even lower localization success rates (results not shown).

We also note that, depending on the NGS experiment or the imputation parameters (e.g. the matching between the reference and imputation sample) for estimated genotype at the causal SNP $C$, the $\rho_c$ may not be lower bounded by the tagging $r_{CG}$, which we discuss further in the Discussion section. This pattern is similar for the rare SNP and more common SNP cases (Figures 4.11 and 4.12).

### 4.4.2 Practical implementation of the post-GWAS re-ranking statistics

The above analytical results demonstrate the need to correct for the joint effects of selection, tagging and genotyping accuracy on the localization success rate. The practical implementation of the proposed re-ranking statistic in equation (1) is as follows. The estimated selection bias $\Delta G$ at the tag SNP $G$ can be obtained using $BR$-squared that provides Bias-Reduced estimates via Bootstrap Resampling at the genome-wide level [39,40]. (The original program, designed to provide estimates for the genetic effect $\beta$, has been modified slightly to provide estimates for the test statistic $T$; see software documentation on author’s website for details). The bootstrap estimator can be applied whether the region of interest was selected by rank or by p-value threshold. Unlike the threshold-based likelihood methods [41-44], the genome-wide bootstrap method incorporates information across the entire GWAS in order to account for the effects of LD and rank on the bias at each SNP. The values of the individual and joint call rates $\delta$s are available from the dataset, and genotype correlation $\hat{r}_{GSi}$ in equation (1) can be estimated from the sample. Correlation between the actual and estimated genotypes at a sequenced SNP $\rho_{Si}$ can be obtained from the mean posterior genotype (e.g. MACH ratio of variances estimate, [24]) or from the full genotype posterior probabilities (e.g. BEAGLE allelic $r^2$ estimate [45]). An R script
that implements Equation 1 is available. The R script calls the BR2 software (http://www.utstat.toronto.edu/sun/Software/BR2/), which provides the essential quantity of delta G (if the original GWAS dataset was used for fine-mapping).

4.4.3 Simulation study design

We conducted extensive simulation studies to empirically evaluate the performance of the re-ranking method under five general scenarios (Table 4.2):

- **Scenario 1: GWAS used for discovery, sequencing/imputation used for fine-mapping around GWAS “hits” using same GWAS sample.** Scenario 1 is a GWAS-focused design based on the WTCCC Type 1 Diabetes substudy data. A significant region is identified by a significant GWAS tag SNP (p < 5x10^{-7}) and followed by fine-mapping with post-GWAS data (sequenced or imputed SNPs) in the region surrounding the tag SNP. The SNP with the largest test statistic in each region is selected as the best candidate causal SNP. Data is simulated as follows.

  - **GWAS Data and Tag SNP:** In order to generate realistic genome-wide correlation structure, we fixed the individual level genotypes from the WTCCC T1D sub-study as the GWAS data (1963 cases and 2938 controls). Among the reported WTCCC T1D significant regions, we randomly selected 12q24 109.82-111.49Mb as the region of interest and designated rs11066410 (MAF 4.8%) as the GWAS tag SNP.

  - **Sequencing/Imputation Data and Causal SNP:** We simulated sequencing data for the case-control study for the region of interest with 10 post-GWAS SNPs (MAF 4.8%), among which one is the causal SNP with OR=1.5. We varied the correlation between the tag and the causal SNP from r=0.78 (causal not well tagged by the
GWAS SNP) to 1 (the GWAS tag SNP is the causal, although this is an unlikely scenario). We introduced random error into post-GWAS SNP genotypes at per-allele rates 2%, 1.5%, 1%, 0.5%, 0.25% or 0%, so that average $\rho_S$ in each dataset was 0.82, 0.86, 0.90, 0.95, 0.97 or 1, respectively.

- **Phenotype:** Phenotype datasets significant ($p < 5 \times 10^{-7}$) at the GWAS tag SNP were simulated using an additive logistic model with causal SNP OR = 1.5.

- **Scenario 2: All GWAS and sequenced/imputed SNPs used for discovery and fine-mapping in the same dataset.** Here we assumed that all GWAS and post-GWAS SNPs are used to identify an associated region ($p < 5 \times 10^{-7}$), and the most significant SNP in the region is then identified as the best candidate causal SNP. GWAS tag SNP data was simulated with MAF 5%. Sequencing data were simulated as described in Scenario 1 with parameter values detailed in Table 4.2. Phenotype datasets significant ($p < 5 \times 10^{-7}$) at any GWAS or post GWAS SNP, were simulated using a logistic model with causal SNP OR = 2.

- **Scenario 3: Discovery and fine-mapping in different datasets.** In this scenario, the region of interest was discovered in a previous study, while sequencing is performed in an independent dataset without conditioning on significance of the GWAS tag SNP in the independent dataset. Genotype and phenotype data were simulated as in Scenario 2, except that phenotype datasets were not selected for significance.

- **Scenario 4: Multiple causal SNPs.** To explore the effect of multiple causal SNPs, we reconsidered scenario 3 but we assumed there are 11 fine-mapping sequenced/imputed SNPs, among which 2 are causal (both OR=2) (the correlation between the two causal SNPs was $r=0.73$, and MAF was 0.05).
• **Scenario 5: Missing Data.** Scenario 5 focuses on the effect of missing data (e.g. imperfect call rate). Genotype and phenotype data were simulated as in Scenario 3, except that genotyping accuracy was perfect. The missing rates were randomly assigned to each SNP so that the missing data proportion was between zero and twice the average error rate (Table 4.4).

The parameter values in Table 4.2 were chosen to best reflect realistic scenarios. For example, in order to address realistic tagging, we examined the Affymetrix 5.0 chip and identified the SNP that best captured each significant WTCCC T1D GWAS SNP. The correlation between the two SNPs ranges from $r = 0.79$ to $r = 1$. For the range of genotyping accuracy, we note that in practice, the average sequencing $\rho$ can vary substantially from study to study. For example, for low-coverage studies, it can vary from 0.63 to 0.99 depending on the coverage, MAF and sample size [25]. When low-coverage sequencing (4X) and imputation are combined, the average $\rho$ can range from 0.89 to 0.99 depending on the reference panel size [24]. Sequencing $\rho$ also depends on MAF; the same error rate in a lower MAF SNP results in a smaller $\rho$.

Even when average $\rho$ is high, $\rho$ at the SNP level can vary widely within a single study. Browning and Browning [45] found that imputation with a phased reference panel of 60 Hapmap CEU samples yielded a median $\rho$ of 0.95, however $\rho$ was less than 0.77 for 20% of individual SNPs. We show that coverage rates can vary widely between SNPs (Figure 4.6) by examining the 1000 Genomes low-coverage whole-genome pilot data from chromosome 1 in the CHB and JPT samples (Figure 4.6; October 2010 release; 1000 Genomes Project, 2010). We mimicked
this variability in our simulations by randomly assigning each SNP in each dataset an error rate that ranged from zero to twice the overall average error rate. No random error was introduced into the genotypes of the tag SNP ($\rho_{G} = 1$), because GWAS genotyping has been estimated to be over 99.8% accurate [13,46]. In order to ensure realistic correlation structure among post-GWAS sequencing/imputation SNPs, we examined all SNPs in the regions surrounding the WTCCC T1D significant SNPs using the HapMap3 dataset. The average correlation between adjacent SNPs in these regions was approximately 0.975.

4.4.4 Simulation Study Results

The main finding of the simulation studies is that GWAS-based region selection or moderate genotyping error can substantially reduce the probability of correctly identifying the causal SNP (Tables 4.3-4.4 and Tables 4.5). Simulations detailed in Table 4.5 demonstrate that the combined tagging and genotyping accuracy effect can reduce the localization success rate by over 30% (Table 4.5).

Re-ranking recovers much of this lost power to identify the causal SNP, increasing the localization success rates by 1.5- to 3-fold in many cases (Table 4.3). When genotyping accuracy is high, the power lost due to tagging is small and so re-ranking tends to have little effect.

For studies using GWAS-based selection (scenario 1), the adverse effects of tagging and genotyping accuracy on localization success rate are strongest when the causal SNP is well tagged (larger $r$) and less accurately sequenced/imputed (smaller $\rho$) (Tables 4.3, 4.4 and 4.5). High-density GWAS followed up with low-coverage sequencing would fall into this category.
Well-tagged causal SNPs tend to suffer from lower localization success rates because the near perfectly genotyped tag often captures the association better than the imperfectly sequenced or imputed causal SNP. Re-ranking corrects this problem, so that the localization success rate does not depend on how well the causal SNP is tagged, except when the tag SNP is in fact the causal SNP. In this case, the tagging and genotyping accuracy effects actually increase the localization success rate. After re-ranking, the localization success rate is similar to levels seen when the tag is not causal. We consider this a minor tradeoff, because the causal SNP is unlikely to be found among the GWAS SNPs for a number of reasons: GWAS SNPs are typically selected independent of the phenotype of interest and post-GWAS SNPs tend to greatly outnumber GWAS SNPs.

When the discovery sample is fine-mapped, but significance is not required at the GWAS-tag SNP (scenario 2), the genotyping accuracy effect alone could considerably reduce power to identify the causal variant (Table 4.3). When an independent sample is used for fine-mapping (scenario 3, Table 4.3), localization success rates are very similar to those seen in scenario 2. In both cases, the re-ranking method improves the probability of correctly identifying the causal SNP. The improvement is most pronounced when genotyping accuracy is low (2- to 4-fold improvement). When there is more than one causal variant (scenario 4), we find that re-ranking effectively increases localization success rates for both causal SNPs (scenario 4, Table 4.3). Imperfect call rates affect localization success rate in a similar manner to imperfect genotyping accuracy (scenario 5, Table 4.4). Equation (5) implies that a missing data rate of 0.80 should affect the distribution of the causal SNP test statistic in the same manner as a
sequencing $\rho$ of 0.89, and this is borne out in our simulations. The re-ranking procedure corrects for both missing data and genotyping error to the same degree.

In some cases, investigators are more interested in delimiting a set of best candidate causal SNPs instead of a single top SNP. We define an alternative localization success rate metric as the probability that the causal SNP is in the top 10% of SNPs by rank (Table 4.6). Briefly, we examine the probability that the causal SNP is among the top 5 SNPs when there are 50 total SNPs (ranked by test statistic or re-ranking statistic). Without re-ranking, the probability that the causal SNP is in the top 10% of SNPs over the region is moderate. Re-ranking provides an improvement up to 1.8-fold.

### 4.5 Application

Machiela et al [47] used the August 2010 release of the 1000 Genomes Project European-ancestry (EUR) panel to impute 11.6 million variants in 2,782 aggressive prostate cancer cases and 4,458 controls. These subjects were genotyped as part of the NCI Breast and Prostate Cancer (BPC3) Cohort Consortium aggressive prostate cancer GWAS [48]; genotyping platforms varied across the seven BPC3 studies, although all used versions of the Illumina HumanHap arrays and most used the Illumina HumanHap 610 Quad array. The correlation between imputed genotype dosage and genotypes thus varied across studies. Imputation and association analyses using imputed genotype dosages were conducted separately for each study, and the association results were combined via fixed-effect meta-analysis. For each imputed SNP, studies with imputation
r^2 < 0.8 were excluded from the meta-analysis test statistic, leaving a total of 5.8 million GWAS
or imputed SNPs.

Fine-mapping in the meta-analysis context ranks SNPs by the meta-analysis test statistic.
Re-ranking requires that we compute the correlation between the meta-analysis test statistic on
the Z-score scale (i.e. normally distributed test statistic) with and without accounting for
genotyping error. Assume $Z_j$ is the normally distributed test statistic for study $j$, and $w_j$ is the
weight for study $j$, the meta analysis test statistic used for the standard naïve ranking is

$$Z = \frac{\sum_j Z_j w_j}{\sqrt{\sum_j w_j^2}}.$$

If $\rho_j$ is an estimate of pair-wise correlation between the actual and imputed genotypes in study $j$
(e.g. allelic-$r^2$ [45], or ratio of variances $r^2$ [24]), it follows that the estimated correlation
between the meta analysis test statistic computed with perfectly genotyped SNPs ($Z_{act}$) and the
meta analysis test statistic computed with the observed imperfectly genotyped SNPs ($Z_{obs}$) is

$$\hat{\rho}_{meta} = \text{cor}(Z_{obs}, Z_{act}) = \frac{\sum_j w_j^2 \hat{\rho}_j}{\sum_j w_j^2}.$$

The re-ranking statistic in the meta-analysis case is

$$T_{S_{i}}^* = \frac{T_{S_{i}} - \bar{r}_{GS_{i}} \frac{\delta_{GS_{i}}}{\sqrt{\delta_{G}\delta_{S_{i}}}} \hat{\Delta}_{G}}{\sqrt{\delta_{S_{i}} \hat{\rho}_{meta,S_{i}}}},$$

where $T_{S_{i}}$ is the meta analysis test statistic $Z$ scaled for variance of 1.
Machiela et al [47] reported seven statistically independent associated sub-regions: five within the 8q24.21 locus (Figure 4.4, 4.16, 4.17, 4.18, 4.19) and one for each of 11q13.3 and 17q24.3 (Figure 4.5, 4.16). We selected all SNPs in LD (r^2 > 0.2) with the index SNP from each region for analyses (Figures 4.4 and 4.5, and Figures 4.16-4.18). In the application, we first ranked SNPs using the naïve test statistics [47]; and excluded any SNP with MAF <0.01; but unlike Machiela et al [47] we did not exclude any studies. Machiela et al selected significant regions by examining all imputed and genotyped SNPs at once and so we corrected for the imputation accuracy effect only (i.e. \( \Delta_c = 0 \)).

Re-ranking identifies new top SNPs for the 8q24.21 and 17q24.3 loci (Figures 4.4 and 4.5 respectively). In addition to the most significant region at 8q24.21 (Figure 4.4), re-ranking also identifies a new top SNP for the third most significant region (Figure 4.16 A and C). For both regions re-ranking also identifies SNPs that may have otherwise been missed due to imperfect imputation. After re-ranking, 2 SNPs in the most significant region at the 8q24.21 locus (Figure 4.4) and 8 SNPs at the 17q24.3 locus (Figure 4.5) move from the lower ranks into the top 10 percent. On the other hand, SNPs in the top 10% are moved down by only a few ranks. In this way, re-ranking keeps highly significant SNPs identified by the naïve ranking and adds a few SNPs that would have otherwise been missed. When the top test statistics are of similar size, re-ranking may identify a new top SNP. When most SNPs are well-genotyped, re-ranking makes only subtle changes (Figures 4.17, 4.18 and 4.19).

There is one poorly imputed SNP at 17q24.3 (rs1014000, r^2=0.20, poorly imputed in all studies) that moves from the naïve rank of 245 to the new rank of 16 after adjustment. This SNP’s apparent association is largely driven by data from a single study: the naïve rank in the
EPIC study is 10. When we remove this study from the meta-analysis, the naïve rank is 306 and the adjusted rank is 119. No other SNP in the top 10% is this drastically affected when this study is removed from the analysis. In the meta-analysis context, we recommend examining top SNPs for heterogeneity among studies when re-ranking produces dramatic results.

4.6 Discussion

Overall, we observed that the tagging and genotyping accuracy effects are non-trivial sources of bias that could obscure association evidence at the causal SNP. The proposed re-ranking procedure is simple to implement and can substantially increase the probability of identifying the causal SNP. For low-coverage sequencing, we recommend the re-ranking method to improve causal SNP identification. For imputation and high-coverage sequencing, we recommend that unfiltered SNPs in association regions be examined to see if correlation varies across SNPs and if so, we recommend adjustment with the re-ranking method. Large changes in rank should be carefully examined for underlying issues such as heterogeneity among meta-analysis studies or differential accuracy between cases and controls, and procedures to correct for these issues should be incorporated.

Re-ranking is most beneficial when genotyping accuracy is moderate to low, that is, the average correlation between the actual and estimated genotypes of post-GWAS (sequenced or imputed) SNPs is less than 0.97. A large number of post-GWAS SNPs in a study may appear to be significant, but when not all were directly genotyped with high accuracy, re-ranking can help select the most probable causal SNPs for follow up. High density genotyping followed by low-coverage sequencing in the same sample can produce misleading results, as demonstrated by our simulations, so we do not recommend this design for identifying causal variants. Our re-ranking
method tends to down-rank the tag SNP. If the tag SNP is suspected to be causal (e.g. based on prior study), we recommend examining the rank of the tag SNP using both the naïve and re-ranked methods when selecting SNPs for further study. Several imputation and sequencing software packages provide accurate estimates of $\rho$ or quantities from which $\rho$ can be computed [24, 45]. Re-ranking depends on accurate estimates of $\rho$. Recalibration of sequencing quality scores can greatly improve accuracy and so we recommend this step prior to re-ranking [27,56].

Re-ranking is especially important when study-specific factors exacerbate the effects of GWAS-based selection and genotyping error. Such factors include: repetitive elements which makes sequenced reads difficult to align [27] and low LD among SNPs or lack of population-specific reference panels which makes some populations particularly difficult to impute (e.g. some African populations [49]) resulting in imputation error as high as 10% for these populations. Low MAF SNPs tend to suffer from both low power (which exacerbates the tagging effect) and high genotyping error. Re-ranking can be applied to rare and low MAF SNPs with allele counts large enough for test statistics to reach asymptotic normality. Very low (1X-2X) and extremely low (0.1X-0.5X) read depth sequencing has received recent attention as a way to maximize cost efficiency and make use of off-target sequencing data [29, 32]. Error rates for such regions would be both very high and highly variable among SNPs and so re-ranking to account for errors in the estimated genotypes would be crucial. When genotyping accuracy is extremely poor, the re-ranking method may not be able to improve the localization success rate enough to ensure useful results. We recommend that investigators consider the accuracy thresholds recommended by the genotype calling or imputation algorithm they are using before re-ranking is applied.
Re-ranking improves the localization success rate when applied to SNPs under the alternative, i.e. SNPs that are themselves causal or in LD with a causal SNP. Including null SNPs in the re-ranking procedure increases the number of SNPs the causal must out-compete, and so we recommend that only SNPs suspected to be under the alternative be included. In our application we included all SNPs that had squared pairwise correlation ($r$) with the index SNP (most significant SNP in the region) greater than 0.2.

Existing methods that incorporate genotype uncertainty into tests for association to reduce power lost due to genotyping error or missing data [e.g. 50-52] do not completely recover lost power, and so the genotyping accuracy effect will remain. The simplest way to deal with genotype uncertainty in a test is to use the expected additive genotype (i.e. the posterior mean or dosage) in the standard linear or logistic regression. In this case, the re-ranking method can be applied using the allele dosages in place of called genotypes as described above. Guan and Stephens [53] compared several frequentist and Bayesian methods that incorporate genotype uncertainty into tests for association. The re-ranking procedure could be extended to any case where the correlation between test statistics or Bayes factors can be worked out.

We expect that re-ranking will play an important role as sequencing costs fall and GWAS platform coverage increases. Ultra-high density GWAS platforms are more likely to include tag SNPs in very high correlation with the causal SNP, which increases power to detect indirect association at the tag SNP. However, without re-ranking, strong tagging also decreases power to correctly identify the causal SNP in subsequent low-coverage sequencing. Advances in GWAS and sequencing platforms will allow researchers to drill down into lower MAFs (< 0.05) and smaller effect sizes. Both low MAF and small effect size yield lower power, which exacerbates
upward bias at the tag [20] and, therefore, the adverse tagging effect. Low MAF SNPs tend to suffer from higher error rates, which exacerbates the genotyping accuracy effect. Association study sample sizes will therefore need to continue to increase, so even as sequencing costs fall, it is anticipated that low-coverage will continue to be the most cost-effective design for many studies, despite the high sequencing error rates [27]. In conclusion, we anticipate that re-ranking to correct for the adverse effects of selection, tagging and differential genotyping accuracy rates among SNPs will continue to be important in candidate causal SNP identification for some time.
### 4.7 Tables

<table>
<thead>
<tr>
<th>Table 4.1. Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequenced or imputed SNPs indexed (i \in 1 \ldots M)</td>
</tr>
<tr>
<td>Causal SNP</td>
</tr>
<tr>
<td>GWAS tag SNP</td>
</tr>
<tr>
<td>Test statistic at sequenced SNP (i), causal SNP, GWAS tag SNP</td>
</tr>
<tr>
<td>Observed value of the test statistic at the tag SNP</td>
</tr>
<tr>
<td>Re-ranking statistic at sequenced SNP (i)</td>
</tr>
<tr>
<td><strong>Correlation between:</strong></td>
</tr>
<tr>
<td>Actual genotypes of casual and tag, causal and sequenced SNP (i), tag and sequenced SNP (i)</td>
</tr>
<tr>
<td>Estimated genotypes for the tag and sequenced SNP (i)</td>
</tr>
<tr>
<td>Actual genotype of the causal SNP and estimated genotype at sequenced SNP (i)</td>
</tr>
<tr>
<td>Call rate (1 - missing data rate) at sequenced SNP (i), tag SNP</td>
</tr>
<tr>
<td>Joint call rate for tag SNP and sequenced SNP (i)</td>
</tr>
<tr>
<td>Correlation between actual and estimated genotypes at: sequenced SNP (i), causal SNP, GWAS tag SNP</td>
</tr>
<tr>
<td>Estimated correlation (sample correlation) of the above</td>
</tr>
<tr>
<td>Tag selection bias ((E[T_G] - E[T_G]))</td>
</tr>
<tr>
<td>Genetic effect at the causal SNP, estimate</td>
</tr>
<tr>
<td>Standard deviation of the estimate at the causal SNP, estimate</td>
</tr>
<tr>
<td>Sample size</td>
</tr>
<tr>
<td>Expected value of the test statistic at the causal SNP re-scaled for sample size</td>
</tr>
<tr>
<td>(\mu_C = \frac{E[T]}{\sigma(\hat{\beta})\sqrt{n}} = \frac{\beta}{\sigma(\hat{\beta})\sqrt{n}})</td>
</tr>
<tr>
<td>Standard normal critical value at significance level (\alpha)</td>
</tr>
<tr>
<td>Standard normal cumulative distribution and density functions</td>
</tr>
</tbody>
</table>
Table 4.2 Parameters and parameter values of the main simulation studies.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Effect of selection, tagging, genotype accuracy</th>
<th>Effect of selection, genotype accuracy</th>
<th>Effect of genotype accuracy</th>
<th>Effect of multiple causal SNPs</th>
<th>Effect of missing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Same as the Discovery sample, conditional on significance at the GWAS tag SNP (p&lt;5x10^{-7})</td>
<td>Same as the Discovery sample, conditional on significance at any of the SNPs in the region (p&lt;5x10^{-7})</td>
<td>Independent sample</td>
<td>Independent sample</td>
<td>Independent sample</td>
</tr>
<tr>
<td>2</td>
<td>Same as the Discovery sample, conditional on significance at the GWAS tag SNP (p&lt;5x10^{-7})</td>
<td>Same as the Discovery sample, conditional on significance at any of the SNPs in the region (p&lt;5x10^{-7})</td>
<td>Independent sample</td>
<td>Independent sample</td>
<td>Independent sample</td>
</tr>
<tr>
<td>3</td>
<td>Same as the Discovery sample, conditional on significance at the GWAS tag SNP (p&lt;5x10^{-7})</td>
<td>Same as the Discovery sample, conditional on significance at any of the SNPs in the region (p&lt;5x10^{-7})</td>
<td>Independent sample</td>
<td>Independent sample</td>
<td>Independent sample</td>
</tr>
<tr>
<td>4</td>
<td>Same as the Discovery sample, conditional on significance at the GWAS tag SNP (p&lt;5x10^{-7})</td>
<td>Same as the Discovery sample, conditional on significance at any of the SNPs in the region (p&lt;5x10^{-7})</td>
<td>Independent sample</td>
<td>Independent sample</td>
<td>Independent sample</td>
</tr>
<tr>
<td>5</td>
<td>Same as the Discovery sample, conditional on significance at the GWAS tag SNP (p&lt;5x10^{-7})</td>
<td>Same as the Discovery sample, conditional on significance at any of the SNPs in the region (p&lt;5x10^{-7})</td>
<td>Independent sample</td>
<td>Independent sample</td>
<td>Independent sample</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample sequenced</th>
<th># of GWAS tag SNPs in the region, G</th>
<th># of post-GWAS SNPs (# of causal SNPs) in the region, M</th>
<th>OR of the causal SNP(s), β</th>
<th>MAF of the tag and post-GWAS SNPs</th>
<th>Correlation between the tag and causal SNP(s), ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample sequenced</td>
<td>1</td>
<td>10 (1)</td>
<td>1.5</td>
<td>4.8%</td>
<td>0.78, 0.83, 0.85, 0.90, 0.93, 0.95, 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.90, 0.93, 0.95, 0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.95</td>
</tr>
</tbody>
</table>

**Correlation between two adjacent non-causal post-GWA SNPs, \( r \):**

<table>
<thead>
<tr>
<th>Sample sequenced</th>
<th>Correlation between two adjacent non-causal post-GWA SNPs, ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>0.975</td>
</tr>
</tbody>
</table>

**Correlation between the actual and called genotypes for sequenced/imputed SNPs, \( \rho_{ij} \):**

<table>
<thead>
<tr>
<th>Sample sequenced</th>
<th>Correlation between the actual and called genotypes for sequenced/imputed SNPs, ( \rho_{ij} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.82, 0.86, 0.90, 0.95, 0.97</td>
</tr>
<tr>
<td></td>
<td>Same as S1</td>
</tr>
<tr>
<td></td>
<td>Same as S1</td>
</tr>
<tr>
<td></td>
<td>Same as S1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Call rates (1-missing data rate), \( \delta \):**

<table>
<thead>
<tr>
<th>Sample sequenced</th>
<th>Call rates (1-missing data rate), ( \delta )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4901 (1963 cases and 2938 controls of the WTCCC T1D study)</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>

**Simulation replicates for each configuration:**

<table>
<thead>
<tr>
<th>Sample sequenced</th>
<th>Simulation replicates for each configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>300</td>
</tr>
<tr>
<td></td>
<td>800</td>
</tr>
<tr>
<td></td>
<td>800</td>
</tr>
</tbody>
</table>

**Localization Success Rate:**

<table>
<thead>
<tr>
<th>Sample sequenced</th>
<th>Localization Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P(the causal SNP is top-ranked)</td>
</tr>
<tr>
<td></td>
<td>Same as S1</td>
</tr>
<tr>
<td></td>
<td>Same as S1</td>
</tr>
<tr>
<td></td>
<td>Defined for each of the 2 causal SNPs as P(the causal SNP ranks in top 2)</td>
</tr>
</tbody>
</table>
Table 4.3 Localization success rates for simulation Scenarios 1, 2, 3, 4

<table>
<thead>
<tr>
<th>Correlation between the tag and causal SNPs, ( r )</th>
<th>Sample size</th>
<th>0.82</th>
<th>0.86</th>
<th>0.90</th>
<th>0.95</th>
<th>0.97</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Naïve</td>
<td>Re-ranked</td>
<td>Naïve</td>
<td>Re-ranked</td>
<td>Naïve</td>
<td>Re-ranked</td>
</tr>
<tr>
<td>0.78</td>
<td>4901</td>
<td>0.20</td>
<td>0.38</td>
<td>0.21</td>
<td>0.42</td>
<td>0.23</td>
<td>0.42</td>
</tr>
<tr>
<td>0.83</td>
<td>4901</td>
<td>0.12</td>
<td>0.35</td>
<td>0.16</td>
<td>0.41</td>
<td>0.20</td>
<td>0.47</td>
</tr>
<tr>
<td>0.85</td>
<td>4901</td>
<td>0.20</td>
<td>0.39</td>
<td>0.23</td>
<td>0.43</td>
<td>0.26</td>
<td>0.51</td>
</tr>
<tr>
<td>0.90</td>
<td>4901</td>
<td>0.09</td>
<td>0.32</td>
<td>0.15</td>
<td>0.43</td>
<td>0.18</td>
<td>0.45</td>
</tr>
<tr>
<td>0.93</td>
<td>4901</td>
<td>0.08</td>
<td>0.41</td>
<td>0.12</td>
<td>0.35</td>
<td>0.19</td>
<td>0.40</td>
</tr>
<tr>
<td>0.95</td>
<td>4901</td>
<td>0.11</td>
<td>0.32</td>
<td>0.09</td>
<td>0.31</td>
<td>0.21</td>
<td>0.42</td>
</tr>
<tr>
<td>Tag is causal</td>
<td>4901</td>
<td>0.93</td>
<td>0.17</td>
<td>0.89</td>
<td>0.26</td>
<td>0.89</td>
<td>0.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 2a</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 3a</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 4a</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.80</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>0.95</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

See Table 4.2 for details of the simulation models; scenario 4 has two causal loci
1963 cases and 2938 controls for Scenario 1; equal number of cases and controls for Scenario 2,3,4
Naïve is standard ranking without correction for selection or genotyping error
Re-ranked is ranking by corrected statistic in Equation 1
In this simulation, the GWAS tag SNP is causal and all post-GWAS SNPs are non-causal
Table 4.4 Localization success rates for simulation Scenarios 5

<table>
<thead>
<tr>
<th>Correlation between the tag and causal SNPs, $r$</th>
<th>Sample</th>
<th>0.80</th>
<th>0.90</th>
<th>0.95</th>
<th>0.98</th>
<th>0.99</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>size$^b$</td>
<td>2500</td>
<td>0.17</td>
<td>0.20</td>
<td>0.26</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>5000</td>
<td>0.21</td>
<td>0.29</td>
<td>0.30</td>
<td>0.32</td>
<td>0.36</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>7500</td>
<td>0.18</td>
<td>0.34</td>
<td>0.33</td>
<td>0.43</td>
<td>0.46</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>0.18</td>
<td>0.38</td>
<td>0.37</td>
<td>0.42</td>
<td>0.48</td>
<td>0.53</td>
</tr>
</tbody>
</table>

- Call Rate (=1-Missing Data Rate)

<table>
<thead>
<tr>
<th>Call Rate (=1-Missing Data Rate)</th>
<th>0.95</th>
<th>0.98</th>
<th>0.99</th>
<th>1.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>0.33</td>
<td>0.45</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>0.36</td>
<td>0.50</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>0.46</td>
<td>0.61</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>0.53</td>
<td>0.66</td>
<td>0.77</td>
<td></td>
</tr>
</tbody>
</table>

---

$^a$ See Table 4.2 for details of the simulation models

$^b$ Equal number of cases and controls

$^c$ Naïve is standard ranking without correction for selection or genotyping error

$^d$ Re-ranked is ranking by corrected statistic in Equation 1
Table 4.5 Trends in power and localization success rate due to tagging and genotyping accuracy effect

<table>
<thead>
<tr>
<th>Tagging Power to Detect</th>
<th>Localization Success Rate</th>
<th>Tagging Power to Detect</th>
<th>Localization Success Rate</th>
<th>Tagging Power to Detect</th>
<th>Localization Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagging $r^3$</td>
<td></td>
<td></td>
<td>Tagging Power to Detect</td>
<td>Localization Success Rate</td>
<td></td>
</tr>
<tr>
<td>0.78</td>
<td>0.24</td>
<td>0.61</td>
<td>0.04</td>
<td>0.36</td>
<td>0.50</td>
</tr>
<tr>
<td>0.85</td>
<td>0.19</td>
<td>0.45</td>
<td>0.09</td>
<td>0.25</td>
<td>0.52</td>
</tr>
<tr>
<td>0.93</td>
<td>0.24</td>
<td>0.40</td>
<td>0.10</td>
<td>0.17</td>
<td>0.42</td>
</tr>
<tr>
<td>0.98</td>
<td>0.23</td>
<td>0.36</td>
<td>0.08</td>
<td>0.08</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Simulation parameters as describe in Table 4.2, Simulation 1 except as noted below

a All SNPs genotyped with perfect accuracy, regions selected by prior information
b GWAS SNPs genotyped perfectly, sequencing SNPs genotyped imperfectly (correlation between called and actual genotypes $\rho = 0.86$), regions selected by prior information
c GWAS SNPs genotyped perfectly, sequencing SNPs genotyped imperfectly (correlation between called and actual genotypes $\rho = 0.86$), regions selected by significant (p<5e-7) GWAS tag SNP
d Correlation between the tag and causal SNPs
e Probability causal SNP significant (p<5e-7)
f Probability causal SNP is most significant in region

In Table 4.5, we simulated data for 3 cases:

1. **Sequence Independent Sample, Genotyping 100% Accurate.** We simulated GWAS and sequencing data as described in Scenario 1 of the main text except that genotyping accuracy was 100% and fine-mapping was not conditional on genome-wide significance.

2. **Sequence Independent Sample, Genotyping 97% Accurate.** We simulate data as in (1) except that genotyping accuracy is 97%.

3. **Sequence GWAS Discovery Sample, Genotyping 97% Accurate** We simulate data as in (2) except that fine-mapping is conditional on genome-wide significance at the tag SNP.

The results show that when fine-mapping is performed in an independent sample with 100% accurate genotyping, the power to detect the original effect at the causal SNP as well as the localization success rate is high. When genotyping error is introduced, both power and localization success rate fall considerably. Re-using the original successful GWAS for the fine-mapping stage increases the
probability that the causal SNP will be genome-wide significant because it is conditional on genome-wide significance at the GWAS tag SNP. However, re-using the GWAS data considerably decreases the probability that the causal SNP will be top-ranked.
Table 4.6 Alternative Localization success rates for simulation Scenarios 2, 3, 4

<table>
<thead>
<tr>
<th>Correlation between the tag and causal SNPs, r</th>
<th>Sample size^b</th>
<th>Average correlation between the actual and estimated genotypes of sequencing SNPs, ρ_{si}</th>
<th>Low-coverage Sequencing</th>
<th>High-coverage Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.82</td>
<td>0.86</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>0.57</td>
<td>0.72</td>
<td>0.57</td>
<td>0.74</td>
</tr>
<tr>
<td>5000</td>
<td>0.55</td>
<td>0.85</td>
<td>0.63</td>
<td>0.89</td>
</tr>
<tr>
<td>7500</td>
<td>0.58</td>
<td>0.92</td>
<td>0.65</td>
<td>0.94</td>
</tr>
<tr>
<td>10000</td>
<td>0.60</td>
<td>0.94</td>
<td>0.66</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.95</td>
<td>0.52</td>
<td>0.70</td>
<td>0.59</td>
<td>0.72</td>
</tr>
<tr>
<td>5000</td>
<td>0.59</td>
<td>0.84</td>
<td>0.62</td>
<td>0.88</td>
</tr>
<tr>
<td>7500</td>
<td>0.59</td>
<td>0.93</td>
<td>0.67</td>
<td>0.95</td>
</tr>
<tr>
<td>10000</td>
<td>0.60</td>
<td>0.95</td>
<td>0.70</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.27</td>
<td>0.42</td>
<td>0.31</td>
<td>0.43</td>
</tr>
<tr>
<td>5000</td>
<td>0.30</td>
<td>0.43</td>
<td>0.33</td>
<td>0.46</td>
</tr>
<tr>
<td>7500</td>
<td>0.28</td>
<td>0.50</td>
<td>0.32</td>
<td>0.49</td>
</tr>
<tr>
<td>10000</td>
<td>0.29</td>
<td>0.50</td>
<td>0.32</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.93</td>
<td>0.29</td>
<td>0.41</td>
<td>0.32</td>
<td>0.40</td>
</tr>
<tr>
<td>5000</td>
<td>0.30</td>
<td>0.44</td>
<td>0.32</td>
<td>0.44</td>
</tr>
<tr>
<td>7500</td>
<td>0.30</td>
<td>0.51</td>
<td>0.34</td>
<td>0.53</td>
</tr>
<tr>
<td>10000</td>
<td>0.29</td>
<td>0.48</td>
<td>0.32</td>
<td>0.53</td>
</tr>
</tbody>
</table>

^a See Table 4.2 for details of the simulation models; scenario 4 has two causal loci
^b equal number of cases and controls for Scenario 2,3,4
^c Naïve is standard ranking without correction for selection or genotyping error
^d Re-ranked is ranking by corrected statistic in Equation 1
^e Localization success rate defined as probability causal SNP is in top 10% of SNPs ranked by test statistic, re-ranking statistic

Simulation Details

In some cases, investigators might be interested in selecting a set of candidate SNPs instead of a single top one. Here we define an alternative localization rate metric as the probability that the causal SNP is among the top 10% of the SNPs by rank. For this table, we simulated data in the same manner as the described in the main text for Scenarios 2-4, except that there are a total of 50 SNPs in the region: 1 tag, 1 causal and 48 non-causal sequencing SNPs.
4.8 Figures

Figure 4.1 Tagging effect decreases localization success rates with or without the selection effect. The expected values of the association test statistics at a tag SNP (red) and the causal SNP (black), shading from 25th – 75th percentiles (A, C), and the localization success rates (B, D) for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.12; OR=1.25; perfect genotyping accuracy) and one tag SNP (MAF=0.12; in varying degree of correlation with the causal SNP, $r=0.2$ to 1; perfect genotyping accuracy) with no selection for significance at the tag SNP (A, B) or selection at the tag SNP requiring the test statistic $T_G$ to be significant with p-value < 0.05 (C, D).
Figure 4.2 Low genotyping accuracy further reduces localization success rates with or without the selection effect. Localization success rates for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.12; OR=1.25; imperfect genotyping accuracy due to sequencing or imputation errors resulting in correlation between the actual and estimated genotypes $\rho_C = 0.80$ (blue dash-dotted) to 1 (black solid) and one tag SNP (MAF=0.12; in varying degree of correlation with the causal SNP, $r_{CG} = 0.2$ to 1 (X-axis); perfect genotyping accuracy with $\rho_G = 1$) with no selection for significance at the tag SNP (A) or selection at the tag SNP requiring the test statistic $T_G$ to be significant with p-value < 0.05 (B).
Figure 4.3 Well-tagged causal SNPs sequenced with low accuracy are unlikely to be correctly identified even as sample size increases. Localization success rates for association studies (sample size from 50:50 cases:controls to 5000:5000 cases:controls, X-axis) of one causal SNP (MAF=0.12; OR=1.25; imperfect genotyping accuracy due to sequencing or imputation errors resulting in correlation between the actual and estimated genotypes $\rho_C = 0.95$) and one tag SNP (MAF=0.12; in high correlation with the causal SNP, $r_{CG} = 0.8$ (purple solid) to 0.98 (red dashed); 100% genotyping accuracy with $\rho_G = 1$) with no selection for significance at the tag SNP.
Figure 4.4 Naïve test statistics and re-ranking statistics for regions surrounding rs78246868 in the 8q24.21 region for association with prostate cancer risk. Naïve test statistics (A), and re-ranking statistics adjusting for genotyping accuracy (B) for SNPs in LD ($r^2 > 0.2$) with rs78246868 (SNP with largest naive test statistic). Circles highlight SNPs whose rank changed considerably after re-ranking. Color indicates pair-wise correlation with the most significant SNP in the region selected based on the naïve ranking (purple diamond, rs78246868). Other shapes indicates genotyping accuracy over all 7 studies as measured by $\rho_{\text{meta}}$. rs78246868 remains the most significant SNP in the region after re-ranking.
Figure 4.5 Naïve test statistics and re-ranking statistics for regions surrounding rs8071558 in the 17q24.3 region for association with prostate cancer risk. Naïve test statistics (A), and re-ranking statistics adjusting for genotyping accuracy (B) for SNPs in LD (r^2 > 0.2) with rs8071558 (SNP with largest naïve test statistic). Circles highlight SNPs whose rank changed considerably after re-ranking. Color indicates pair-wise correlation with the most significant SNP in the region selected based on the naïve ranking (purple diamond, rs8071558). Other shape indicates genotyping accuracy over all 7 studies as measured by ρ_{meta}, rs8071558 is the most significant SNP in the region after re-ranking.
4.9 Appendices

4.9.1 Appendix 4 Derivation of Distribution of GWAS and Tag SNP Test Statistics

Let $S_1, \ldots, S_k$ be the sequencing SNPs in the disease-associated region, one of which is the causal SNP, denoted as $C$. The GWAS tag SNP is denoted as $G$. Let $T_G$ be the test statistic at the tag SNP, and $T_{S_1}, T_{S_2}, \ldots, T_{S_k}$ be the vector of test statistics for the sequencing SNPs. We define the pairwise correlations as follows: $\rho_{CS_i}$ is the pairwise correlation between the causal SNP and sequencing SNP $i$, $\rho_{GS_i}$ is the pairwise correlation between the GWAS tag SNP and sequencing SNP $i$, $\rho_{S_iS_j}$ is the pairwise correlation between sequencing SNPs $i$ and $j$, $\rho_{CG}$ is the pairwise correlation between the causal SNP and the GWAS tag SNP ($\rho_{CG} > 0$ without loss of generality). One of the sequencing SNPs is causal, so for one of $i=1 \ldots k$, $T_{S_i} = T_C$, $\rho_{CS_i} = 1$ and $\rho_{GS_i} = \rho_{CG}$. The multivariate normal (MVN) joint distribution of the test statistics at the tag and sequencing SNPs is

$$
\begin{pmatrix}
    T_G \\
    T_{S_1} \\
    T_{S_2} \\
    \vdots \\
    T_{S_k}
\end{pmatrix}
\sim N
\begin{pmatrix}
\sqrt{n} \mu_C \\
\rho_{CG} \\
\rho_{CS_1} \\
\rho_{GS_1} \\
\rho_{CS_2} \\
\rho_{GS_2} \\
\vdots \\
\rho_{CS_k} \\
\rho_{GS_k}
\end{pmatrix}
\begin{pmatrix}
1 & \rho_{GS_1} & \rho_{GS_2} & \cdots & \rho_{GS_k} \\
\rho_{CG} & 1 & \rho_{S_1S_2} & \cdots & \rho_{S_1S_k} \\
\rho_{CS_1} & \rho_{GS_1} & 1 & \cdots & \rho_{S_1S_k} \\
\rho_{GS_2} & \rho_{S_1S_2} & \cdots & \cdots & \rho_{S_2S_k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho_{CS_k} & \rho_{S_1S_k} & \rho_{S_2S_k} & \cdots & 1
\end{pmatrix}.
$$

Eqn A

Note that although the GWAS tag SNP is first in the vector, the physical location of the tag SNP is likely to be in the middle of the sequencing SNPs. In Equation A, we assume no genotyping error and no prior selection of the region for a significant tag SNP association.

The expected value of the test statistic at a particular SNP is proportional to its correlation with the causal SNP. The expected value is largest when correlation with the causal SNP is 1, i.e. the SNP is causal or in
perfect LD with the causal SNP.

**Tagging Effect**

Selection of significant regions via GWAS tag SNPs induces upward bias in the magnitude of observed tag SNP associations. When GWAS samples are sequenced, the bias at the tag carries through to sequencing SNPs in high LD with the tag. This effect attenuates as correlation between the tag and sequencing SNPs breaks down [20].

Conditional on the value of the observed test statistic $T_{G_{\text{obs}}}$ at the tag SNP, the joint distribution of the test statistics at the sequencing SNPs is

$$
\begin{pmatrix}
T_{S_1} \\
T_{S_2} \\
\vdots \\
T_{S_k}
\end{pmatrix}
| T_G = T_{G_{\text{obs}}}
\sim N\left(\sqrt{n} \mu_C \left(\begin{array}{c}
\rho_{GS_1} \\
\rho_{GS_2} \\
\vdots \\
\rho_{GS_k}
\end{array}\right) + \left(\begin{array}{c}
\rho_{GS_1} \\
\rho_{GS_2} \\
\vdots \\
\rho_{GS_k}
\end{array}\right)(T_{G_{\text{obs}}} - \rho_{GC} \mu_C \sqrt{n}), \Sigma\right)
$$

Eqn B

where the elements of $\Sigma$ are $\Sigma_{ij} = \rho_{S_iS_j} - \rho_{GS_i} \rho_{GS_j}$, $i = 1..k, j = 1..k$.

We assume, without loss of generality that the tag SNP genotype is coded so that $T_{G_{\text{obs}}} > 0$. If we select for a large tag SNP test statistic, then the expected value of $T_{G_{\text{obs}}} - \rho_{GC} \mu_C \sqrt{n} > 0$. We call $T_{G_{\text{obs}}} - \rho_{GC} \mu_C \sqrt{n}$ the tagging bias term. The tagging bias term carries through to the sequencing SNPs in proportion to the correlation between the tag and sequencing SNPs, $\rho_{GS_k}$. The combination of distortion at the tag and attenuation due to LD distorts the signal (i.e. the expected value of the test statistic at each SNP) so that SNPs in high LD with the tag are more likely to be top-ranked, even when they are non-causal. We call this phenomenon the tagging effect. The re-ranking procedure first
estimates the effect of selection on the tag SNP test statistic \( T_{G_{\text{obs}}} - T_{G_{\text{mut}}} \), then estimates the degree to which inflation at the tag transfers to the sequencing SNP \( \rho_{TS} \frac{\delta_{GS}}{\sqrt{\delta_C \delta_S}} \) and adjusts each test statistic accordingly. The quantity \( \frac{\delta_{GS}}{\sqrt{\delta_C \delta_S}} \) adjusts the correlation for missing data (see below for details).

The tagging effect is most pronounced when there is low power at the tag SNP. For example, consider the case where we only examine the region if the tag SNP test statistic is significant at level \( \alpha \). Then, using Inverse Mills Ratio, the expected value of the test statistic is

\[
Q = E[T_{G_{\text{obs}}} \mid T_{G_{\text{obs}}} > Z_\alpha] = \frac{\phi(Z_\alpha - \sqrt{n \mu_C} \rho_{GC})}{\Phi(-Z_\alpha - \sqrt{n \mu_C} \rho_{GC})}.
\]  

Eqn C

This quantity increases when the critical value \( Z_\alpha \) is extreme compared to the expected value of the tag SNP test statistic \( \sqrt{n \mu_C} \rho_{GC} \), i.e. when there is low power to detect the tag SNP. The vector of expected values at each test statistic is,

\[
\sqrt{n \mu_C} \begin{pmatrix} \rho_{CS1} \\ \rho_{CS2} \\ \vdots \\ \rho_{CSk} \end{pmatrix} + \begin{pmatrix} \rho_{GS1} \\ \rho_{GS2} \\ \vdots \\ \rho_{GSk} \end{pmatrix} Q. 
\]  

Eqn D

This vector is composed of two terms. The first term is the genetic effect, which maximizes at the causal SNP. The second term is the tagging effect which maximizes at the tag SNP and attenuates as correlation between the tag and sequencing SNPs breaks down. In addition, the second term increases for all SNPs as power at the tag SNP decreases. This second term corresponds to increased probability that SNPs that are in highest LD with the tag achieve the top ranks, regardless of whether they are causal or non-causal.
Genotyping Accuracy Effect

Here we show how genotyping error rates and call rates that vary among SNPs can change the joint distribution of the test statistics so that SNPs with higher error rates are less likely to be top-ranked, regardless of whether they are causal or not. Genotyping algorithms make a trade off between call rates and accuracy by applying quality thresholds [63]. One may reduce a SNP’s error rate by filtering out individual calls not meeting stringent quality thresholds; however this increases the missing data rate for that SNP. Since both low accuracy and low call rates lower a SNP’s probability of being top-ranked, filtering out low quality calls will not correct for the distortion in the ranking of test statistics. We refer to the combined effect of genotyping error and excluded data as the genotyping accuracy effect.

Here we describe the distribution of the test statistic with genotyping error and missing data. Where $0 < \rho_M < 1$ is the correlation between the true genotype and the observed genotype for sequenced (or imputed) SNP $i$, and $\delta_i$ is the call rate (proportion of non-missing genotypes), the distribution of the test statistic using the observed genotypes is

$$T_{S_i} \sim N\left(\rho_M \sqrt{\delta_i n_C \rho_{CS_i}}, 1\right).$$

Eqn E

The smaller $\rho_M$ and $\delta_i$ are, the lower the probability that SNP $S_i$ will be top-ranked, regardless of whether it is causal or non-causal.

4.9.2 Appendix 5  Tagging and Coverage of GWAS Platforms

In general, both tagging as well as imputation quality can be lower than investigators expect for the following three reasons.

(1) Tagging $r^2$ is often lower than expected. The coverage of GWAS platforms reported by manufacturers tends to be inflated. Manufacturers use the HapMap samples both to design
GWAS platforms and to estimate coverage, which results in overestimated coverage rates [64]. Numerous studies have shown that coverage of common variation is substantially lower than reported by microarray manufacturers [65]; this is especially true for non-European samples [66].

(2) Imputation accuracy is not lower-bounded by tagging r². A SNP may be very well tagged, but imputation accuracy depends on how well the model that is derived from the reference panel matches the GWAS sample. For example, if an imputed SNP is perfectly correlated with a tag SNP in the GWAS sample but not in the reference panel, then imputation r² can be less than 1. Jiang et al [64] tested the accuracy of imputation in Chinese samples and found that despite Illumina’s claim that the OmniExpress chip covers 91% of common SNPs (MAF > 5%) at r² > 0.8 in Chinese populations, only 73% of common SNPs were imputed at r² > 0.8.

(3) Sequencing/imputation SNPs are best captured by tag SNPs with similar MAFs, as correlation is bounded when MAFs are unequal. In our simulations, we assume that the tag and sequencing SNPs have similar MAF so that correlation can be allowed to vary from 0.78 to 0.98.
4.9.3 Appendix 6  Additional Figures

Figure 4.6 Distribution of SNP-specific read depth using the 1000 Genomes low-coverage pilot data on 351,456 SNPs from chromosome 1 in the CHB and JPT samples (October 2010 release; www.1000genomes.org/data).
Figure 4.7 Tagging effect decreases localization success rates with or without the selection effect, rare SNP. The expected values of the association test statistics at a tag SNP (red) and the causal SNP (black), shading from 25th – 75th percentiles (A,C), and the localization success rates (B, D) for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.02; OR=1.5; perfect genotyping accuracy) and one tag SNP (MAF=0.02; in varying degree of correlation with the causal SNP, $r = 0.2$ to 1; perfect genotyping accuracy) with no selection for significance at the tag SNP (A, B) or selection at the tag SNP requiring the test statistic $T_G$ to be significant with p-value < 0.05 (C, D).
Figure 4.8 Tagging effect decreases localization success rates with or without the selection effect, high frequency SNP. The expected values of the association test statistics at a tag SNP (red) and the causal SNP (black), shading from 25th – 75th percentiles (A,C), and the localization success rates (B, D) for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.25; OR=1.25; perfect genotyping accuracy) and one tag SNP (MAF=0.25; in varying degree of correlation with the causal SNP, $r$= 0.2 to 1; perfect genotyping accuracy) with no selection for significance at the tag SNP (A, B) or selection at the tag SNP requiring the test statistic $T_G$ to be significant with p-value < 0.05 (C, D).
Figure 4.9 Genotyping accuracy effect further reduces localization success rates with or without the selection effect, rare SNP. Localization success rates for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.02; OR=1.5; imperfect genotyping accuracy due to genotyping, sequencing or imputation errors resulting in correlation between the actual and estimated genotypes $\rho_G = 0.80$ (blue dash-dotted) to 1 (black solid)) and one tag SNP (MAF=0.12; in varying degree of correlation with the causal SNP, $r_{CG} = 0.2$ to 1 (horizontal axis); perfect genotyping accuracy with $\rho_G = 1$) with no selection for significance at the tag SNP (A) or selection at the tag SNP requiring the test statistic $T_G$ to be significant with p-value < 0.05 (B).
Figure 4.10 Genotyping accuracy effect further reduces localization success rates with or without the selection effect, high frequency SNP. Localization success rates for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.25; OR=1.25; imperfect genotyping accuracy due to genotyping, sequencing or imputation errors resulting in correlation between the actual and estimated genotypes $\rho_C = 0.80$ (blue dash-dotted) to 1 (black solid)) and one tag SNP (MAF=0.12; in varying degree of correlation with the causal SNP, $r_{CG} = 0.2$ to 1 (horizontal axis); perfect genotyping accuracy with $\rho_G = 1$) with no selection for significance at the tag SNP (A) or selection at the tag SNP requiring the test statistic $T_G$ to be significant with p-value $< 0.05$ (B).
Figure 4.11 Well-tagged causal SNPs sequenced with low accuracy are unlikely to be correctly identified even as sample size increases, rare SNP. Localization success rates for association studies (50:50 cases:controls to 5000:5000 cases:controls, horizontal axis) of one causal SNP (MAF=0.02; OR=1.5; imperfect genotyping accuracy due to genotyping, sequencing or imputation errors resulting in correlation between the actual and estimated genotypes $\rho_C = 0.95$) and one tag SNP (MAF=0.12; in high correlation with the causal SNP, $r_{CG} = 0.8$ (purple solid) to 0.98 (red dashed); 100% genotyping accuracy with $\rho_G = 1$) with no selection for significance at the tag SNP.
Figure 4.12 Well-tagged causal SNPs sequenced with low accuracy are unlikely to be correctly identified even as sample size increases, high frequency SNP. Localization success rates for association studies (50:50 cases:controls to 5000:5000 cases:controls, horizontal axis) of one causal SNP (MAF=0.25; OR=1.25; imperfect genotyping accuracy due to genotyping, sequencing or imputation errors resulting in correlation between the actual and estimated genotypes $\rho_C = 0.95$) and one tag SNP (MAF=0.12; in high correlation with the causal SNP, $r_{CG} = 0.8$ (purple solid) to 0.98 (red dashed); 100% genotyping accuracy with $\rho_G = 1$) with no selection for significance at the tag SNP.
Figure 4.13 Tagging effect decreases localization success rates with or without the selection effect, 3 SNPs: 1 tag, 1 causal, 1 non-causal sequencing SNP. The expected values of the association test statistics at a tag SNP (red), the causal SNP (black), a non-causal sequencing SNP (green), shading from 25\textsuperscript{th} – 75\textsuperscript{th} percentiles (A,C), and the localization success rates (B, D) for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.02; correlation between causal and non-causal sequencing SNPs = 0.90, OR=1.5; perfect genotyping accuracy) and one tag SNP (MAF=0.12; in varying degree of correlation with the causal SNP, \( r \) = 0.2 to 1; perfect genotyping accuracy) with no selection for significance at the tag SNP (A, B) or selection at the tag SNP requiring the test statistic \( T_G \) to be significant with p-value < 0.05 (C, D).
Figure 4.14 Tagging effect decreases localization success rates with or without the selection effect 5 SNPs: 1 tag, 1 causal, 3 non-causal sequencing SNPs. The expected values of the association test statistics at a tag SNP (red), the causal SNP (black) and the maximum test statistic of the 3 non-causal sequencing SNPs (green), shading from 25th – 75th percentiles (A,C), and the localization success rates (B, D) for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.02; OR=1.5; perfect genotyping accuracy) and one tag SNP (MAF=0.12; correlation between causal and non-causal sequencing SNPs = 0.90, in varying degree of correlation with the causal SNP, r = 0.2 to 1; perfect genotyping accuracy) with no selection for significance at the tag SNP (A, B) or selection at the tag SNP requiring the test statistic $T_G$ to be significant with p-value < 0.05 (C, D).
Figure 4.15 Tagging effect decreases localization success rates with or without the selection effect 7 SNPs: 1 tag, 1 causal, 5 non-causal sequencing SNPs. The expected values of the association test statistics at a tag SNP (red), the causal SNP (black) and the maximum test statistic of the 3 non-causal sequencing SNPs (green), shading from 25th – 75th percentiles (A,C), and the localization success rates (B, D) for association studies (1000 cases and 1000 controls) of one causal SNP (MAF=0.02; OR=1.5; perfect genotyping accuracy) and one tag SNP (MAF=0.12; correlation between causal and non-causal sequencing SNPs = 0.90, in varying degree of correlation with the causal SNP, \( r = 0.2 \) to 1; perfect genotyping accuracy) with no selection for significance at the tag SNP (A, B) or selection at the tag SNP requiring the test statistic \( T_G \) to be significant with p-value < 0.05 (C, D).
Figure 4.16. Location of five statistically independent associated sub-regions at 8q24.21 locus for the BPC3 aggressive prostate cancer GWAS. Location, index SNP (SNP with largest naive test statistic) and zoom in plot figure number is indicated for each sub-region.
Figure 4.17 Naïve test statistics and re-ranking statistics for regions surrounding rs1016343 on 8q24.21 and rs34255287 on 11q13.3 for association with prostate cancer risk. Naïve test statistics (A, B), and re-ranking statistics adjusting for genotyping accuracy (C, D) for SNPs in LD ($r^2 > 0.2$) with rs1016343 (SNP with largest naive test statistic) (A, C) or rs34255287 (SNP with largest naive test statistic) (B, D). Circles highlight SNPs whose rank changed considerably after re-ranking. Color indicates pair-wise correlation with the most significant SNP in the region. Shape indicates genotyping accuracy over all 7 cohorts as measured by $\rho_{META}$, purple diamond is index SNP (most significant SNP from naïve meta-analysis rs1016343 for A and C and rs34255287 for B and D).
Figure 4.18 Naïve test statistics and re-ranking statistics for regions surrounding rs7816007 and rs6983267 on 8q24.21 for association with prostate cancer risk. Naïve test statistics (A, B), and re-ranking statistics adjusting for genotyping accuracy (C, D) for SNPs in LD ($r^2 > 0.2$) with rs7816007 (SNP with largest naive test statistic) (A, C) or rs6983267 (SNP with largest naive test statistic) (B, D). Color indicates pair-wise correlation with the most significant SNP in the region. Shape indicates genotyping accuracy over all 7 cohorts as measured by $\rho_{\text{META}}$, purple diamond is index SNP (most significant SNP from naive meta-analysis, rs7816007 for A and C, rs6983267 for B and D).
Figure 4.19 Naïve test statistics and re-ranking statistics for regions surrounding rs382434 on 8q24.21 for association with prostate cancer risk. Naïve test statistics (A), and re-ranking statistics adjusting for genotyping accuracy (B) for SNPs in LD ($r^2 > 0.2$) with rs382434 (SNP with largest naive test statistic, rs382434). Circles highlight SNPs whose rank changed considerably after re-ranking. Color indicates pair-wise correlation with the most significant SNP in the region. Shape indicates genotyping accuracy over all 7 cohorts as measured by $\rho_{META}$, purple diamond is index SNP (most significant SNP from naive meta-analysis).
4.10 References


13. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-78


37. Sinnott JA and Kraft P. (2012) Artifact due to differential error when cases and controls are imputed from different platforms. Human Genetic 131:111-119


Chapter 5. Overall Discussion

In this dissertation, I have explored two phenomena that can induce bias in effect estimates and reduce power to localize disease-causing variants: selection for significant GWAS tag SNPs and differential genotyping error between next generation sequencing (NGS) SNPs. I developed novel analytical methods that recover much of this lost accuracy and power. The difference these methods make to the interpretation of evidence in real studies has been illustrated by application to datasets from actual genetic association studies. As quantified by our simulation studies, these methods provide substantial improvement over the standard analysis.

When the causal SNP is detected through imperfect LD with a GWAS tag SNP, the effect at the causal SNP can be estimated either directly through imputed or sequenced SNPs, or indirectly through the GWAS tag SNP genotypes. The first scenario is explored in the third chapter of this thesis. The second scenario is explored by Spencer et al [1]. The Spencer paper shows that while the winner’s curse induces upward bias in effect estimates, imperfect LD between the tag and causal SNPs induces downward bias. The interplay between these two factors can leave the effect estimate either upward biased, downward biased or unbiased (i.e. the two effects exactly oppose each other) for the true causal SNP effect. The direction and magnitude of bias depends on the underlying effect size, the significance threshold, the MAFs at the tag and causal SNPs, and the tag SNP effect sizes. For this reason, it is important to understand the limits of using a tag SNP to estimate the effect at an imperfectly captured causal SNP.
5.1 Performance and computational efficiency

Sun et al [2] and Poirier et al [3] further explored the performance of the bootstrap bias-reduction method. The former apply the bootstrap method to a psoriasis case-control GWAS and a diabetes-related quantitative trait, and find that the bootstrap estimates tend to be closer to the replication values than the naive estimates. Poirier et al use simulation to explore performance of the bootstrap in the case of a time-to-event phenotype. Poirier finds that, in general, the bias-reduced estimates tend to be closer than the naive to the true underlying effect size, and that they tend to be somewhat conservative especially at high power. While the work in Chapter 2 found that the bias-reduced estimates tended to be slightly conservative at high power, this is also true in the time-to-event setting.

While the BR-Squared software is computationally efficient, researchers may still view applying a bootstrap method as onerous. For researchers applying the method to a single dataset, BR-Squared software requires very little time to run. However, it does require that researchers have the original genotype and phenotype information for each sample. For researchers conducting large simulation studies, the computational time required may appear prohibitive. For example, Baca\-nu and Kendler [4] propose an estimation method that captures information from suggestive signals (nearly significant test statistics) that includes bias-correction. The authors explain that they had wanted to compare their method with BR-squared, as the method performs very well. However they chose to simulate test statistics instead of GWAS genotypes and phenotypes. While this type of simulation is simpler to program, BR-squared software requires the GWAS genotype and phenotype datasets and so they did not include BR-squared in their comparison. Ferguson et al [5] proposed a new Empirical Bayes
bias-reduction method which they compare with several other methods. While they include BR-squared in their application, they leave it out of the simulation study without providing a reason. While simulating GWAS data can appear to be daunting, software packages that simulate genotype and phenotype data are available. Recently, the National Cancer Institute released a catalogue of genetic simulation software packages to help investigators identify available packages that can generate data for their simulation studies [6]. Several of the packages listed in January 2012 appear to be able to simulate genotype and phenotype data suitable for analysis with BR-squared. By reducing the programming burden associated with simulating realistic genotype and phenotype datasets, these resources can help investigators to easily compare their methods with BR-squared. Further improvements to computational efficiency could also help remove perceived barriers to the adoption of BR-squared software. Sambo and Camillo [7] discuss computational intensity of bootstrap-based methods, including BR-Squared software. They propose two levels of optimization that exploit the repeated subjects in each bootstrap and repeated sets of subjects across bootstraps in order to eliminate duplicated computations. High-performance computing clusters can also reduce the time required to run simulations by allowing replicates to be run in parallel, and these clusters are available at most research institutions.

5.2 Replication study design

Replication is of particular concern with genetic association studies. Non-replication is a familiar problem for GWAS [8]. Many high-profile journals require replication in an independent sample in order to publish a significant GWAS finding [9]. Replication study sample sizes calculated using inflated effect estimates tend to be underpowered. Investigators may assume non-replicating associations are false positives; however they may in fact be true
positive associations that simply failed to replicate. Inaccurate localization can also result in failed replication because of reduced power at the incorrectly identified “causal” SNP. As demonstrated in our third paper, SNPs with the greatest selection bias are more likely to be top-ranked, even if they are not causal. This means that the SNPs with the largest bias are more likely to be used to compute the replication study sample size, which leads to low power. Selecting the wrong SNP(s) for follow up can also result in non-replication. If only a certain number of sequenced or imputed SNPs can be selected for replication (e.g. via a custom microarray), then bias due to selection or genotyping accuracy can cause the correct causal SNP to be left out. Finally, replication samples are often drawn from a different population in order to demonstrate the original association was not due to systematic bias [10]. If the LD structure in the replication population differs from the original population, the non-causal tag SNPs may not be in high enough correlation with the true causal SNPs for the replication to reach statistical significance.

5.3 Complex criteria for association

Complex criteria are often used in association testing and ranking procedures. Complex criteria can help to boost the power to detect an association, for example the maximum of several test statistics [11,12] or Bayes optimal ranks [13]. In order to compute bias-reduced estimates for SNPs selected via complex criteria, BR-squared software would need to be extended. Extending the bootstrap would simply require adding the complex criteria to the bootstrap selection step. The corresponding extension for the likelihood method would require a distribution for the genetic effect parameter conditional on the complex criteria, which may or may not be tractable. One type of complex ranking is currently implemented in BR-Squared
software: ranking by minimum p-value from the trend and genotype tests. In order to apply the re-ranking method, the correlation between these complex test statistics for each SNP would need to be derived or accurately approximated.

Tests that minimize loss of power due to genotyping uncertainty have been proposed [14,15]. However, power lost due to genotyping error cannot be completely recovered, and so it is still important to adjust the ranking for the genotyping accuracy effect. The simplest way to incorporate uncertainty into a test is to use the expected genotype count (i.e. the posterior mean) in the standard logistic or linear regression. In this case, the re-ranking method can be applied using the allele dosages in place of called genotypes and the correlations between the tag SNP genotype and the sequencing SNPs allele dosage genotypes. Guan et al [16] compare several frequentist and Bayesian methods that incorporate genotype uncertainty into the test for association. In these cases in which SNPs are selected by more complex methods, the re-ranking procedure can be extended to any case where the correlation between test statistics or Bayes factors can be worked out.

5.4 Common study design factors exacerbate bias

Algorithms originally designed for imputation have recently been leveraged to improve the accuracy of genotype calling for very low coverage sequencing data [17]. Imputation algorithms greatly improve accuracy for well-tagged SNPs, but improvements are modest for SNPs that are not well-tagged. As shown in chapter 4, SNPs that suffer from poor tagging and low accuracy tend to be down-ranked relative to well-tagged, well-called SNPs. When the SNPs that are poorly tagged are also poorly genotyped, then the loss of power to localize causal SNPs described in chapter 4 would be exacerbated. So long as the imputation algorithm produces
accurate estimates of the correlation between actual and imputed SNPs, the re-ranking method should help recover lost power.

5.5 Low MAF and rare SNPs

Low MAF SNPs (1-5%) and rare SNPs (MAF < 1%) have received much recent attention for several reasons: common causal SNPs (MAF > 5%) identified by GWAS explain only a small proportion of heritability for complex traits [20]; theories of quantitative variation predict that many causal variants will have lower MAFs [21, 22]; and falling costs have made the sample sizes and sequencing coverage required to accurately sequence rare/low-MAF SNPs economically feasible. The bias-reduction and re-ranking methods can be expected to be particularly important for low MAF SNPs. Low-MAF causal SNPs tend to suffer from both low power and high genotyping error, and so they tend to suffer from high selection bias and low power to localize. Both the bias-reduction and re-ranking methods should be particularly helpful when the causal SNP is uncommon. The methods assume allele counts large enough for test statistics to reach asymptotic normality. Where sample sizes are large enough for rare variants to meet this requirement, the bias-reduction and re-ranking method can include rare variants. SNPs with very low allele counts (e.g. count < 5) should be considered for exclusion from the re-ranking method. The simulations in Chapter 4 focus on SNPs with MAF of 5% and a single LD pattern over the associated locus. In order to better characterize power to identify candidate causal SNPs future work could include a simulation study covering a wider range of MAFs and LD patterns. An alternative measure of linkage disequilibrium is D', which may be of interest to geneticists due to its close relationship with the recombination rate between two SNPs [23].
Recently, collapsing and aggregating methods have been developed for rare variants. The effect estimate for the collapsing variable may be complicated to interpret, but could still be used to estimate the sample size required for a replication study. In this case, the bootstrap bias-reduction method could be extended to give a more accurate estimate of sample size required for an appropriately powered sample. As power to detect rare variants tends to be very low, selection bias could be substantial and the bootstrap bias-reduction method has the potential to greatly reduce this bias. Liu and Leal [24] have extended the original Sun and Bull [25] method to construct a bias-reduced estimator for the average genetic effect of rare variants included in the aggregate rare variant test. In keeping with my work, they found that the estimate of genetic effect for the rare variant aggregate statistics suffers from upward bias due to selection. However, the rare variants in the aggregate statistic will be different in each random sample drawn from the population, and so the corresponding genetic effect will not necessarily be comparable across studies. Liu and Leal do not implement any procedure to correct for correlation between the within- and out-of-samples, which may provide further improvement. Considering differing directions of effect at each rare SNP and expanding the type of aggregation method to a more general framework could also be useful. The biological significance of the genetic effect parameter is difficult to define for aggregation methods, and so further exploration of the most relevant scale for estimation may be of value to the genetic epidemiologists who will use the method.

5.6 Continuing importance of selection and genotyping accuracy effects

Low coverage sequencing, and the associated challenges described in this thesis, will continue to play an important role in genetic association studies. Custom fine-mapping
genotyping arrays such as the Immunochip for immune-related disease and Metabochip for metabolic diseases contain a dense set of SNPs for fine mapping. However, these arrays are less informative than sequencing and not optimized for non-European samples [26]. Fine-mapping arrays can be complemented by sequencing and/or imputation in order to obtain base-pair level information [e.g 27]. Whole genome sequencing is approximately twice the cost of exome-sequencing [26] and so large sample whole-genome sequencing will likely remain feasible only at low coverage. As the number of undiscovered large effect size SNPs dwindles, researchers will turn their attention increasingly to causal SNPs with smaller effect sizes, which tend to suffer from higher tagging bias due to low power. Association study sample sizes will therefore need to continue to increase, so even as sequencing costs fall, it is anticipated that low coverage will continue to be a cost-effective design for many studies [28]. Given a fixed budget, there is a trade off between read depth and sample size depending on the study objectives. Sampson et al [29] find that power to detect associations is highest when sample size is maximized, even as read depth falls to 1x-2x. Re-ranking to adjust for bias would be especially crucial for such studies as the error rates would be both very high and highly variable among SNPs. Re-ranking can adjust for a moderate amount of error, but if data quality is truly poor, e.g. in regions that are extremely difficult to sequence, re-ranking should not be applied.

5.7 Extensions

This work could be extended in several directions. The bootstrap procedure gives bias-reduced effect estimates for single SNPs in single-stage GWAS study designs. GWAS studies are often followed by sequencing or imputation, and effects at sequenced or imputed SNPs may only be reported if the test statistic reaches statistical significance at either a nearby tag SNP, or
the sequenced or imputed SNP, or both. The bootstrap could be easily extended to provide bias-reduced estimates for SNPs not found on GWAS platforms (e.g. sequenced or imputed SNPs). Successful GWAS are often followed by replication studies in which significant SNPs are tested in a new sample, the replication significance threshold being less stringent than the original due to the reduced multiplicity. In some cases only the significant SNPs are genotyped in the replication phase (e.g. Todd et al [30]), in other cases both the initial and replication phases include a full GWAS panel even though only significant SNPs from the original analysis are tested in the replication stage (e.g. Okada et al [31]). Conditional on successful replication, the SNP’s effect estimates would still suffer from bias because replicated SNPs are only reported if significance crosses the replication study threshold (typically p < 0.05). In some cases a joint analysis is used to increase power to detect an effect [32] and the precision of the effect estimate [33]. In that case, the degree of selection bias would depend on selection criteria at the first and second stage. If the replication sample is genotyped using a full GWAS panel but only SNPs crossing the replication study significance threshold are reported, then the genome-wide bootstrap method, described in chapter 1, could be extended to mimic both stages of selection and provide a bias-reduced joint effect estimate. If the replication samples are only genotyped at a small subset of SNPs, then the single-SNP bootstrap, described in Sun and Bull [25] could be extended instead. The re-ranking method could also be extended, provided that correlation between the combined test statistics at different SNPs could be worked out or approximated. When several GWAS are combined for meta-analysis, different significance criteria for each GWAS can result in a complex selection effect. This is even more complex when a meta-analysis builds on a previous meta-analysis by adding new GWAS data.
The methods presented in this dissertation provide efficient procedures to improve localization and estimation of effect sizes in genetic association studies. Work by other researchers has confirmed performance and efficiency of the bootstrap bias-reduction method. As uptake of next generation sequencing continues, and researchers drill down into SNPs with low MAFs and small effect sizes, I expect these methods to be particularly important.

5.8 References for Discussion


7. Sambo F and Di Camillo F (2012) "Minimizing time when applying bootstrap to
contingency tables analysis of genome-wide data" Learning and Intelligent Optimization
Papers pp 175-189

8. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA and
Hirschhorn JN (2008) Genome-wide association studies for complex traits: consensus,


studies Statistical Science 24:561-573

of true associations in genome-wide case-control association studies. BMC Proceedings
1:S165

for case-control association studies. 123:617-623

13. Louis TA and Ruczinski I (2010) Efficient evaluation of ranking procedures when the
number of units is large, with application to snp identification. Biometrical Journal 52:34-49

with missing genotype data. American Journal of Human Genetics 82:444-452

method for genome-wide association studies by imputation of genotypes. Nature
Genetics 39(7):906-13


genome-wide association scans with replication studies. Genetic Epidemiology 33:406–418
6 References, Alphabetical Order


Bazerman, MH and Samuelson WF (1983) I won the auction but don't want the prize. J Conflict Resolut 27:618-634


Browning BL and Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide
association studies. American Journal of Human Genetics 85:847-861


Faye L, Sun L, Dimitromanolakis A and Bull SB (2009) A comprehensive look at the likelihood and bootstrap approaches to overcome the winner’s curse in GWAS. (Abstract 126) Genetic Epidemiology 33:788-789.


Fridley BL, Abo R, Brisbin A and Jenkins GD (2011) Sample selection study designs to
follow-up GWAS signals with targeted sequencing. Genetic Epidemiology 36:148

Garner C (2007) Upward bias in odds ratio estimates from genome-wide association
studies. Genetic Epidemiology 31:288–295

Genetic Epidemiology 35:261-268.

Georges, M (2011) The long and winding road from correlation to causation. Nature
Genetics 43:180-181

approximate conditional likelihood approach. The American Journal of Human Genetics
82:1064–1074

Göring HH, Terwilliger JD and Blangero J (2001) Large upward bias in estimation of
locus-specific effects from genome-wide scans. The American Journal of Human
Genetics 69:1357-1369

mapping. PLoS Genetics 4(12):e1000279

(2007) Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2
protects against type 2 diabetes. Nature Genetics 39:977-983

Hao K and Wang X (2004) Incorporating individual error rate into association test of
unmatched case-control design. Human Heredity 58:154-63

Heckman JJ (1979) Sample selection bias as a specification error. Econometrica 47:153-161


Li Q, Yu K, Li Z and Gang Z (2008) MAX-rank: a simple and robust genome-wide scan
for case-control association studies. 123:617-623


Louis TA and Ruczinski I (2010) Efficient Evaluation of ranking procedures when the number of units is large, with application to SNP identification. Biometrical Journal 52:34-49


Paterson AD, Waggott D, Boright AP, Hosseini SM, Shen E et al. (2010) A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both HbA1c and glucose. Diabetes 59:539–549

Poirier J, Faye LL, Dimitromanolakis A, Paterson AD, Sun L, and Bull SB (2013) A general procedure to address the winner’s curse in genetic association studies: bias-reduction in analysis of time-to-event traits. (under review, to update when accepted)


Pritchard JK (2001) Are rare variants responsible for susceptibility to complex disease?
American Journal of Human Genetics 69:124-137

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81:559–575


Sinnott JA and Kraft P. (2012) Artifact due to differential error when cases and controls are imputed from different platforms. Human Genetic 131:111-119


11q23 and replicates risk loci at 8q24 and 18q21. Nature Genetics 40:631-637


Udler MS, Tyrer J and Easton DF (2010) Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. Genetic Epidemiology 34:463-468


Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-78

Wiltshire S, Morris AP and Zeggini E (2008) Examining the statistical properties of fine-scale mapping in large-scale association studies. Genetic Epidemiology 32:204-14

WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678


