Title
Efficient Design and Analysis of Genome-wide Association Studies

Author
KOSTEM, EMRAH

Publication Date
2013

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Efficient Design and Analysis of Genome-wide Association Studies

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Emrah Kostem

2013

ABSTRACT OF THE DISSERTATION

Efficient Design and Analysis of Genome-wide Association Studies

by

Emrah Kostem

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2013

Professor Eleazar Eskin, Chair

The recent advances in genomic technologies, have made it possible to collect large-scale information on genetic variation across a diverse biological landscape. This has resulted in an exponential influx of genetic information and the field of genetics has become data-rich in a relatively short amount of time. These developments have opened new avenues to elucidate the genetic basis of complex diseases, where the traditional disease study approaches had little success.

In recent years, the genome-wide association study (GWAS) approach has gained widespread popularity for its ease of use and effectiveness, and is now the standard approach to study complex diseases. In GWAS, information on millions of single-nucleotide polymorphisms (SNPs) is collected from case and control individuals. SNP genotyping is cost-effective and due to their abundance in the genome, SNPs are correlated to their neighboring genetic variation, which makes them tags for genomic regions. Typically, each SNP is statistically tested for association to disease, and the genomic regions tagged by the significant SNPs are believed to be harboring the functional variants contributing to disease.

In order to reduce the cost of GWAS and the redundancy in the information col-

lected, an informative subset of the SNPs, or tag SNPs, are genotyped. Typically, the genomic regions harboring the significantly associated tag SNPs may be large and contain many additional polymorphisms. At this stage of the study it may not be clear which specific genes or polymorphisms are in fact most strongly associated to disease. We present a novel framework for designing cost-effective follow-up association studies to further characterize such regions by genotyping additional SNPs to identify all the associated polymorphisms. This identification of all associated polymorphisms provides a catalog of all possible functional variants, and the values of the actual association statistics at these polymorphisms may provide information to identify causal variants. We present the utility of our method in identifying significant associations and causal variants using simulated and real GWAS datasets.

Although GWAS have been widely used to study associations of SNPs to disease phenotypes, there has been growing interest in applying the GWAS approach to high-throughput biological phenotypes, such as gene expression. In these studies, the goal is to identify genomic regions that affect gene expression levels, known as expression quantitative trait loci (eQTL). A challenge in applying GWAS to eQTL studies is that there are tens of thousands of measurements, each representing the expression level of one gene, for each sample tested, as opposed to values for one or two clinical traits. This results in a tremendous computational burden when performing the analysis, requiring computation for billions of tests and demands substantial computational resources. We present a novel two-stage approach to efficiently identify all of the significant associations without testing all the SNPs. In the first-stage, a small number of informative SNPs across the genome are tested. Based on their observed associations, our approach locates the regions that may contain significant SNPs and only tests additional SNPs from those regions. We demonstrate that this method increases the computational speed of eQTL studies by a factor of ten, and can be applied to reduce the computational burden of a wide range of association statistics.

Finally, we develop a novel approach to address a problem that has been of fundamental interest to geneticists for decades. The contribution of genetics to a trait, termed as heritability, is often measured by the amount of variation in the trait that is due to genetics. Heritability, quantifies the role of genetics in a trait and provides insight about disease etiology. Traditionally, heritabilities were estimated in studies of individuals with known relatedness such as classical twin studies. Recently, estimating the heritability of a trait from unrelated individuals using GWAS data, and further, partitioning the heritability into the contributions of genomic regions has received a lot of attention. Existing methods partition the heritability by jointly estimating the contributions of all regions. However, these methods are computationally intractable and may be inaccurate when the number of regions is large. In this work, we present an alternative approach that partitions the total heritability into the contributions of an arbitrary number of regions, while performing these computations in parallel. We demonstrate that our method is more accurate and computationally efficient than existing approaches.

The dissertation of Emrah Kostem is approved.

Adnan Darwiche

David Heckerman

Jake Lusis

Stott Parker

Eleazar Eskin, Committee Chair

University of California, Los Angeles

2013

To my family: Nilufer, Ayca, Guler and Levent.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Dr. Eleazar Eskin for his support and guidance in evolving my research skills. I have learnt a lot from him on when and how to follow my intuition. I also thank my committee members, Dr. Adnan Darwiche, Dr. David Heckerman, Dr. Jake Lusis and Dr. Stott Parker for their inspiration, mentoring and feedback that made this dissertation better. I also acknowledge my past and present "ZarLab" colleagues for their helpful discussions and being representers of great things.

I would like to thank all my collaborators in Lusis Lab, especially Dr. Brian Bennett, Dr. Brian Parks and Dr. Mete Civelek for sharing their insight and support. I will certainly miss our fun gatherings under the roof of Jake and his wife Margarete.

I would like to thank Emel and Ayhan Ergul for their assistance in helping me to open vital doors. I also thank my friends Semih Kazazoglu, Dr. Erk Subasi, Dr. Baris Bagci and Iskender Aksan for being there at hard times.

I cannot thank enough to my wife Ayca for her constant support, patience and stressing on my behalf. The last but not the least, I am grateful to my family for their love, encouragement and support across the ocean. With so many more reasons, I dedicate this dissertation to them.

VITA

1998–2002	B.Sc. Electrical and Electronics Engineering, Middle East Technical University, Ankara, Turkey.
2002–2004	M.Sc.Electrical and Computer Engineering, University of Massachusetts, Lowell, MA.
2004–2005	Scientific Programmer, La Jolla Institute for Allergy and Immunology, San Diego, CA.
2005–2006	Computer Engineer, PassmoreLab, San Diego, CA.
2006–2007	Software Engineer, Sony-SOE, San Diego, CA.
2007–2013	Doctoral Student and Research Assistant, University of California, Los Angeles, CA.

PUBLICATIONS

"Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions." **Kostem E**, Eskin E. Am J Hum Genet. 2013 - In press.

"Efficiently identifying significant associations in genome-wide association studies." **Kostem E**, Eskin E. J Comput Biol. 2013 - In press.

"Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice." Parks BW, Nam E, Org E, **Kostem E**, Norheim F, Hui ST, Pan C, Civelek M, Rau CD, Bennett BJ, Mehrabian M, Ursell LK, He A, Castellani LW, Zinker B, Kirby M, Drake TA, Drevon CA, Knight R, Gargalovic P, Kirchgessner T, Eskin E, Lusis AJ. Cell Metab. 2013 Jan 8;17(1).

"Hybrid mouse diversity panel: a panel of inbred mouse strains suitable for analysis of complex genetic traits." Ghazalpour A, Rau CD, Farber CR, Bennett BJ, Orozco LD, van Nas A, Pan C, Allayee H, Beaven SW, Civelek M, Davis RC, Drake TA, Friedman RA, Furlotte N, Hui ST, Jentsch JD, **Kostem E**, Kang HM, Kang EY, Joo JW, Korshunov VA, Laughlin RE, Martin LJ, Ohmen JD, Parks BW, Pellegrini M, Reue K, Smith DJ, Tetradis S, Wang J, Wang Y, Weiss JN, Kirchgessner T, Gargalovic PS, Eskin E, Lusis AJ, LeBoeuf RC. Mamm Genome. 2012 Oct;23(9-10)

"High-resolution association mapping of atherosclerosis loci in mice." Bennett BJ, Orozco L, **Kostem E**, Erbilgin A, Dallinga M, Neuhaus I, Guan B, Wang X, Eskin E, Lusis AJ. Arterioscler Thromb Vasc Biol. 2012 Aug;32(8)

"Mouse genome-wide association and systems genetics identify Asxl2 as a regulator of bone mineral density and osteoclastogenesis." Farber CR, Bennett BJ, Orozco L, Zou W, Lira A, **Kostem E**, Kang HM, Furlotte N, Berberyan A, Ghazalpour A, Suwanwela J, Drake TA, Eskin E, Wang QT, Teitelbaum SL, Lusis AJ. PLoS Genet. 2011 Apr;7(4)

"Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms." **Kostem E**, Lozano J, Eskin E. Genetics. 2011 Jun;188(2)

"Fine mapping in 94 inbred mouse strains using a high-density haplotype resource." Kirby A, Kang HM, Wade CM, Cotsapas C, **Kostem E**, Han B, Furlotte N, Kang EY, Rivas M, Bogue MA, Frazer KA, Johnson FM, Beilharz EJ, Cox DR, Eskin E, Daly MJ. Genetics. 2010 Jul;185(3)

"A high-resolution association mapping panel for the dissection of complex traits in mice." Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, Neubauer M, Neuhaus I, Yordanova R, Guan B, Truong A, Yang WP, He A, Kayne P, Gargalovic P, Kirchgessner T, Pan C, Castellani LW, **Kostem E**, Furlotte N, Drake TA, Eskin E, Lusis AJ. Genome Res. 2010 Feb;20(2)

CHAPTER 1

Introduction

1.1 Introduction

A wave of technological developments is reshaping the world of life sciences and medicine, ushering in the genomic era. We can now simultaneously collect high-throughput molecular information from thousands of individuals, constructing datasets that were unimaginable just a few decades ago. Genotyping technologies can deliver information on millions of DNA polymorphisms across the genome, and we can monitor the activity levels of tens of thousands of genes in an individual using microarrays.

The cost of genomic technologies is falling exponentially over time[Met10, Ste10, Mar11]. For example, the Human Genome Project[Con04], the first to successfully sequence the human genome (2003), took 13 years and cost $2.7 billion, whereas the current cost of sequencing a genome is approaching $1000 and takes less than a week. Within the next ten years, it is expected to drop to as low as $10 and take just a few minutes to genotype all 3 billion nucleotides in the human genome.

The rapid drop in cost and corresponding increase in throughput of genomic technologies provide unprecedented opportunities to study and understand the genetic basis of complex diseases such as cancer, heart disease, diabetes and autoimmune diseases. Complex diseases are common in the population and account for the majority of human mortality and healthcare spending across the world. Elucidating the genetic risk factors of complex diseases and understanding their etiology allows for cost-effective

medical services through early disease diagnosis and clinical treatments tailored to the individual, a concept known as personalized medicine.

Complex diseases do not show the "simple" inheritance pattern observed in Mendelian diseases, where alterations in a single gene or a unique locus is causal for a phenotype[LS94, RM96, CB01, GNA02]. In a complex disease, multiple genes are involved, each with low-penetrance, where each gene modestly increases the probability of disease and does not completely determine disease status. These factors often render the traditional genetic dissection approaches, such as linkage analysis[LS94, RM96], ineffective tools to study complex diseases.

Linkage analysis democratized disease research by allowing researchers to study human diseases systematically without any prior knowledge about their mechanism. In a linkage study, information on hundreds of polymorphisms across the genome are collected from a carefully selected cohort of closely related individuals, such as affected offsprings and their unaffected parents. The inheritance pattern of disease and the transmission pattern of each polymorphism is statistically analyzed to identify the polymorphisms that segregate with disease phenotype with statistically significant levels. Intuitively, this approach is most powerful when one of the polymorphisms observed is within the causal gene, which is unlikely since only a small number of polymorphisms are collected. Therefore, this approach maps disease loci defined by large chromosomal regions flanked by the significant polymorphisms, referred to as the candidate regions. Using labor intensive and time consuming wet lab experiments[Col92, Col95], researchers then fine map disease loci by discovering novel polymorphisms within the candidate regions and repeat the analysis until disease-genes are identified. Using linkage studies, the genetic basis of hundreds of Mendelian diseases have been successfully discovered, such as Huntingtons disease[Wal07], late-onset Alzheimers disease[PBG91] and some forms of breast cancer[HLN90].

With the advent of genotyping technologies, multinational projects, such as the HapMap Project[Int05] and the 1000 Genomes Project[The10], have started to catalog common genetic variation across various ethnic populations. Among all types of genetic variation, single nucleotide polymorphisms (SNPs), defined as locations in the DNA sequence that are polymorphic in the population, are the most abundant in the genome. Even though a SNP is not the underlying causal variant in a disease, due to their abundance, SNPs are correlated with genetic variation that may directly be functional in disease, and thus be used as genomic markers. The availability of reference datasets on genetic variation and genotyping technologies have made it possible to develop commercial SNP arrays that can cost-effectively collect information on millions of SNPs across the genome.

These developments have changed the perspective of research in complex diseases and enabled the genome-wide association study (GWAS) era[DR95, RM96, Int05, HS09]. In a GWAS, information on millions of SNPs are collected from thousands of unrelated individuals with and without the disease, termed as the cases and the controls. Typically, each SNP is statistically tested for disease association by comparing its minor allele frequency (MAF) between the cases and the controls, and the regions harboring significant association are believed to be functional in disease. In general, GWAS provides better resolution and power than linkage analysis to map disease-associated genes[RM96, CB01]. Furthermore, GWAS allows a simpler study design since unrelated individuals is a more abundant resource than families with a particular disease inheritance pattern. The GWAS approach has also been utilized to identify regions of the genome which harbor variation affecting gene expression or expression quantitative trait loci (eQTLs)[Boc03, RK06, CLA09]. In eQTL studies, tens of thousands of gene expression levels are measured and the GWAS approach is applied to each gene expression level. Hundreds of GWASs have been performed on dozens of complex diseases and have successfully discovered many novel loci involved in

disease[HSJ09].

Although GWAS is a very powerful tool, it comes with some drawbacks. Systematic relatedness, or population structure, among the individuals may lead to spurious associations and increase the false positive rate[PSR00, VP05]. Typically, two models of population structure are considered, population stratification and cryptic relatedness. In the population stratification model, individuals come from populations with different MAF distributions on SNPs and different disease prevalence. Hence, if a population is over-represented among the cases, a SNP that segregates this population from the others may become disease-associated even though it is not. In the cryptic relatedness model, spurious associations may arise due to shared common ancestry among a subset of the individuals. These individuals are enriched for genomic regions that are identical by descent (IBD). If such individuals are over-represented either in the cases or the controls, a SNP within an IBD region may become spuriously associated to disease. Many methods have been proposed to address the population structure problem[PR99, SFY01, RG01, ZFG06, KSS10, LLL11, ZS12]. These methods may substantially increase the computational burden of computing the association statistic.

Recently, another application of GWAS has gained tremendous recognition in order to determine the role of genetics compared to the environment in complex diseases, termed as the heritability of disease. Heritability, in the ”narrow-sense”[VHW08], quantifies the influence of the additive genetic effects relative to the environment. Traditionally, heritabilities were estimated in studies of individuals with known relatedness such as classical twin studies[DSD12]. The GWAS era has created new possibilities for estimating heritability from unrelated individuals using their inferred relatedness from the observed SNP data[ZK12]. These estimates obtained from GWAS datasets explain a large fraction of the estimates of heritability obtained from twin studies[SSP03]. More recently, partitioning the heritability[YMP11] into the contribu-

tions of genomic regions has received a lot of attention with important applications to study complex traits.

1.2 Contributions and Organization of the Thesis

In Chapter 2, I introduce a method for designing follow-up association studies to identify all disease-associated polymorphisms. In order to reduce the cost of GWAS, commercial genotyping arrays collect information on a subset of the SNPs, termed tag SNPs, across the genome. Each tag SNP is a marker for a genomic region which may contain dozens of genes and many additional polymorphisms. Therefore, once the significantly disease-associated tag SNPs are identified, it may not be clear to the investigator which specific genes or polymorphisms are most associated to disease. In addition, biological validation on all such candidates may be costly, labor intensive and time consuming. How to cost-effectively follow-up and further investigate the regions represented by the significant tag SNPs presents a challenge. The method I introduce aims to design a cost-effective follow-up study to identify all the disease-associated polymorphisms within these regions. A complete set of all associated polymorphisms can be seen as a catalog of all possible functional variants, and the actual values of the association statistics at these polymorphisms provide information about which of these polymorphisms may be causal in disease.

In Chapter 3, I introduce a method that improves the computational efficiency of analyzing expression quantitative trait loci (eQTL) studies. A challenge in applying GWAS to gene expression data is that there are tens of thousands of measurements, each representing the expression level of one gene, for each sample tested, as opposed to values for one or two clinical traits. This results in a tremendous computational burden when performing the analysis, requiring computation for billions of tests and demands substantial computational resources, particularly when applying novel statis-

tical methods to account for the population structure. The introduced method is a two-stage testing procedure that identifies all of the significant associations more efficiently than testing all the SNPs. In the first-stage, a small number of informative SNPs across the genome are tested. Based on their observed associations, our approach locates the regions that may contain significant SNPs and only tests additional SNPs from those regions. I demonstrate that this method increases the computational speed of eQTL studies by a factor of ten and can be applied to reduce the computational burden of a wide range of association statistics.

In Chapter 4, I introduce a method that improves the accuracy and efficiency of estimating the contributions of genomic regions to heritability from GWAS data. Partitioning the heritability accounted for by common SNPs into the contributions of genomic regions has received a lot of attention with important applications for understanding the genetic architecture of complex diseases. Current methods partition the total heritability by jointly estimating the contributions of all regions. However, these methods are computationally intractable and may be inaccurate when the number of regions is large. The method I introduce is an alternative approach that partitions the total heritability into the contributions of an arbitrary number of regions. In addition, I show that our method can characterize the population structure better by estimating the effects of population stratification and cryptic relatedness more accurately.

In Chapter 5, I conclude by summarizing the contributions of the dissertain and discuss additional research directions.

CHAPTER 2

Increasing Power of Genome-wide Association Studies by Collecting Additional SNPs

2.1 Motivation

In order to reduce the cost of genome-wide association studies (GWAS) and the redundancy in the information collected, an informative subset of the SNPs, termed tag SNPs, are genotyped in GWAS. Tag SNPs are selected by utilizing the correlation structure between the SNPs, referred to as linkage disequilibrium (LD). Tag SNP selection under different criteria has been very well investigated[BYP05, Str04, Str05, CDG06, CGM03, HKS05, LA04, PLW05, QGA06, SRS06, CER04].

However, genomic regions which are in LD with the most significantly associated tag SNPs are often large and may contain many additional polymorphisms. At this stage of the study it may not be clear to the investigator which specific genes or polymorphisms lead to increase in disease risk. Additionally, biological validation on all such candidates may be costly and time consuming. However, the regions harboring these candidates is characterized by identifying all of the associated polymorphisms. A complete set of all associations can be seen as a catalog of all possible functional variants and the actual values of the association statistics at these polymorphisms provide information about which of these polymorphisms may be causal.

How to cost-effectively follow-up and further investigate the regions represented

by the significant tag SNPs presents a challenge. Given all the SNPs within these regions, or candidate SNPs, one way to identify the associated SNPs is to collect genotype information on every candidate SNP. However, this approach is highly inefficient as only a small percentage of the candidate SNPs are likely to be associated. Ideally, we would like to know which of the candidate SNPs are the associated SNPs before genotyping them. We introduce a follow-up study approach in which a subset of the candidate SNPs, or follow-up SNPs, which are likely to be associated is selected and genotyped in the original case-control individuals. We propose a follow-up SNP selection method with the goal of maximizing the number of statistically associated SNPs among the follow-up SNPs. We assume that the candidate SNPs are catalogued in a reference human genetic variation dataset, such as the HapMap Project. The intuition behind our method is that a candidate SNP which is strongly correlated to a significantly associated tag SNP is likely to be associated as well (i.e., such as perfect correlation). We formalize this intuition to compute the probability of each candidate SNP being associated, and select the follow-up SNPs accordingly.

Our approach may also be used in conjunction with the fine-mapping efforts which obtain the complete sequence information for regions of interest. In a typical GWAS, thousands of case-control individuals are employed to achieve a reasonable statistical power in the study and sequencing thousands of individuals is still a difficult and costly task compared to genotyping SNPs. Therefore, a small number of individuals can be sequenced in these regions to catalogue the candidate SNPs and the follow-up SNPs selected using our method can be genotyped in all of the case-control individuals.

Below we formalize two intuitive approaches which we refer to as the distance and the correlation-based traditional follow-up SNP selection approaches. The traditional follow-up SNP selection approaches choose candidate SNPs which are within a certain distance or correlated above a minimum correlation cut-off value to the most signifi-

cant tag SNPs. The distance-based traditional approach assumes that the neighboring SNPs are strongly correlated, yet because of the complexity of the LD landscape, SNPs close to each other may not necessarily be correlated, and assuming so may fail to identify the associated SNPs. Similarly, the correlation-based traditional approach may fail as a consequence of using the same minimum correlation cut-off value for every candidate and tag SNP. Predicting whether a candidate SNP is associated or not depends on the particular values of the observed tag SNP statistic, pairwise correlation, and the effect size of the candidate SNP. Our method outperforms traditional approaches both in simulated and real GWAS data. In our simulations we use the SNPs available from the HapMap Project and the widely used Affymetrix 500K SNP array as the candidate and tag SNPs respectively. We generate various simulated candidate regions to compare the performance of the follow-up SNP selection approaches. For performance evaluation under real GWAS, we use data available from the Wellcome Trust Case Control Consortium[WTC07]. In each of the seven disease GWAS we use half of the observed SNPs as the tag SNPs and the remaining as the candidate SNPs.

2.2 Preliminaries

Given a biallelic SNP m_i with true population minor allele frequency p_i, we denote the true case and control minor allele frequencies with p_i^+ and p_i^- and the observed frequencies with $\hat{p}_i^+$ and $\hat{p}_i^-$. For the simplicity of our equations we will work with balanced case and control panels of size $N/2$, which yields N chromosomes in each panel.

The following association statistic S_i is evaluated at SNP m_i for large enough N,

$$S_i \equiv \frac{\hat{p}_i^+ - \hat{p}_i^-}{\sqrt{2/N}\sqrt{\hat{p}_i(1-\hat{p}_i)}} \sim \mathcal{N}\left(\frac{p_i^+ - p_i^-}{\sqrt{2/N}\sqrt{p_i(1-p_i)}}, 1\right) \text{ where } \hat{p}_i = \frac{\hat{p}_i^+ + \hat{p}_i^-}{2}. \tag{2.1}$$

S_i is normally distributed with mean $\lambda_i\sqrt{N}$ (the non-centrality parameter), and unit

variance, where

$$\lambda_i\sqrt{N} = \frac{p_i^+ - p_i^-}{\sqrt{2p_i(1-p_i)}}\sqrt{N}. \tag{2.2}$$

A SNP m_i is associated with the disease trait if its non-centrality parameter (NCP) is not zero, $\lambda_i\sqrt{N} \neq 0$. The NCP of a SNP is unknown and the association of a SNP is inferred statistically. We refer to SNP m_i as *statistically* associated or significant, if under the null distribution ($\lambda_i\sqrt{N} = 0$) of the association statistic S_i, the observed statistic $\hat{s}_i$ is in the rejection region defined by the significance level α, i.e. $|\hat{s}_i| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$, where Φ^{-1} is the quantile function of the standard normal distribution. Note that even though a SNP can be associated, it may not be detected as statistically associated.

Additionally, given that the SNP m_i is associated ($\lambda_i\sqrt{N} \neq 0$), m_i can be either a causal SNP or correlated to a causal SNP. Assuming there is a single causal SNP m_c and m_i is not the causal SNP, then the following relation holds between the NCPs of m_c and m_i[PP01]:

$$\lambda_i\sqrt{N} = r_{ic}\lambda_c\sqrt{N}, \tag{2.3}$$

where r_{ic} is the correlation coefficient between the two SNPs.

2.3 Problem Formulation

Consider T tag SNPs were genotyped in a GWAS and the association statistic of each tag SNP is computed, $\hat{s}_1, ..., \hat{s}_T$. Given K candidate SNPs, the follow-up SNP selection problem is to choose k follow-up SNPs from the K candidate SNPs to genotype in the original GWAS case/control panels, using the observed statistics of T tag SNPs and the pairwise correlation between the SNPs.

Assume that in a follow-up study the chosen follow-up SNPs are genotyped in the case/control individuals of the original GWAS. In this scheme an ideal follow-up SNP

selection, referred to as *oracle*, selects the candidate SNPs which are in fact significant in the original GWAS. We evaluate the performance of a follow-up SNP selection method with the precision criteria, which is the proportion of the follow-up SNPs that are significant, or true positives (TP). Finally, at a given significance level α, if TP(α) denotes the number of the true positives among k follow-up SNPs, then the precision can be expressed as follows:

$$\text{Precision}(\alpha) = \frac{\text{TP}(\alpha)}{k}.$$

2.4 Traditional Follow-up SNP Selection Approaches

To the best of our knowledge although there is no existing method addressing the follow-up SNP selection, we formalize two intuitive approaches which we refer to as the distance and the correlation-based traditional follow-up SNP selection approaches in which each candidate SNP is paired with a tag SNP.

Under the correlation and distance-based traditional follow-up SNP selection approaches each candidate SNP is paired with a tag SNP. The tag SNP that pairs a candidate SNP is selected as the tag SNP with the highest pairwise correlation or within a certain distance of the candidate SNP. For each candidate/tag SNP pair, $\hat{s}_t$ denotes the observed association statistic of the tag SNP, r_{it} denotes the pairwise correlation and d_{it} denotes the distance between the candidate and the tag SNPs.

Under the correlation-based traditional approach, follow-up SNPs are selected as follows. First the candidate/tag SNP pairs are sorted according to the significance of their tag SNP's observed association statistic, from the most significant to the least significant. If two pairs have the same significance of tag SNP statistic, then the pair with stronger pairwise correlation precedes the other one. To select the follow-up SNPs a minimum pairwise correlation cut-off value, $r_{min} > 0$, is given such that the

top k pairs carrying stronger pairwise correlation than r_{min}, are selected.

The distance-based traditional approach selects the follow-up SNPs with respect to the distance between the candidate and tag SNPs. A distance window, d_{max}, is given such that starting from the most significant tag SNP, candidate SNPs which are within the distance window are paired with the tag SNP. In addition, pairs with the same significance of tag SNP statistic are sorted with respect to how close their distance is to their tag SNP. The top k pairs are selected to determine the follow-up SNPs.

2.5 A Statistical Framework to Analyze Follow-up SNP Selection

We introduce a simple statistical framework for analyzing the follow-up SNP selection problem. Although this model is an oversimplification, it captures the essence of follow-up SNP selection and is easy to analyze. We pair each candidate SNP with the tag SNP with the highest correlation. We assume that the candidate/tag SNP pairs are independent and focus on the joint distribution of the association statistics in a pair. We can estimate the covariance between two SNPs as shown by Han et al.[HKE09] and calculate the joint distribution of their association statistics as follows:

$$\begin{bmatrix} S_i \\ S_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \lambda_i\sqrt{N} \\ r_{it}\lambda_i\sqrt{N} \end{bmatrix}, \begin{bmatrix} 1 & r_{it} \\ r_{it} & 1 \end{bmatrix} \right). \tag{2.4}$$

Note that $S_i \sim \mathcal{N}\left(\lambda_i\sqrt{N}, 1\right)$ and $S_t \sim \mathcal{N}\left(r_{it}\lambda_i\sqrt{N}, 1\right)$. The value of the candidate SNP non-centrality parameter (NCP), $\lambda_i\sqrt{N}$, depends on whether or not the SNP is causal. We introduce a new parameter, c_i, as the probability of the candidate SNP being causal and assume it attains a certain NCP of $\lambda_c\sqrt{N}$. Under these assumptions the joint distribution can be expressed as follows:

$$\begin{bmatrix} S_i \\ S_t \end{bmatrix} \sim \begin{cases} \mathcal{N}\left(\begin{bmatrix} \lambda_c\sqrt{N} \\ r_{it}\lambda_c\sqrt{N} \end{bmatrix}, \begin{bmatrix} 1 & r_{it} \\ r_{it} & 1 \end{bmatrix} \right) & \text{if the candidate SNP is causal} \\ \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r_{it} \\ r_{it} & 1 \end{bmatrix} \right) & \text{otherwise.} \end{cases} \tag{2.5}$$

Although the joint distribution depends on two unknown parameters, the NCP of the causal SNP, $\lambda_c\sqrt{N}$ and the probability of the candidate SNP being the causal SNP, c_i, it will be shown in the results section that variation in these parameters has small effect on the choices of follow-up SNPs.

2.6 Follow-up SNP Selection under the Proposed Framework

Using the joint distribution given in equation (2.5) the conditional distribution of the candidate SNP statistic given the observed tag SNP statistic can be expressed as follows:

$$S_i |\ S_t = \hat{s}_t \sim \begin{cases} \mathcal{N}\left(\lambda_c\sqrt{N}\,(1 - r_{it}^2) + r_{it}\hat{s}_t, 1 - r_{it}^2\right) & \text{if the candidate SNP is causal} \\ \mathcal{N}\left(r_{it}\hat{s}_t, 1 - r_{it}^2\right) & \text{otherwise.} \end{cases} \tag{2.6}$$

Let $\phi(x\,;\mu,\sigma^2)$ denote the density of a univariate normal distribution with mean μ and variance σ^2. The density function of the conditional distribution of the candidate SNP statistic, f, can expressed as the following mixture density:

$$f\left(s_i \mid S_t = \hat{s}_t\right) = c_i\phi\left(s_i\,;\lambda_c\sqrt{N}\,(1 - r_{it}^2) + r_{it}\hat{s}_t, 1 - r_{it}^2\right) + (1 - c_i)\,\phi\left(s_i\,;r_{it}\hat{s}_t, 1 - r_{it}^2\right). \tag{2.7}$$

Using the above equation we can express the probability of a candidate SNP being statistically associated given the observed value of its tag SNP statistic as:

$$P\left(|S_i| > \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \mid \hat{s}_t\right), \tag{2.8}$$

which will be referred to as $\pi_i(\hat{s}_t)$. The selection of follow-up SNPs is achieved by computing the $\pi_i(\hat{s}_t)$ for each candidate SNP and ranking them descending with respect to their $\pi_i(\hat{s}_t)$. We will refer to the follow-up SNP selection method where each candidate SNP is paired with the tag SNP that has the highest pairwise correlation as RFSS (*R*ank-based *F*ollow-up *S*NP *S*election).

Consider the example given in Figure 2.1, the selection of two follow-up SNPs from four candidate/tag SNP pairs, where the true framework parameters are $\lambda_c\sqrt{N} = 5.73$ and $c_i = 10^{-6}$. We assume there are one million candidate SNPs (only four shown in the example) and the significance level is 10^{-8}, which takes into account the multiple testing correction. The NCP value of $\lambda_c\sqrt{N} = 5.73$ corresponds to 50% power at the causal SNP. Under the correlation-based traditional approach, for the given tag SNP statistics and pairwise correlations, two follow-up SNPs can be selected using three different values for the minimum pairwise correlation cut-off value, $r_{min} = \{0.50, 0.90, 0.92\}$. For each candidate SNP the $\pi_i(\hat{s}_t)$ value can be calculated as: $\pi_1 = 0.084$, $\pi_2 = 0.0004$, $\pi_3 = 0.0007$ and $\pi_4 = 0.008$. The two optimal follow-up SNPs are m_1 and m_4 which have the highest $\pi_i(\hat{s}_t)$ values. This example shows that the correlation-based traditional approach may fail to identify the optimal follow-up SNP selection under all of the possible correlation cut-off values.

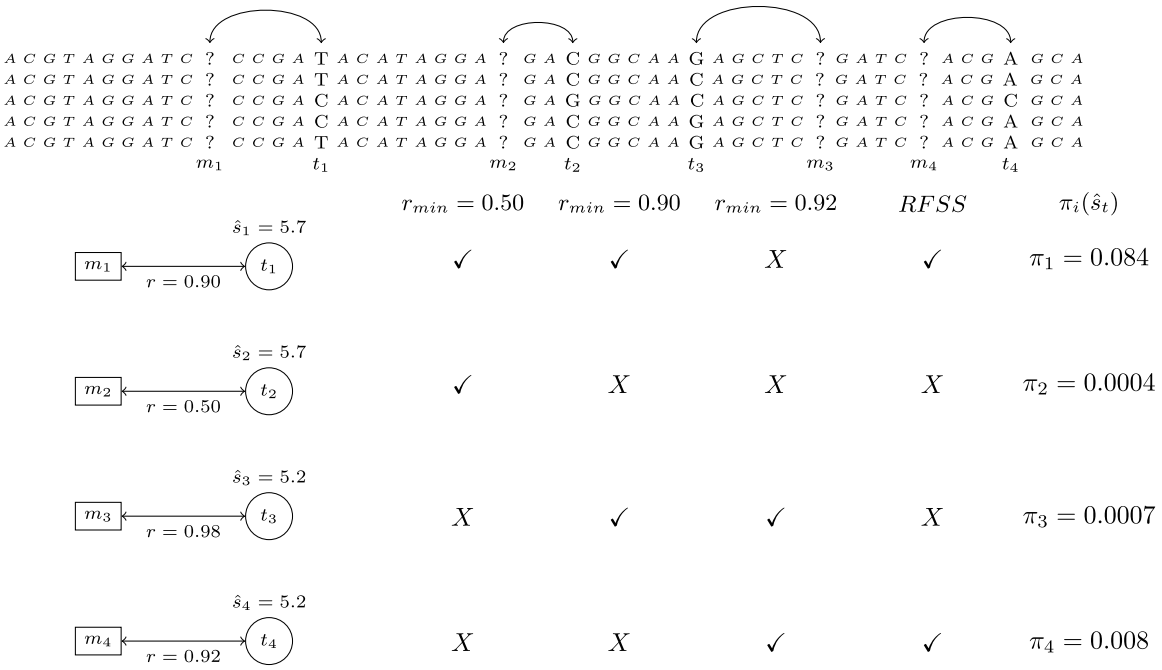


Figure 2.1: Consider a genome-wide follow-up SNP selection with 10^6 candidate SNPs, where only four candidate SNPs m_1, m_2, m_3, m_4 and their tag SNPs t_1, t_2, t_3, t_4 are shown. Assuming $\lambda_c\sqrt{N} = 5.73$, $c_i = 10^{-6}$ and $\alpha = 10^{-8}$, the correct ranking of the four candidate SNPs is m_1, m_4, m_3 and m_2. Consider selecting two follow-up SNPs among the four candidate SNPs. The correlation-based traditional approach can realize this selection using three different minimum correlation cut-off values (r_{min}). Under the columns $r_{min} = 0.50$, $r_{min} = 0.90$ and $r_{min} = 0.92$, the follow-up SNPs selected under each cut-off value are indicated with a checkmark. Under the column $\pi_i(\hat{s}_t)$ the probability of each candidate SNP being statistically significant conditioned on its observed tag SNP is given. Unlike the proposed method, the correlation-based traditional approach fails to identify the optimal selection (m_1 and m_4) under all possible thresholds.

2.7 Extending the Statistical Framework to Incorporate Multiple tag SNPs and the Neighboring candidate SNPs

We now extend the RFSS approach by grouping each candidate SNP with multiple tag SNPs with the highest correlations, and relaxing the assumption of the candidate SNPs being independent. In this scheme, we incorporate two additional sources of information. First, in addition to the best tag SNP, the observed statistics of the top highly correlated tag SNPs are utilized. Second, even though a candidate SNP may not be causal, it may still be associated (e.g., $\lambda_i\sqrt{N} \neq 0$) as a result of being correlated to a candidate SNP that is causal. We refer to the multivariate extension of the RFSS approach as *mRFSS*, and present its performance improvements in the results section.

Given a candidate SNP m_i, we consider its L most strongly correlated tag SNPs. Let $\mathbf{R}_{i\mathbf{L}}$ denote the $L \times 1$ vector of the correlation coefficients between m_i and the L tag SNPs. Similarly, let $\mathbf{S_L}$ and $\lambda_{\mathbf{L}}\sqrt{N}$, respectively, be the $L \times 1$ vectors of the association statistics and non-centrality parameters (NCPs) of the tag SNPs, and $\mathbf{\Sigma_L}$ be the $L \times L$ matrix of their pairwise correlation coefficients. The joint distribution of the association statistics of the candidate SNP m_i and the L tag SNPs follows a multivariate normal distribution, which can be expressed as follows:

$$\begin{bmatrix} S_i \\ \mathbf{S_L} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \lambda_i\sqrt{N} \\ \lambda_{\mathbf{L}}\sqrt{N} \end{bmatrix}, \begin{bmatrix} 1 & \mathbf{R}_{i\mathbf{L}}^t \\ \mathbf{R}_{i\mathbf{L}} & \mathbf{\Sigma_L} \end{bmatrix} \right). \tag{2.9}$$

In equation (2.9) the NCPs of the candidate SNP m_i and the tag SNPs are unknown. Suppose the causal SNP m_c is known, where it correlates to m_i by r_{ic}, and to the tag SNPs by the $L \times 1$ vector $\mathbf{R}_{c\mathbf{L}}$. Using the indirect association rule (2.3), we can express the NCPs of m_i and the tag SNPs as follows: $\lambda_i\sqrt{N} = r_{ic}\lambda_c\sqrt{N}$ and $\lambda_{\mathbf{L}}\sqrt{N} = \mathbf{R}_{c\mathbf{L}}\lambda_c\sqrt{N}$.

Although the causal SNP is unknown, we can consider each candidate SNP m_k

as the causal SNP with probability c_k, and use the indirect association rule to resolve the unknown NCPs. For each candidate SNP m_k, where $k \in \{1, \ldots, K\}$, let r_{ik} and $\mathbf{R}_{k\mathbf{L}}$ denote the correlation coefficient of m_k to m_i and the L most strongly correlated tag SNPs to m_i. Then joint distribution of $(S_i, \mathbf{S_L})$ can then be expressed as a 2-level hierarchical model, which uses the indirect association to compute the NCPs.

$$\begin{bmatrix} S_i \\ \mathbf{S_L} \end{bmatrix} \sim \begin{cases} \mathcal{N}\left(\begin{bmatrix} r_{i1}\lambda_c\sqrt{N} \\ \mathbf{R}_{1\mathbf{L}}\lambda_c\sqrt{N} \end{bmatrix}, \begin{bmatrix} 1 & \mathbf{R}_{i\mathbf{L}}^t \\ \mathbf{R}_{i\mathbf{L}} & \mathbf{\Sigma_L} \end{bmatrix} \right) & \text{if candidate SNP } m_1 \text{ is causal} \\ \quad \vdots & \\ \mathcal{N}\left(\begin{bmatrix} r_{iK}\lambda_c\sqrt{N} \\ \mathbf{R}_{K\mathbf{L}}\lambda_c\sqrt{N} \end{bmatrix}, \begin{bmatrix} 1 & \mathbf{R}_{i\mathbf{L}}^t \\ \mathbf{R}_{i\mathbf{L}} & \mathbf{\Sigma_L} \end{bmatrix} \right) & \text{if candidate SNP } m_K \text{ is causal} \\ \\ \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \mathbf{R}_{i\mathbf{L}}^t \\ \mathbf{R}_{i\mathbf{L}} & \mathbf{\Sigma_L} \end{bmatrix} \right) & \text{otherwise.} \end{cases} \tag{2.10}$$

Consequently, the density function, f, of the conditional distribution of the candidate SNP statistic S_i given the vector of observed tag SNP statistics $\hat{\mathbf{s}}_\mathbf{L}$ can be written as follows:

$$\begin{aligned} f\left(s_i \mid \mathbf{S_L} = \hat{\mathbf{s}}_\mathbf{L}\right) = & \sum_{k=1}^{K} c_k \phi\left(s_i\,; \lambda_c\sqrt{N}\left(r_{ik} - \mathbf{R}_{i\mathbf{L}}^t\mathbf{\Sigma_L^{-1}}\mathbf{R}_{k\mathbf{L}}\right) + \mathbf{R}_{i\mathbf{L}}^t\mathbf{\Sigma_L^{-1}}\hat{\mathbf{s}}_\mathbf{L}, 1 - \mathbf{R}_{i\mathbf{L}}^t\mathbf{\Sigma_L^{-1}}\mathbf{R}_{i\mathbf{L}}\right) \\ & + \left(1 - \sum_{k=1}^{K} c_k\right) \phi\left(s_i\,; \mathbf{R}_{i\mathbf{L}}^t\mathbf{\Sigma_L^{-1}}\hat{\mathbf{s}}_\mathbf{L}, 1 - \mathbf{R}_{i\mathbf{L}}^t\mathbf{\Sigma_L^{-1}}\mathbf{R}_{i\mathbf{L}}\right). \end{aligned} \tag{2.11}$$

Note that in Equation (2.11) there is one mixture component for each candidate SNP m_k being the causal SNP (with weight c_k), and a mixture component that corresponds to having no causal SNPs with a weight $\left(1 - \sum_{k=1}^{K} c_k\right)$.

Although the model (2.10) considers K candidate SNPs, in practice it can be simplified by using only the M neighboring candidate SNPs with the highest pairwise correlations to m_i, since the remaining $K - M$ candidate SNPs do not contribute to the sum in Equation (2.11). In our experiments, we have used ten most strongly correlated neighboring candidate SNPs, $M = 10$. Additionally, the choice of the most strongly correlated tag SNPs affects the computational cost and the numerical stability of the method. If any two tag SNPs are strongly correlated with each other the matrix $\mathbf{\Sigma_L}$ may become nearly singular, therefore in practice we discarded any tag SNP with correlation > 0.9 to any of the already included tag SNPs. Finally, in our experiments we have chosen the number of tag SNPs to be at most ten, $L \leq 10$.

2.8 Performance of the correlation-based Traditional Approach

In this section we analyze the expected performance (EP) of the correlation-based traditional approach on a single follow-up SNP. That is, how often a candidate SNP is observed as statistically associated given that it is selected as a follow-up SNP. The selection of a candidate SNP as a follow-up SNP depends on the ordering of all the candidate/tag SNP pairs and the given minimum correlation cut-off value, r_{min}. In order to simplify the interdependence, we introduce a new parameter called the minimum statistic cut-off value, $\hat{s}_{min}$, where $\hat{s}_{min} > 0$. As a rule, a candidate SNP is selected as a follow-up SNP if the observed statistic of its tag SNP, $\hat{s}_t$, is above $\hat{s}_{min}$. We assume there is a one to one mapping between r_{min} and $\hat{s}_{min}$, such that for every r_{min}, there exists a $\hat{s}_{min}$ value.

The follow-up SNP selection rule defined for the correlation-based traditional approach using $\hat{s}_{min}$ is comparable to RFSS, which uses a conditional probability threshold, π^*, as a follow-up SNP selection rule. It can be shown that the probability of a candidate SNP being significant given the observed value of its tag SNP statistic,

$P\left(|S_i| > \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \mid \hat{s}_t\right)$, is a monotonic function of $\hat{s}_t$. That is, for a given candidate/tag SNP pair, for every π^* it is possible to determine a unique $\hat{s}_i^*$ value such that $\pi^* = P\left(|S_i| > \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \mid \hat{s}_i^*\right)$, where the candidate SNP is selected if $|\hat{s}_t| > \hat{s}_i^*$. Therefore, we can compare the two selection rules based on $\hat{s}_{min}$ and $\hat{s}_i^*$, where the correlation-based traditional approach uses the same $\hat{s}_{min}$ for every candidate/tag SNP pair, and RFSS determines a $\hat{s}_i^*$ for each candidate/tag SNP pair based on $\lambda_c\sqrt{N}$, c_i, r_{it} and α. Below we show how the expected performance changes with respect to $\hat{s}_{min}$.

The EP under the correlation-based traditional approach can be computed directly from the corresponding joint distribution of the association statistics given in equation (2.5) as $P\big(|S_i| > \Phi^{-1}(1-\frac{\alpha}{2}) \mid |S_t| > \hat{s}_{min}\big)$. This probability can be expressed as

$$\frac{P(|S_i| > \Phi^{-1}(1-\frac{\alpha}{2}),\ |S_t| > \hat{s}_{min})}{P(|S_t| > \hat{s}_{min})}. \tag{2.12}$$

In Figure 2.2, the densely shaded region represents where the follow-up SNP is significant. Likewise, in the lightly shaded region, the candidate SNP is selected as a follow-up SNP however it is not significant. Therefore EP can be expressed as the ratio of the probability in the densely shaded region to all shaded regions as given in equation (2.13).

For given c_i, $\lambda_c\sqrt{N}$ and α, we can write the EP as a function of r_{it} and $\hat{s}_{min}$,

$$EP\left(\hat{s}_{min}, r_{it}\right) = \frac{c_i A\left(d_1\right) + (1-c_i) A\left(d_2\right)}{c_i B\left(\hat{s}_{min} - r_{it}\lambda_c\sqrt{N}\right) + (1-c_i) B\left(\hat{s}_{min}\right)},$$

where

$$\begin{aligned}
d_1 &= \left\{(\hat{s}_i, \hat{s}_t) \,\middle|\, |\hat{s}_i| > \Phi^{-1}\left(1-\tfrac{\alpha}{2}\right) - \lambda_c\sqrt{N}\ ,\ |\hat{s}_t| > \hat{s}_{min} - r_{it}\lambda_c\sqrt{N}\right\},\\
d_2 &= \left\{(\hat{s}_i, \hat{s}_t) \,\middle|\, |\hat{s}_i| > \Phi^{-1}\left(1-\tfrac{\alpha}{2}\right)\ ,\ |\hat{s}_t| > \hat{s}_{min}\right\},\\
A(d) &= \iint_d \frac{1}{2\pi\sqrt{1-r_{it}^2}}\, e^{\left(\frac{-\left(\hat{s}_t^2 + \hat{s}_i^2 - 2r_{it}\hat{s}_i\hat{s}_t\right)}{2\left(1-r_{it}^2\right)}\right)}\, d\hat{s}_i\, d\hat{s}_t,\\
B(x) &= 1 - \Phi\left(x\right) + \Phi\left(-x\right).
\end{aligned} \tag{2.13}$$

In Figure 2.3 each of $\lambda_c\sqrt{N}$, c_i and r_{it} is varied while keeping the other two parameters fixed and the effect of variation in EP is shown. When the variation of EP is compared between the change in (a) $\lambda_c\sqrt{N}$ or in (b) c_i to (c) r_{it}, we observe that the largest variation in EP is due to the pairwise correlation. A second observation is that varying $\lambda_c\sqrt{N}$ or c_i approximately corresponds to a shift of EP in the horizontal axis. This suggests that if the same $\lambda_c\sqrt{N}$ and c_i parameters are used to calculate the EPs of any two candidate/tag SNP pairs, the order of the pairs with respect to their EPs will always be the same. That is, as long as the values of $\lambda_c\sqrt{N}$ and c_i are close to the true values, the selection of the follow-up SNPs is robust to uncertainties in these parameters. This property is utilized in RFSS and mRFSS, where we show the concordance of the selections under different combinations of the framework parameters.

2.9 Performance Comparison under Simulated Data

We evaluate and compare the performance of the traditional approaches and our proposed methods, RFSS and mRFSS, using simulated association studies generated using the ENCODE regions from the HapMap Project. We use the ENCODE SNPs as the candidate SNPs and the Affymetrix 500K array as the tag SNPs. There are 10 ENCODE regions, each 500K base pairs long, which are genotyped separately under the four HapMap populations. The correlation structure and the minor allele frequency of the SNPs in these regions vary depending on the specific population. A summary of the number of candidate and tag SNPs in each region and population is given in Table 2.1.

We simulate the follow-up study of a genome-wide association study (GWAS) as follows: assuming there is a single causal SNP, the region where the significant tag SNPs are located, is simulated by an ENCODE region. Using each ENCODE region

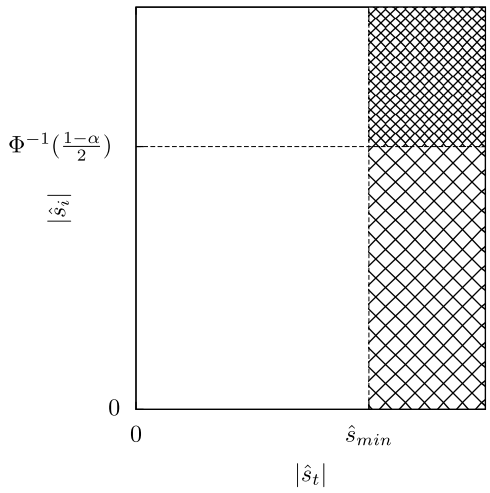
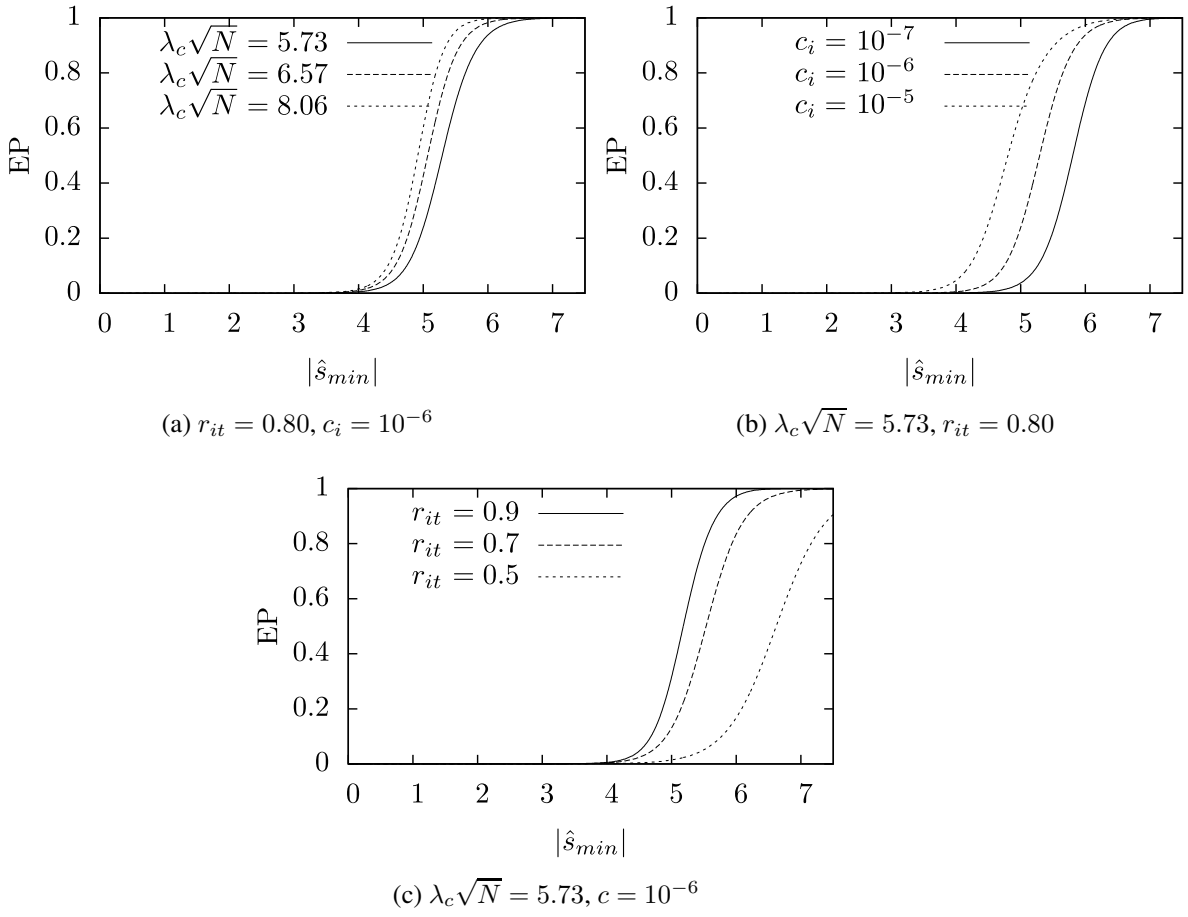


Figure 2.2: Regions with shades represent where a candidate SNP is selected as a follow-up SNP based on the observed statistic of its tag SNP. In the region with dense-shades the follow-up SNP is statistically associated, whereas in the light-shade region it is not.

ENCODE Region	The number of SNPs in each ENCODE region per population							
	CEU		CHB		JPT		YRI	
	candidate	tag	candidate	tag	candidate	tag	candidate	tag
ENm010.7p15.2	687	45	569	34	591	40	833	48
ENr112.2p16.3	1191	65	1031	65	1046	73	1805	84
ENr131.2q37.1	1187	121	993	123	981	122	1459	130
ENr113.4q26	1311	57	979	56	979	55	1467	59
ENm013.7q21.13	1621	182	1225	151	1228	157	1800	181
ENm014.7q31.33	1657	203	1424	189	1422	193	1892	203
ENr321.8q24.11	758	78	685	90	701	87	1179	103
ENr232.9q34.11	673	45	656	46	645	47	1001	52
ENr123.12q12	1193	92	1135	90	1394	86	1134	83
ENr213.18q12.1	795	68	644	58	676	62	1216	76

Table 2.1: Summary of the number of candidate and tag SNPs in the ENCODE regions in each population.



(a) $r_{it} = 0.80$, $c_i = 10^{-6}$

(b) $\lambda_c\sqrt{N} = 5.73$, $r_{it} = 0.80$

(c) $\lambda_c\sqrt{N} = 5.73$, $c = 10^{-6}$

Figure 2.3: The effect of 2.3a non-centrality parameter, 2.3b probability of the candidate SNP being causal and 2.3c the pairwise correlation in the Expected Precision (EP) is shown. The NCPs of 5.73, 6.57 and 8.06 correspond to 50%, 80% and 99% statistical power at the causal SNP. The unknown parameters, non-centrality parameter of the causal SNP, $\lambda_c\sqrt{N}$, and probability of a candidate SNP being causal, c_i, have smaller impact in the performance compared to the pairwise correlation. ($\alpha = 10^{-8}$)

in each population, we simulate 2000 association study panels. In half of these panels, we implant a causal SNP, which is randomly selected among the candidate SNPs, and in the rest of the panels there are no causal SNPs. In order to generate the association statistics of the SNP, we simulate the case/control genotypes of the SNPs. In each panel, we generate the genotypes by sampling haplotypes depending on whether or not there is a causal SNP in the panel. If there is no causal SNP, we randomly sample case/control individuals' haplotype from the specific haplotype pool of the ENCODE region from the corresponding population. If there is a causal SNP, we divide the haplotype pool into two separate pools based on the causal SNP's allele. The true case and control minor allele frequencies of the causal SNP are calculated as follows. We assume the HapMap frequency of the causal SNP under the corresponding population is the true control frequency, and determine the true case minor allele frequency such that the non-centrality parameter (NCP) of each causal SNP yields 50% statistical power. In other words, when the causal SNP is genotyped, half of the time it is detected as significant. Once the true case and control minor allele frequencies are determined at the causal SNP, we sample case and control individuals using these probabilities from the corresponding haplotype pool. We use a genome-wide significance level α of 10^{-8} and under 50% statistical power the NCP of the causal SNP is $\lambda_c\sqrt{N} = 5.73$.

In different HapMap populations and ENCODE regions the correlation structure among the SNPs varies greatly. We compare the performance of the methods for each of the ENCODE regions in each population. In Figure 2.4 an example performance comparison is given in the ENm010.7p15.2 ENCODE region using the CEU HapMap population. The precision of each method is plotted along the size, k, of the follow-up SNPs (Figure 2.4(a)). We give the ideal follow-up SNP selection as the "oracle" as a reference to compare each method to. The vertical line indicates the total number of statistically associated candidate SNPs in the simulation data.

The performance of the correlation-based traditional approach depends on the minimum correlation cut-off value, r_{min}, used for selecting the follow-up SNPs. If r_{min} is high, such as $r = 0.9$, the correlation-based traditional approach performs well for a small number of follow-up SNPs and performance degrades rapidly as more follow-up SNPs are collected. However significant candidate SNPs may not be strongly correlated to the most significant tag SNPs, and by using a high r_{min} value such candidate SNPs cannot be selected. On the other hand when r_{min} is low (Figures 2.4(a) or 2.6), the traditional approach selects the significant candidate SNPs that are correlated weakly to the most significant tag SNPs, however a price is paid by selecting many follow-up SNPs most of which are not significant. Hence when the number of follow-up SNPs is small, the precision is significantly lower. The distance-based traditional approach performs worse than the correlation-based counterpart which can be observed in Figure 2.4(a) as shown for the distance windows of 1k and 10k base pairs. Whether or not a candidate SNP is statistically associated depends on the values of the observed tag SNP statistic and the pairwise correlation, and due to the complexity of the LD landscape SNPs located near each other may not be strongly correlated.

RFSS makes assumptions about the $\lambda_c\sqrt{N}$ and c_i which effects the estimates of the decision rule. We evaluate how incorrect assumptions on these parameters affect the performance by varying the value of these parameters in our simulations. Figure 2.4(b) shows that even with incorrect assumptions about these parameters, performance of our method is nearly identical to the performance using correct parameters.

In Table 2.2 we shown the summary of performance comparisons under all generated association studies. In each population in every ENCODE region, the precision of each method is reported two times, first when the number of follow-up SNPs is equal to the total number of the significant candidate SNPs, N_s, and second when number of follow-up SNPs is twice this value. For each population, the two sets of precisions

obtained over all ENCODE regions are averaged respectively. The proposed RFSS and mRFSS methods significantly performs better than the traditional approaches for the relevant sizes of follow-up SNPs.

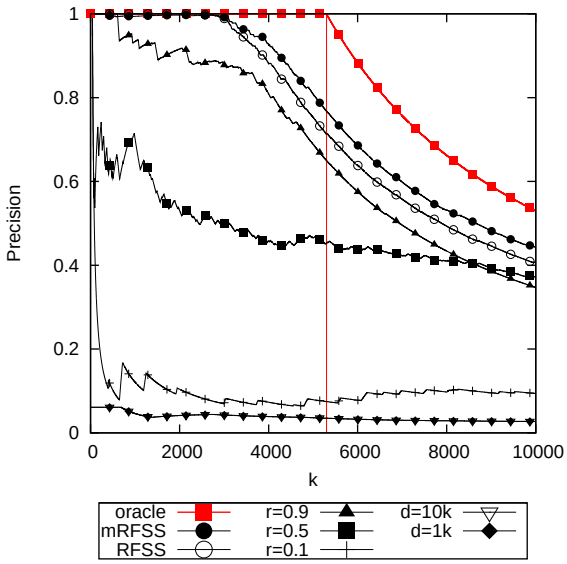
2.10 Performance Comparison using Incorrect HapMap Population Correlations

We further experiment with the effect of using incorrect correlations on the performance of each method. Among the four HapMap populations, the correlations among the SNPs vary the most between the CEU and YRI populations. We use the correlation values from the YRI population to select follow-up SNPs from the simulation data generated in the ENm010.7p15.2 ENCODE region using the CEU population.

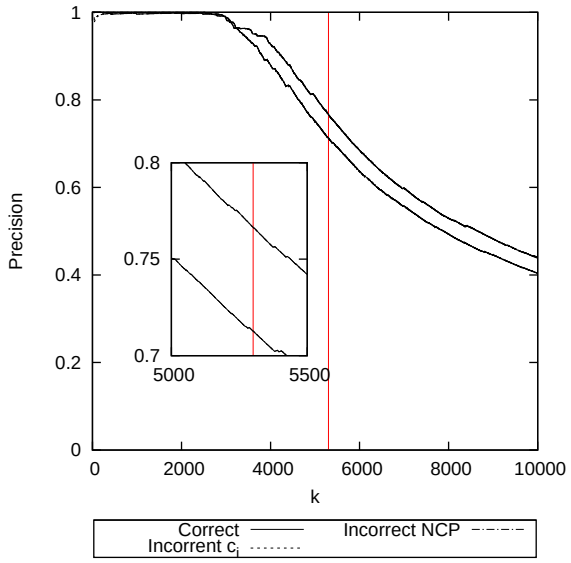
In Figure 2.5 the performance comparison of the traditional and the proposed methods is given. We observe that each method performs worse than their performance when the correct correlations are used. However, the proposed methods performed better than the traditional approaches.

2.11 Performance on Discovering Causal SNPs

Next, we compare the performance of the traditional and the proposed approaches in discovering the causal SNPs. Note that neither the traditional nor the proposed methods are designed for this goal. Nevertheless, we compare their performance on what percentage of the significant causal SNPs that are present in the data are selected as follow-up SNPs. Here we assume only the follow-up SNPs that are observed as significant are further analyzed, which may lead to the discovery of whether or not a SNP is causal.



(a) Performance comparison the proposed and the traditional methods along the number, k, of follow-up SNPs collected. The vertical line indicates the number of statistically significant candidate SNPs in the simulation data.



(b) "Correct" assumes the parameters used in the simulations. In "Incorrect NCP" the causal SNP is believed to have 80% statistical power. In "Incorrect c_i" c_i is ten times less likely.(Box inside is zoomed in for clarity.)

Figure 2.4: Sample performance evaluation under the ENCODE region ENm010.7p15.2 in CEU. In (a) the correlation and distance-based traditional approaches are compared to the optimal approach. In (b) the effect of using wrong parameters in the performance of the proposed methods is shown.

Population	Average precision on significant candidate SNPs over all ENCODE regions per population													
	k = number of significant candidate SNPs (N_s)							$k = 2N_s$						
	RFSS	mRFSS	Traditional Approaches					RFSS	mRFSS	Traditional Approach				
			r=0.1	r=0.5	r=0.9	d=1k	d=10k			r=0.1	r=0.5	r=0.9	d=1k	d=10k
CEU	0.83	0.90	0.10	0.47	0.77	0.05	0.08	0.46	0.49	0.10	0.39	0.39	0.05	0.05
CHB	0.83	0.91	0.11	0.48	0.77	0.07	0.06	0.46	0.50	0.10	0.38	0.40	0.07	0.06
JPT	0.78	0.89	0.10	0.45	0.70	0.06	0.06	0.43	0.49	0.09	0.36	0.36	0.06	0.05
YRI	0.57	0.79	0.04	0.39	0.43	0.02	0.02	0.33	0.47	0.04	0.28	0.22	0.02	0.02

Table 2.2: Performance results of the traditional approaches and the proposed methods. Under each ENCODE region and population, the precisions of each method are recorded when the number of follow-up SNPs, k, is equal to the number of the significant candidate SNPs at the corresponding simulation N_s and two times this value, $k = 2N_s$. For each method the recorded precisions are then averaged over the ENCODE regions per each population.

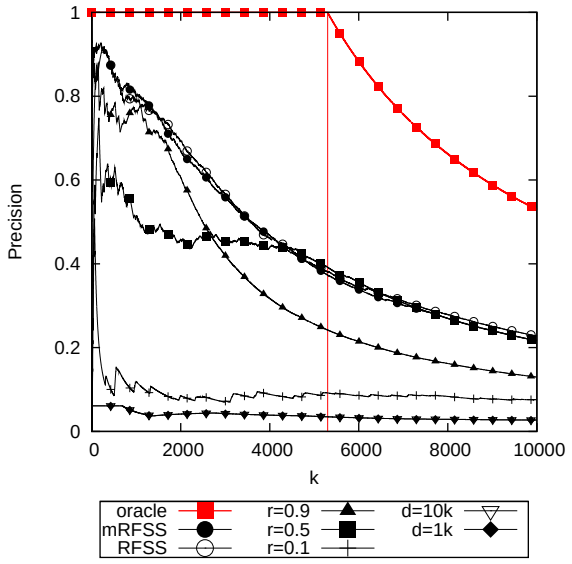


Figure 2.5: Performance comparison of the traditional and proposed methods when incorrect correlation coefficients between the SNPs are used. Simulation is generated in the ENCODE region ENm010.7p15.2 in CEU population and the correlation coefficients from the YRI population are used.

In Figure 2.6 we show the candidate SNPs in the order they are selected as follow-up SNPs in each method. We plot the hidden statistic of each follow-up SNP as a green circle and indicate the causal SNPs with a black ring. The red horizontal lines mark the significance threshold for the association statistic under the significance level $\alpha = 10^{-8}$. We use two extreme minimum correlation cut-off values of $r_{min} = 0.1$ and $r_{min} = 0.9$ for the correlation-based traditional approach. The blue circles indicate the statistic of the tag SNP that corresponds to each follow-up SNP. The plot for the distance-based traditional approach is omitted as it performs similar to the $r_{min} = 0.1$. Additionally, the likelihood of each candidate SNP being significant is shown with a blue line in the plots the proposed methods. The traditional approaches select the follow-up SNPs despite of the significance level or the assumed statistical power at the causal SNP. However, the proposed approaches rank the candidate SNPs based on these parameters where the likelihood of each candidate SNP being significant changes accordingly. We observe that when multi tag SNPs are used this ranking is more successful. Both of our proposed methods identify significantly higher number of causal SNPs, where the mRFSS method prioritizes causal candidate SNPs earlier in the selection.

In Table 2.3 the summary of the average performance comparisons on each HapMap population is shown. In each population and ENCODE region, the performance of the methods are recorded twice where the number of the follow-up SNPs, k, equals to the number of significant candidate SNPs in the corresponding simulation, $k = N_s$ and two times this value $k = 2N_s$. In each population, we then average the performance of each method over the ENCODE regions. The proposed approaches lead to the discovery of significantly more causal SNPs than the traditional approaches.

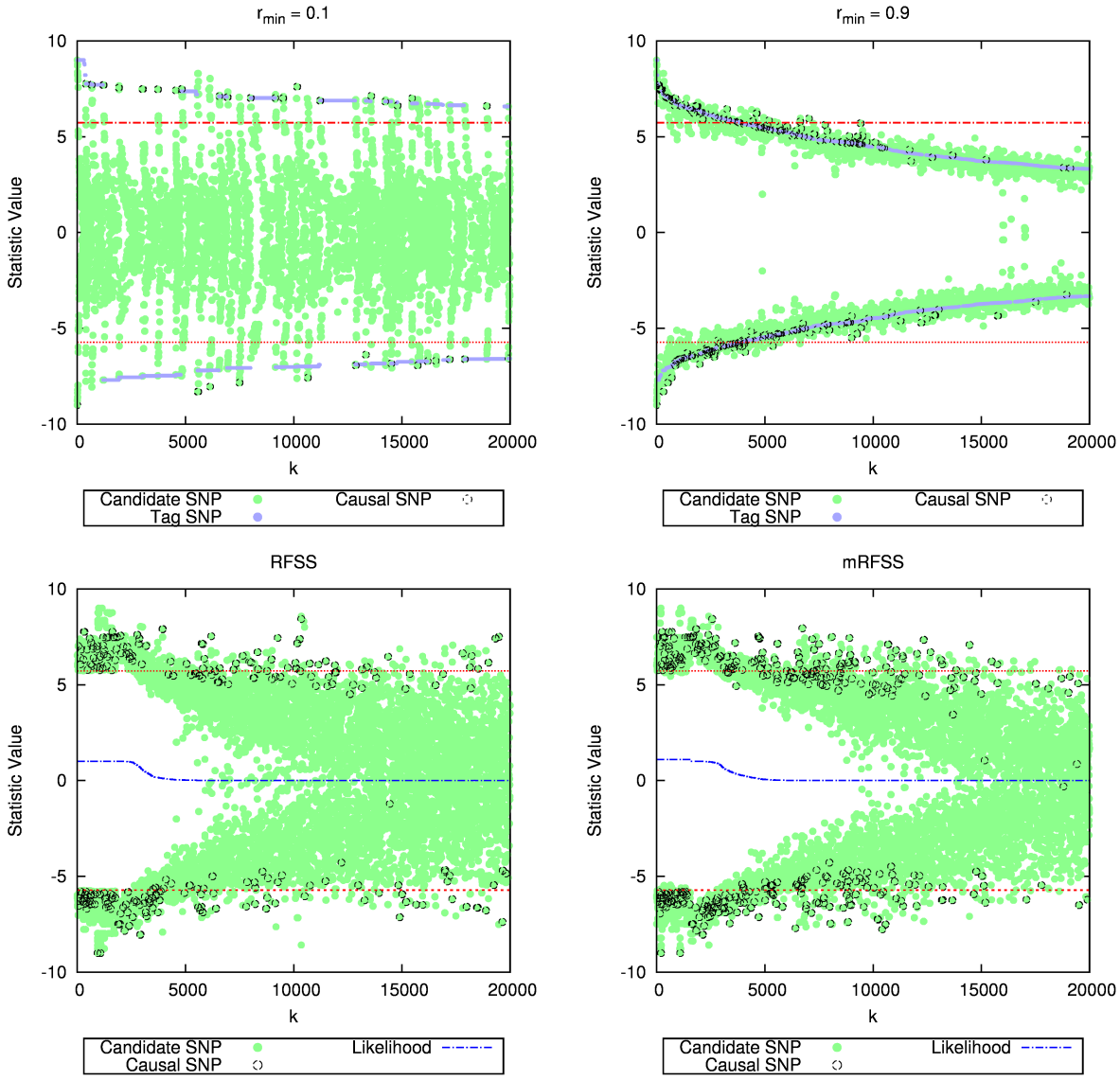


Figure 2.6: Green circles indicate the chosen candidate SNPs under each method in the ENCODE region ENm010.7p15.2 in CEU population simulation dataset. Red horizontal lines indicate the significance threshold and black circles indicate the causal SNPs. The traditional approach is shown via two minimum correlation cut-off values, $r_{min} = 0.1$ and $r_{min} = 0.9$, where $r_{min} = 0.1$ approximates the distance-based approach. Both of the proposed methods identify significantly higher number of causal SNPs, where the mRFSS method prioritizes causal candidate SNPs much more effectively.

2.12 Performance Comparison under Real GWAS Data

We compare the performance of the correlation-based traditional approach to the proposed methods using real GWAS studies. We use the data available from the WTCCC GWAS on seven human diseases which provides genotype data on 2000 case individuals per disease and 3000 shared controls. Each individual is believed to be of European origin (CEU) and genotyped with the Affymetrix 500K array. Bipolar Disorder (BD) and Hypertension (HT) are excluded from the analysis as no statistically significant associations are observed. The performance comparisons are evaluated on the remaining five diseases, Coronary Artery Disease (CAD), Crohn's Disease (CD), Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D) and Type 2 Diabetes (T2D). We use approximately half of the genotyped Affymetrix SNPs as candidate SNPs and the rest as the tag SNPs. Using each SNP's unique reference identifier number (rsID), the SNPs with odd rsIDs are used as the candidate SNPs.

The performance of the correlation-based traditional approach is recorded for minimum correlation cut-off values of 0.9, 0.5 and 0.1. For the proposed methods we assumed the statistical power at the causal SNP to be 50%, and used the value of 10^{-6} for the probability of a candidate SNP being causal. In Figure 2.7, a sample performance comparison is given between the correlation-based traditional approach and the proposed methods under Rheumatoid Arthritis (RA).

In Table 2.4 the summary of performance comparisons in all five diseases is given. In each disease the precision of each method is reported twice, first when the number of follow-up SNPs is equal to the total number of statistically associated candidate SNPs, and second when number of follow-up SNPs is two times the total number of statistically associated candidate SNPs. The proposed RFSS and mRFSS methods consistently outperform the correlation-based traditional approach.

Population	Average percentage of the significant **causal** candidate SNPs collected over all ENCODE regions per population													
	$k =$ number of significant candidate SNPs (N_s)							$k = 2N_s$						
	RFSS	mRFSS	Traditional Approaches					RFSS	mRFSS	Traditional Approach				
			r=0.1	r=0.5	r=0.9	d=1k	d=10k			r=0.1	r=0.5	r=0.9	d=1k	d=10k
CEU	0.54	0.69	0.05	0.25	0.48	0.02	0.02	0.71	0.93	0.09	0.48	0.51	0.04	0.04
CHB	0.55	0.74	0.04	0.24	0.50	0.02	0.02	0.72	0.96	0.09	0.46	0.53	0.05	0.05
JPT	0.57	0.77	0.05	0.25	0.50	0.03	0.03	0.72	0.96	0.09	0.49	0.52	0.06	0.05
YRI	0.34	0.65	0.02	0.19	0.24	0.06	0.01	0.46	0.89	0.04	0.33	0.24	0.01	0.01

Table 2.3: The average performance selecting the significant causal SNPs are compared between the traditional approaches and the proposed methods. Under each ENCODE region and population, the percentage of the significant causal SNPs recalled are recorded when the number of follow-up SNPs, k, is equal to the number of the significant candidate SNPs at the corresponding simulation N_s and two times this value, $k = 2N_s$. For each method the recorded percentages are averaged over the ENCODE regions and shown per population.

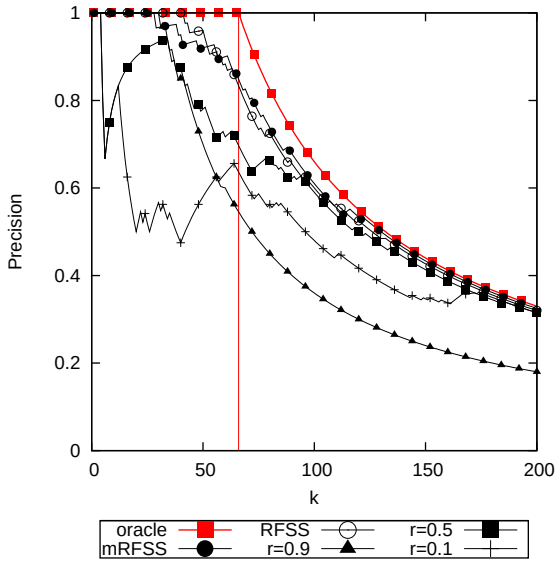


Figure 2.7: Performance evaluation under Rheumatoid Arthritis (RA).

2.13 Concordance of the RFSS Selections under Uncertainty in the Framework Parameters

We determine the rank of the follow-up SNPs which represents the order they will be picked, using different values for the framework parameters, $\lambda_c\sqrt{N}$ and c_i, in the WTCCC data. The concordance between two such rankings indicates the invariance between using different values. In the two rankings, the top k candidate SNPs are examined and the proportion of the candidate SNPs which have the same rsIDs is recorded. The concordance of a ranking to itself is always one for all k, hence two rankings are highly similar if their concordance is close to one.

In Figure 2.8 show the concordance plots of RFSS in Rheumatoid Arthritis (RA), varying $\lambda_c\sqrt{N}$ between 5.73 and 6.57 and c_i between 10^{-5} and 10^{-8}. The concordance of the mRFSS performs similar to the RFSS, and hence not shown. The vertical line indicates the total number of statistically significant candidate SNPs under the significance level $\alpha = 10^{-8}$. We observe that the follow-up SNP selection under RFSS is highly concordant between different values used for the framework parameters, verifying empirically that even though our proposed RFSS method depends on unknown parameters, follow-up SNP selection is highly concordant within the range of likely values.

2.14 Discussion

Currently, genome-wide association studies are initiating the journey of disease gene discovery. The results of a GWAS are effective in guiding investigators to the genomic regions that may contain causal variants. However such regions may contain many polymorphisms, and biologically validating all of them is not an efficient use of resources. We introduced a follow-up study approach which may be useful in better

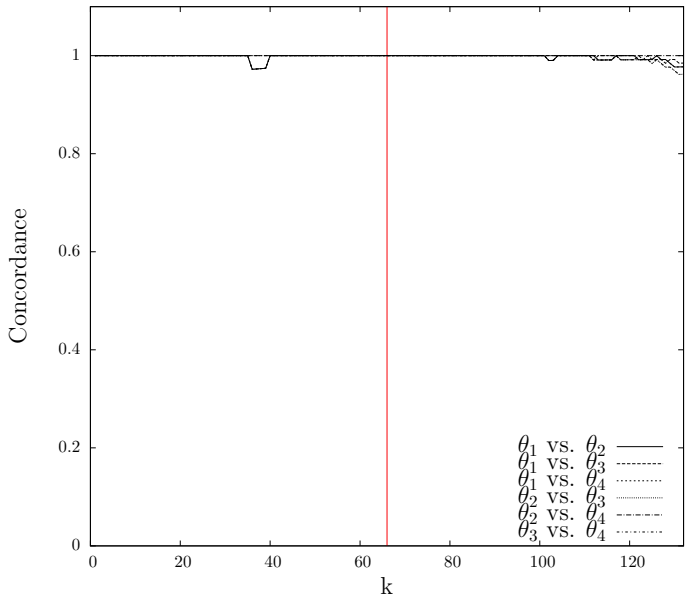


Figure 2.8: Concordance of the selected follow-up SNPs in Rheumatoid Arthritis (RA) between different framework parameters. $\theta_1 \equiv (\lambda_c\sqrt{N} = 5.73\ ,\ c_i = 10^{-5})$. $\theta_2 \equiv (\lambda_c\sqrt{N} = 6.57\ ,\ c_i = 10^{-8})$. $\theta_3 \equiv (\lambda_c\sqrt{N} = 5.73\ ,\ c_i = 10^{-8})$. $\theta_4 \equiv (\lambda_c\sqrt{N} = 6.57\ ,\ c_i = 10^{-5})$.

characterizing the associated regions and may provide investigators a clear direction for biological validations.

Our method makes certain assumptions for predicting whether or not a candidate SNP is significantly associated, which depends on the observed tag SNP statistics, pairwise correlation between the tag and candidate SNPs and the effect size of the causal SNP. We model the effect size of a candidate SNP assuming that the effect size of the causal SNP is known and the probability of each candidate SNP being causal is given. These parameters affect the probability of a candidate SNP being associated, thus whether or not the candidate SNP should be selected for the follow-up study. We empirically show that although the true values for these parameters are not known, surprisingly, within the range of likely values, predicting the association of a candidate SNP is mainly influenced by the known parameters, and errors in the assumptions on the unknown parameters usually do not change the predictions of our method. Additionally, our method does not require any genotype data either on the candidate or the tag SNPs, which makes it easy to apply on the currently available GWAS results.

Our approach requires knowledge of the correlation between the test statistics at each marker and the relation between non-centrality parameters at causal variants and correlated markers. For the standard association statistic presented in this chapter, the correlation between statistics is simply the correlation (r) between the markers. For other association statistics which have the same correlation structure, we can directly apply our method. For statistics which have a different correlation structure, we can still apply our method if we replace equations (2.7) and (2.11) with the correlation structure specific to the association study.

Several groups have performed sequencing in regions implicated in association studies in a small number of individuals to discover many new polymorphisms in those

regions. These studies then follow up a subset of those polymorphisms by genotyping them in the entire case and control population. Usually many polymorphism are discovered, each with different correlation structure with respect to the tag SNPs. Our approach can be directly applied to select which subset of these discovered polymorphisms to collect by using the correlations estimated from the sequenced data.

Study	Observed precision on significant candidate SNPs in WTCCC studies									
	$k = N_s$					$k = 2N_s$				
	RFSS	mRFSS	Traditional Approach			RFSS	mRFSS	Traditional Approach		
			r=0.1	r=0.5	r=0.9			r=0.1	r=0.5	r=0.9
CAD	1.00	1.00	1.00	0.88	0.88	0.50	0.50	0.50	0.50	0.44
CD	0.79	0.92	0.21	0.43	0.64	0.50	0.50	0.14	0.39	0.36
RA	0.82	0.85	0.63	0.69	0.54	0.48	0.49	0.37	0.46	0.27
T1D	0.79	0.81	0.58	0.66	0.56	0.46	0.46	0.42	0.43	0.28
T2D	1.00	1.00	1.00	1.00	0.60	0.50	0.50	0.50	0.50	0.30

Table 2.4: Performance comparison of the correlation-based traditional approach and the proposed methods. The SNPs genotyped in these studies are separated as candidate and tag SNPs based on whether their rsIDs are odd or even. Under each disease, the observed precision values are recorded when the number of follow-up SNPs, k, is equal to the number of the significant candidate SNPs (N_s), and two times this value, $k = 2N_s$.

CHAPTER 3

Efficiently Identifying Significant Associations in Genome-wide Association Studies

3.1 Motivation

The expression quantitative trait loci (eQTL) studies are already very popular[BYC02, BK05, KFT07] and with rapidly decreasing costs of genomic technologies[WGS09, MP11] will likely become more popular in the future. These include, several major efforts collecting expression from multiple-tissues in human[CSE05, SNF07, ETZ08, SBB07, Bak12] and mouse[CLS05, BWD05]. More broadly, application of the GWAS approach to phenotypes measured by other genomic technologies such as those reported by the ENCODE consortium[The04, The07, The11, The12] will face similar computational challenges.

In this chapter, we introduce a novel two-stage method which can be applied to reduce the computational burden of a wide range of association studies including those employ case-control, quantitative trait and mixed-model statistical testing methodologies. In each trait, typically only a small percentage of the SNPs are significantly associated and the SNPs neighboring a significant association have elevated statistics. Intuitively, one can first test an informative subset of the SNPs, termed proxy SNPs, across the genome to quickly locate these regions and test the SNPs therein. This way, many of the regions with no associations can be discarded from the analysis to reduce

the computational burden.

Our novel method for genome-wide rapid association testing (GRAT), guarantees to identify all of the significant associations with high-probability while reducing the total number of tests. The proposed method chooses the proxy SNPs and determines which additional SNPs to test based on the observed proxy SNP statistics and the patterns of linkage disequilibrium (LD) in the region. The key insight underlying GRAT is that by taking advantage of how the statistics at SNPs in LD with each other behave, we can estimate the probability that an untested SNP has a significant association and use this probability to eliminate SNPs from consideration only if they are highly unlikely to have significant associations. We have selected a set of proxy SNPs for the 1000 Genomes Project and any study which imputes their genotyped SNPs in the GWAS to the 1000 Genomes Project SNPs can readily use our approach. We also provide our method for choosing proxy SNPs, which can be applied to any reference dataset. We show through simulations and analysis of real eQTL datasets that the proposed two-stage procedure identifies the significant associations while only testing approximately 10% of the SNPs. GRAT's efficient software implementation reduces the computational time for computing large-scale association studies by a factor of 30 compared to currently used state of the art methods. When our method is applied to association studies that utilize linear mixed models, the speed-up is cumulative with recent efforts that decrease the computational burden of computing the actual association statistic such as EMMAX, FaST-LMM and GEMMA[KSS10, LLL11, ZS12].

3.2 Preliminaries

For the simplicity of description, we consider a balanced case-control genome-wide association study (GWAS) with $N/2$ individuals (N copies of each chromosome) per panel. For our actual experiments, we will use association statistics for quantitative

phenotypes, but the approach assuming case-control phenotypes is equivalent. For SNP m_i, p_i denotes its population minor allele frequency (MAF); p_i^+ and p_i^- denote its population case and control MAFs; $\hat{p}_i^+$ and $\hat{p}_i^-$ denote its observed case and control MAFs in the GWAS. Given the relative risk of the SNP, γ_i, in the disease and the prevalence of the disease, F, in the population, it can be shown that the case and control MAFs of the SNP follow,

$$p_i^+ = \frac{\gamma_i p_i}{(1-\gamma_i)p_i + 1},\ p_i^- = \frac{p_i - F p_i^+}{1-F}. \tag{3.1}$$

A SNP is defined as not associated if $p_i^+ = p_i^-$.

In case-control GWASs the following statistic is widely used, which is normally distributed for large N with mean $\lambda_i\sqrt{N}$ (the non-centrality parameter), and unit variance,

$$S_i = \hat{s}_i = \frac{\hat{p}_i^+ - \hat{p}_i^-}{\sqrt{2\hat{p}_i(1-\hat{p}_i)}}\sqrt{N} \sim \mathcal{N}\left(\lambda_i\sqrt{N}, 1\right), \tag{3.2}$$

$$\text{where } \lambda_i = \frac{p_i^+ - p_i^-}{\sqrt{2p_i(1-p_i)}} \text{ and } \hat{p}_i = \frac{\hat{p}_i^+ + \hat{p}_i^-}{2}.$$

Given the significance level α and the observed value of the test statistic $\hat{s}_i$, the SNP is deemed as significant, or statistically associated, if $|\hat{s}_i| > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$, where $\Phi^{-1}(.)$ is the quantile function of the standard normal distribution. For simplicity, we use the notation: $t_\alpha \equiv \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$. Typically, in a GWAS the significance level is chosen as $\alpha = 10^{-8}$.

3.3 Problem Formulation

We propose the following two-stage testing procedure for identifying the significant associations within a set of SNPs $\mathcal{M}$. Given a subset of the SNPs $\mathcal{T} \subset \mathcal{M}$, referred to as the proxy SNPs, for each proxy SNP, $m_t \in \mathcal{T}$, its association statistic, $\hat{s}_t$, is computed. In the second stage, a decision rule is exercised for each remainder SNP,

$m_i \in \mathcal{M} \backslash \mathcal{T}$, in order to determine whether or not to compute the association statistic of the remainder SNP. The decision rule for a remainder SNP m_i is defined using a proxy SNP, $m_t \in \mathcal{T}$, and a threshold, s_t^*, for its observed statistic $\hat{s}_t$. If the observed statistic of the proxy SNP is more extreme than the threshold value, $\hat{s}_t > s_t^*$ the remainder SNP is tested.

3.4 Proposed Two-stage Framework and its Performance

In a GWAS, the performance of the two-stage approach can be summarized by the total number of SNPs tested (NT), and the percentage of the significant SNPs identified, or the recall rate (RR). The total number of tests is the sum of the tests performed on the proxy SNPs, plus the remainder SNPs that are tested as a result of the decision rules. We use a standard GWAS simulation model[KLE11] to evaluate a given set of proxy SNPs and decision rules based on their *expected* performance within the simulated data.

The simulation model considers the probability of each SNP being causal, c_i, and the non-centrality parameter (NCP) of the causal SNP, $\lambda_c \sqrt{N}$. For simplicity, we give a brief explanation of the simulation procedure for a single causal SNP using a genomic reference dataset such as HapMap. Using the given probabilities of each SNP being causal, at most a single causal SNP is randomly selected. Given the disease prevalence F and the NCP of the causal SNP $\lambda_c \sqrt{N}$, the case and control MAFs, p_c^+ and p_c^- are determined. Next, the HapMap haplotypes are divided into two pools according to the minor and major allele of the causal SNP, and case-control panels are sampled using p_c^+ and p_c^-.

For each simulation dataset, each association statistic is computed to identify which SNPs are significant in the dataset. We then apply the two-stage method to observe the

NT and RR. The expected recall rate (ERR) and the expected number of SNPs to be tested (ENT) then can be computed by repeatedly simulating datasets, applying the two-stage approach and averaging the observed NT and RR value.

3.5 Finding the Optimal Decision Rules for Given Proxy SNPs

For a given set of proxy SNPs, one can determine the decision rules empirically by evaluating the performance of using different threshold values on the remainder SNPs in the simulated data. The empirical approach can be cumbersome and instead we derive an analytical framework for estimating the expected performance, which eliminates the need for generating simulated data and saves time. Furthermore, using this analytical framework we show how to determine the optimal decision rules for the remainder SNPs given a set of proxy SNPs.

A SNP that is disease-associated can be either causal in the disease or in LD with the causal SNP. Given that SNP m_i is the causal SNP, the non-centrality parameter (NCP) of a correlated SNP m_t, $\lambda_t\sqrt{N}$, is proportional to the NCP of the causal SNP, $\lambda_c\sqrt{N}$, by their correlation coefficient, r, where $\lambda_t = r\lambda_c$. It can be shown that the joint distribution of the association statistics of the causal SNP m_i and the non-causal SNP m_t follows a bivariate normal distribution[HKE09].

We follow a conservative approach in which each remainder SNP m_i is paired with the proxy SNP that is most strongly correlated, referred to as the *best*-proxy, and denoted by $m_{b(i)}$. For each remainder SNP m_i, we denote the association statistic of its best-proxy $m_{b(i)}$ with $s_{b(i)}$ and test SNP m_i if its best-proxy SNP association statistic is more extreme than a given threshold, $s_{b(i)} > s^*_{b(i)}$. For simplicity, we assume only the

remainder SNP can be causal and express the density function of the joint distribution,

$$f\left(s_i, s_{b(i)}\right) = c_i \phi\left(\begin{bmatrix} s_i \\ s_{b(i)} \end{bmatrix} ; \begin{bmatrix} \lambda_c\sqrt{N} \\ r\lambda_c\sqrt{N} \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right) + (1 - c_i)\phi\left(\begin{bmatrix} s_i \\ s_{b(i)} \end{bmatrix} ; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right), \tag{3.3}$$

where $\phi(\boldsymbol{x}\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The first term corresponds to having the remainder SNP as causal, with probability c_i, and the second term to not casual with probability $1 - c_i$.

Assume we are given K proxy SNPs, where $\mathcal{T} = \{m_1, \ldots, m_K\}$. The expected number of SNPs to be tested (ENT) can be expressed as the fixed cost of testing K proxy SNPs, plus the expected number of decision rules that are triggered,

$$\mathrm{ENT}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}) = K + \sum_{i=K+1}^{M} \Pr\left(|S_{b(i)}| > s^*_{b(i)}\right). \tag{3.4}$$

We *approximate* the expected recall rate (ERR) as the ratio of the expected number of significant SNPs that the two-stage approach discovers, to the expected number of significant SNPs in a GWAS,

$$\mathrm{ERR}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}) = \frac{\displaystyle\sum_{t=1}^{K} \Pr\left(|S_t| > t_\alpha\right) + \sum_{i=K+1}^{M} \Pr\left(|S_i| > t_\alpha, |S_{b(i)}| > s^*_{b(i)}\right)}{\displaystyle\sum_{i=1}^{M} \Pr\left(|S_i| > t_\alpha\right)}, \tag{3.5}$$

where the first and the second terms in the numerator correspond to the expected number of significant SNPs obtained from testing the proxy SNPs and the remainder SNPs, respectively. Further, we refer to the second term as the expected recall function, which

can be computed using the joint distribution,

$$\begin{aligned} \mathrm{ER}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}) &= \sum_{i=K+1}^{M} \Pr\left(|S_i| > t_\alpha, |S_{b(i)}| > s^*_{b(i)}\right), \\ \Pr\left(|S_i| > t_\alpha, |S_{b(i)}| > s^*_{b(i)}\right) &= \iint_{\Omega_i} f\left(s_i, s_{b(i)}\right) ds_i \, ds_{b(i)}, \end{aligned} \tag{3.6}$$

where $\Omega_i = \left\{ (s_i, s_{b(i)}) \mid |s_i| > t_\alpha \, , |s_{b(i)}| > s^*_{b(i)} \right\}$.

We are interested in determining the decision rules that lead to the least expected number of SNPs to be tested (ENT), while the expected recall rate (ERR) satisfies a given target value, ρ, which can be expressed as an optimization problem,

$$\begin{aligned} &\text{minimize} \quad \mathrm{ENT}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}), \\ &\text{such that} \quad \mathrm{ERR}(s^*_{b(K+1)}, \ldots, s^*_{b(M)}) \geq \rho. \end{aligned} \tag{3.7}$$

In the next section we show that the problem is convex and outline an efficient iterative solution.

3.6 Convexity of the Optimization Problem

The derivative of the expected number of tests (ENT)(3.4) from a single remainder SNP with respect to the decision threshold follows,

$$\begin{aligned} \frac{\partial}{\partial s^*} \mathrm{ENT}(s^*) &= \frac{\partial}{\partial s^*} \left[1 - \int_{-s^*}^{s^*} f(s_t) \, ds_t \right] = -f(s^*) - f(-s^*) \\ &= -c_i \left[\phi\left(s^* - r\lambda_c \sqrt{N}\right) + \phi\left(s^* + r\lambda_c \sqrt{N}\right) \right] - 2(1 - c_i)\phi\left(s^*\right). \end{aligned} \tag{3.8}$$

Note that the second derivative is negative, hence convex. Therefore, the expected number of SNPs to be tested (ENT) is the sum of convex functions and is also convex.

Let us denote the expected recall function by $\mathrm{ER}(s^*) = \Pr\left(|S_i| > t_\alpha, |S_t| > s^*\right) = \rho_i$. Its derivative follows

$$\frac{\partial}{\partial s^*}\mathrm{ER}(s^*) = \frac{\partial}{\partial s^*}\left[\int_{-\infty}^{-s^*}\int_{-\infty}^{-t_\alpha} f(s_i, s_t)\, ds_i\, ds_t\right] + \frac{\partial}{\partial s^*}\left[\int_{-\infty}^{-s^*}\int_{t_\alpha}^{\infty} + \int_{s^*}^{\infty}\int_{-\infty}^{-t_\alpha} + \int_{s^*}^{\infty}\int_{t_\alpha}^{\infty}\right]$$

$$= -\int_{-\infty}^{-t_\alpha}\left(f(s_i, s_t = -s^*) + f(s_i, s_t = s^*)\right)\, ds_i - \int_{t_\alpha}^{\infty}\left(f(s_i, s_t = -s^*) + f(s_i, s_t = s^*)\right)\, ds_i. \tag{3.9}$$

Note that given,

$$f\left(x, y\right) = \phi\left(\begin{bmatrix} x \\ y \end{bmatrix}\ ;\ \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix},\ \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right) \tag{3.10}$$

it can be shown that a cross section of the joint distribution at $y = a$ follows,

$$f\left(x, y = a\right) = \frac{\sqrt{1-r^2}}{\sqrt{2\pi}}\exp\left(-\frac{(a-\mu_y)^2}{2}\right)\phi\left(x\ ;\mu_x + r(a-\mu_y), 1-r^2\right). \tag{3.11}$$

Therefore, using the joint distribution of the statistics of a remainder SNP and its best-proxy, equation (3.9) can be expressed as,

$$\mathrm{ER}^{'}(s^*) =$$

$$\frac{-1}{\sqrt{2\pi}}\Bigg[c_i\exp\left(-\frac{(s^* + r\lambda_c\sqrt{N})^2}{2}\right)\left(\Phi\left(\frac{-t_\alpha - \lambda_c\sqrt{N}(1-r^2) + rs^*}{\sqrt{1-r^2}}\right) + 1 - \Phi\left(\frac{t_\alpha - \lambda_c\sqrt{N}(1-r^2) + rs^*}{\sqrt{1-r^2}}\right)\right)$$

$$+c_i\exp\left(-\frac{(s^* - r\lambda_c\sqrt{N})^2}{2}\right)\left(\Phi\left(\frac{-t_\alpha - \lambda_c\sqrt{N}(1-r^2) - rs^*}{\sqrt{1-r^2}}\right) + 1 - \Phi\left(\frac{t_\alpha - \lambda_c\sqrt{N}(1-r^2) - rs^*}{\sqrt{1-r^2}}\right)\right)$$

$$+2(1-c_i)\exp\left(-\frac{(s^*)^2}{2}\right)\left(\Phi\left(\frac{-t_\alpha + rs^*}{\sqrt{1-r^2}}\right) + \Phi\left(\frac{-t_\alpha - rs^*}{\sqrt{1-r^2}}\right)\right)\Bigg]. \tag{3.12}$$

It can be shown that $\mathrm{ER}(.)$ is a monotonic function of the best-proxy statistic threshold, s^*. Therefore, there exists a unique ρ_i such that $\mathrm{ER}^{-1}(\rho_i) = s^*$, where $\mathrm{ER}^{-1}(.)$ is the inverse of the expected recall function. Using this property, the problem can be simplified by linearizing the constraint function, which reads

$$\begin{aligned} &\text{minimize} \quad \sum_{i=K+1}^{M} \Pr\left(|S_{b(i)}| > \mathrm{ER}^{-1}(\rho_i)\right), \\ &\text{such that} \quad \sum_{i=K+1}^{M} \rho_i = \rho^*. \end{aligned} \tag{3.13}$$

Note that

$$\frac{\partial}{\partial \rho} \mathrm{ER}^{-1}(\rho) = \frac{1}{\mathrm{ER}^{'}\left(\mathrm{ER}^{-1}(\rho)\right)}, \tag{3.14}$$

hence the derivative of the expected number of test from a single remainder SNP with respect to ρ follows,

$$g = \frac{\partial}{\partial \rho} \Pr\left(|S_t| > \mathrm{ER}^{-1}(\rho)\right) = \frac{-f(\mathrm{ER}^{-1}(\rho)) - f(-\mathrm{ER}^{-1}(\rho))}{\mathrm{ER}^{'}\left(\mathrm{ER}^{-1}(\rho)\right)}. \tag{3.15}$$

Using the method of Lagrange multipliers, it can be shown that at the optimum solution the expected number of tests from each remainder SNP has the same derivative value, g^*. In GRAT, we determine g^* by using binary-search such that for each remainder SNP m_i, g^* uniquely maps to ρ_i^*, where $\sum \rho_i^* = \rho^*$.

3.7 Performance on a Single SNP Pair

We apply the proposed method to a pair of SNPs, a causal SNP and non-causal proxy SNP, to verify whether or not the target sensitivity is reached for any value of the pairwise correlation. For each value of the correlation, we sample thousands of joint statistics for the SNP pair and record how many times the causal SNP is significant. The power at the causal SNP is set to $\mathcal{P}_c = 50\%$ using a genome-wide significance level of $\alpha = 10^{-8}$.

We compute the threshold of the proxy SNP statistic for different target sensitivities in each pairwise correlation using a small prior probability for the causal SNP, $c_i =$

10^{-5}. In each correlation value, we apply the decision rules to the samples and record the recall rate of significant causal SNP in each target sensitivity.

In Figure 3.1, the observed recall rates are shown for different values of target sensitivity and pairwise correlation. The target sensitivities are shown as horizontal lines and are followed closely by the observed recall rates. The variation around a target value is due to the asymptotic distribution of the test statistic, and diminishes as the sample size increases.

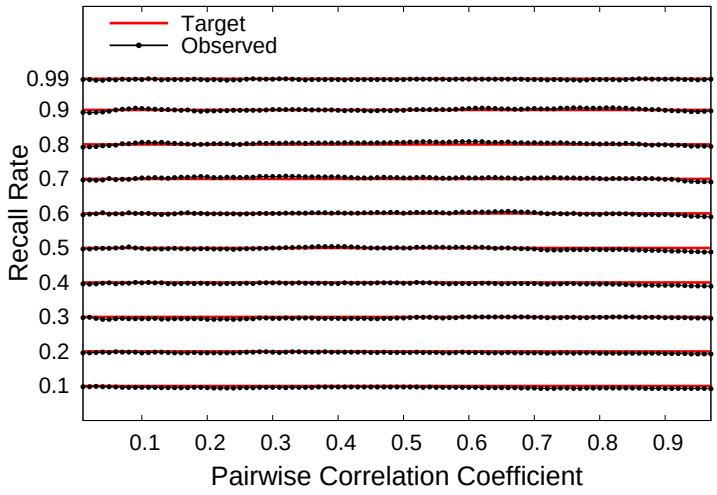


Figure 3.1: Performance of the method using a single pair of SNPs. The observed recall rate of the significant causal SNP is shown for different target sensitivity and pairwise correlation values.

3.8 Choosing the Optimal Proxy SNPs

The expected number of SNPs to be tested (ENT) in the two-stage approach depends on the number of proxy SNPs and which SNPs are chosen as proxies. It can be shown that the problem of finding the optimal set of proxy SNPs, among all possible sets of proxy SNPs, the set of proxy SNPs that gives the minimum ENT is an NP-Hard

problem. Therefore, we propose a heuristic algorithm for choosing the proxy SNPs using a greedy approach, which incrementally builds the set of proxy SNPs.

Starting with an empty set, let $\mathcal{T}_k$ denote the current set of proxy SNPs with size k, where ENT_k and ERR_k denote the values of its ENT and ERR. ($\mathrm{ENT}_0 = +\infty$ and $\mathrm{ERR}_0 = -\infty$). Each remainder SNP m_i is a candidate to extend the current set of proxy SNPs to become $\{\mathcal{T}_k \cup m_i\}$, which performs $\mathrm{ENT}_{k+1}^{(i)}$. The remainder SNP with the least $\mathrm{ENT}_{k+1}^{(i)}$ is chosen for extending the current set of proxy SNPs:

$$\mathcal{T}_{k+1} = \mathcal{T}_k \cup \underset{m_i \in \mathcal{M} \setminus \mathcal{T}_k}{\operatorname{argmin}} \left(\mathrm{ENT}_{k+1}^{(i)} \right). \tag{3.16}$$

While the extended set $\mathcal{T}_{k+1}$ improves the ENT, i.e., $\mathrm{ENT}_{k+1} < \mathrm{ENT}_k$, the algorithm continues.

For each candidate set of proxy SNPs, the algorithm solves the optimization problem (3.7) to compute $\mathrm{ENT}_{k+1}^{(i)}$. This leads to a quadratic computational complexity in the number of collected SNPs and in practice makes it hard to scale to large numbers. We further introduce a heuristic extension to the above greedy-approach to reduce this complexity. While extending the current set of proxy SNPs $\mathcal{T}_k$ to $\mathcal{T}_{k+1}$, the optimization problem (3.7) is solved $M - k$ times. In particular, solving the optimization problem (3.7) corresponds to finding the gradient, g^*, at which the ENT function is minimized while satisfying the constraints (see Appendix). We assume that for $\mathcal{T}_k$ and $\mathcal{T}_{k+1}$ the gradient values of their ENT functions are close enough, $g_k^* \approx g_{k+1}^*$. Therefore, while extending the current proxy set, we compute the ENT of each candidate set, $\mathrm{ENT}_{k+1}^{(i)}$, using the gradient value from the previous step, g_k^*. This way, rather than solving the optimization problem $M - k$ times for each possible proxy SNP at each step k, the gradient is updated once after the new set $\mathcal{T}_{k+1}$ is determined. Using this approach the optimization problem (3.7) is solved a total of K times, where K is the size of the final set of proxy SNPs.

3.9 Updating the Remainder SNP Thresholds in Linear Mixed Models

We consider the following linear mixed model (LMM) formulation,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{g} + \boldsymbol{e}, \tag{3.17}$$

where $\boldsymbol{y}$ is the $(n \times 1)$ vector of phenotypic values, $\boldsymbol{X}$ is the $(n \times p)$ matrix of fixed-effects, which includes the mean, covariates and the SNP to be tested, $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of fixed-effect weights, $\boldsymbol{g}$ is the variance component accounting for the population structure and $\boldsymbol{e}$ is the iid noise. We assume the random effects, $\boldsymbol{g}$ and $\boldsymbol{e}$, follow multivariate normal distribution, $\boldsymbol{g} \sim \mathcal{N}\left(0, \sigma_g^2 \boldsymbol{K}\right)$, $\boldsymbol{e} \sim \mathcal{N}\left(0, \sigma_e^2 \boldsymbol{I}\right)$, where $\boldsymbol{K}$ is the known, $(n \times n)$, genetic similarity matrix, $\boldsymbol{I}$ is the $(n \times n)$ identity matrix with unknown magnitudes σ_g^2 and σ_e^2. We follow the approach taken in EMMAX[KSS10] and estimate σ_g^2 and σ_e^2 in the null model, with no SNP effect, and use these parameters while testing the SNPs. That is, when each SNP is tested, the covariance of $\boldsymbol{y}$ is kept fixed, $\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{\Sigma} = \hat{\sigma}_g^2 \boldsymbol{K} + \hat{\sigma}_e^2 \boldsymbol{I}$, where $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ are the restricted log likelihood (REML) estimates[KSS10, LLL11].

In GRAT, the threshold value for each remainder SNP is computed after the covariance matrix $\boldsymbol{\Sigma}$ is estimated and the alternate model is transformed by the inverse square root of this matrix,

$$\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2 \boldsymbol{I}\right), \tag{3.18}$$

where the residuals are iid. For two SNPs m_i and m_j, let $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ be their $(n \times 1)$ allelic indicator vectors. When the SNPs are tested individually in the above model, the same transformation is applied to the genotype vectors, which may moderately change the pairwise correlation between the SNPs. The transformed genotype vectors

are $\tilde{\boldsymbol{x}}_i = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_i$ and $\tilde{\boldsymbol{x}}_j = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_j$ and their correlation coefficient is,

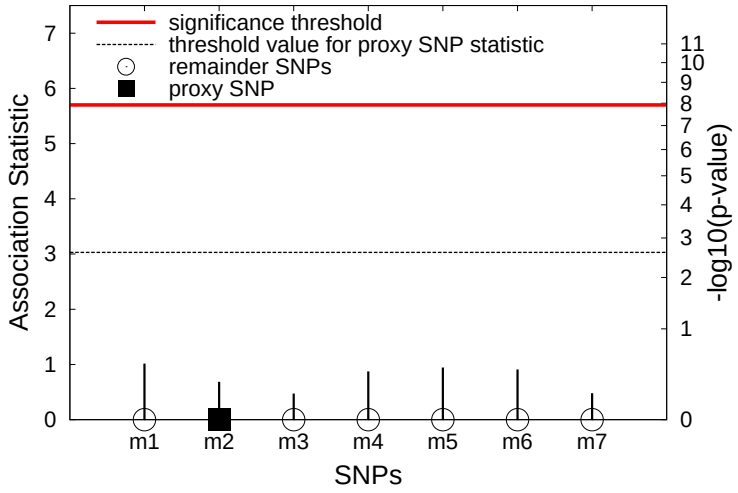
$$\tilde{r}_{ij} = \frac{\mathrm{Cov}(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j)}{\sqrt{\mathrm{Var}(\tilde{\boldsymbol{x}}_i)}\sqrt{\mathrm{Var}(\tilde{\boldsymbol{x}}_j)}}. \tag{3.19}$$

3.10 GRAT: Genome-wide Rapid Association Testing

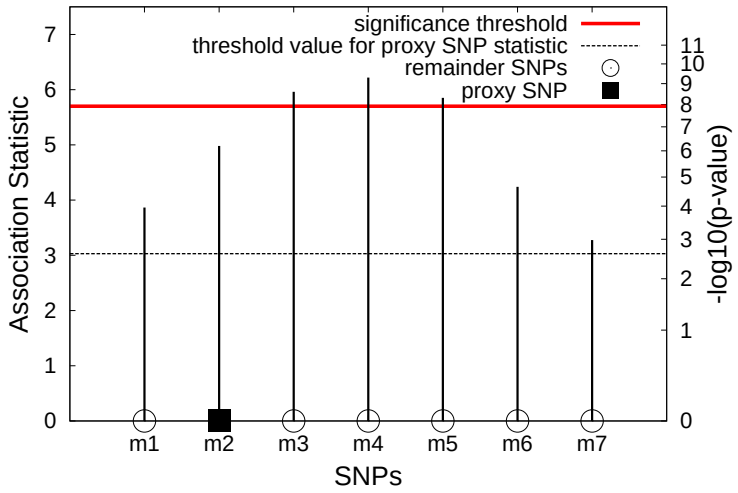
In Figure 3.2, we consider two possible scenarios for a genomic region in a GWAS. In (a) the region contains no significant associations and in (b) the region contains a causal SNP. In (a) and (b), the statistics for each SNP are shown, denoting what could have been observed in each scenario had all the SNPs in the region been tested. Let m_2 be the proxy SNP for this region to decide whether or not to test the rest of the SNPs. We refer to the SNPs other than the proxy SNP (m_1, m_3, m_4, m_5, m_6 and m_7) as the "remainder SNPs". If the observed statistic of the proxy SNP is stronger than a threshold value, which in this example is 3.0, the remainder SNPs are tested.

In the first-stage, only the proxy SNP is tested and its association statistic is observed. In (a), where the region contains no associations, the statistic of the proxy SNP is 0.7. The observed statistic of the proxy is less than the threshold value ($0.7 < 3.0$) and hence none of the remainder SNPs within the region are tested. In (b), the region contains associations and the proxy SNP captures this information. The observed statistic of the proxy SNP is stronger than the threshold value ($5.0 > 3.0$), which leads to testing each of the remainder SNPs in the region. This results in identifying all the significant SNPs (m_3, m_4, m_5).

In Methods, we introduce a novel approach for choosing the proxy SNPs and the threshold values, which provide guarantees that all statistically significant associations will be discovered while computing the least amount of association tests. Due to the complexity of linkage disequilibrium (LD) across the genome, we use a separate threshold value for each remainder SNP rather than using a common threshold value



(a) A region with no associations.



(b) A region with significant associations.

Figure 3.2: An example of applying GRAT in two hypothetical regions. First, the proxy SNP (rectangle) is tested and its statistics is compared to the threshold (dashed line). If the statistic is above the threshold, the remaining SNPs in the region are tested.

for all the remainders SNPs in an LD region. This is performed by pairing each remainder SNP with its most strongly correlated proxy SNP and a threshold value is used for the pair to decide whether or not to test the remainder SNP. We have precomputed

the proxy SNPs for the 1000 Genomes Project and studies imputing to SNPs in this reference can benefit from our method. Even though the LD structure among the SNPs in the study and the reference dataset may be different, our method guarantees to discover all significant associations with high-probability. This is achieved by updating the threshold values using the LD structure observed in the study. We term our novel two-stage testing procedure as Genome-wide Rapid Association Testing (GRAT).

GRAT can be applied to a wide range of statistical models, such as case-control studies, quantitative traits and linear mixed models (LMM). In particular, the LMM approach has recently become popular due to its effective control of population structure. Computing the LMM association statistic is computationally expensive and recently its efficient computation has attracted great interest[KSS10, LLL11, ZS12]. The speed-up due to GRAT is cumulative with these efforts.

3.11 Application to a Large-scale eQTL Study

We compared the performance of GRAT to the standard approach of testing all the SNPs using a large-scale eQTL study[SMD12] that contains $47,292$ gene expression traits on 80 HapMap ASN (East Asian ancestry) individuals that are fully sequenced in the 1000 Genomes Project[The10]. We obtained the genotype data from the MACH website[LWD10] and retained approximately 5.9 million SNPs that are filtered for Hardy-Weinberg equilibrium (HWE) and minor allele frequency (MAF) greater than 5%. We eliminated SNPs with lower MAF frequency since they could not be genome-wide significant due to the sample size.

We performed the standard analysis using PLINK[PNT07] which took approximately 2600 hours. We used a conservative genome-wide significance threshold level, $\alpha = 10^{-8}$, to label the significant SNPs and observed $85,219$ significant associations.

We repeated the association analysis by applying GRAT using the proxy SNPs precomputed for the 1000 Genomes Project ASN population SNPs. The number of proxies is $276,702$, which means GRAT tests approximately 5% of the SNPs in the first stage.

Applying GRAT to the whole eQTL dataset took 35 hours using the same computational resources (single core of an Opteron CPU). In addition to the proxies, GRAT tested 8.5% of the SNPs in the second stage, reducing the computational cost down to analyzing 13.5% of all the SNPs with the rest of the speedup coming from a faster implementation compared to PLINK. GRAT identified all of the significant associations and speeded up the computation by a factor of 75.

3.12 Application to Linear Mixed Model Association

We applied GRAT to a linear mixed model (LMM) association of the eQTL dataset. A challenge in applying GRAT to LMMs is that GRAT utilizes the fact that the joint distribution of traditional association statistics for correlated markers is directly dependent on the correlation between the markers as shown in Pritchard & Przeworski[PP01]. Unfortunately, when applying LMMs, this relation no longer holds. We derive an analogous relationship between LMM statistics that takes into account both the correlation between the markers and the kinship matrix. Utilizing this relationship, we apply GRAT to LMMs using an efficient implementation[LLL11].

We performed the standard analysis, testing each SNP in each expression trait, which identified $66,818$ significant associations ($\alpha = 10^{-8}$). We applied GRAT using the proxy SNPs precomputed for the 1000 Genomes Project ASN population. In two-stages, GRAT statistically tested a total of 9.1% of the SNPs, identifying all of the significant associations, demonstrating that GRAT can speed up LMM association by a factor of 10.

3.13 Simulations Using the 1000 Genomes Project

To obtain a more robust estimate of the performance, we applied GRAT to thousands of simulated GWAS studies. We simulated the studies using common SNPs (minor allele frequency $> 5\%$) available from the 1000 Genomes Project[The10] using the phased SNP genotypes obtained from the MACH website[LWD10] on four populations: African (AFR), East Asian (ASN), Ad Mixed American (AMR) and European (EUR) ancestries.

We divided each chromosome into panels of 1000 SNPs and simulated case-control GWASs by randomly selecting 5% of the panels as the alternate panels, in which we simulated a causal SNP, and the remaining panels as the null panels, without any causal SNPs. In each alternate panel, we randomly selected the causal SNP and set its statistical power to be $\mathcal{P}_c = 50\%$ at the significance level $\alpha = 10^{-8}$. Using this procedure, we simulated 500 GWASs in each population.

We applied GRAT to each simulated GWAS and recorded the recall rate of the significant SNPs and total number of tests performed. In Table 3.1, we show the performance of GRAT in each population averaged over the simulations. GRAT practically identified all significant associations and reduced the number of tests by 10 folds. Across the simulations, from the total $3,718,126$ significant associations GRAT only missed 1052 significant associations.

3.14 Comparison to Tradition tag-SNP Based Association Testing

Choosing an informative subset of the SNPs, termed tag-SNPs, under various criteria has been extensively investigated[Str04, BYP05, Str05, CDG06, CGM03, HKS05, LA04, PLW05, QGA06, SRS06, CER04]. The main goal of these methods is to reduce the cost of GWASs by genotyping a subset of the SNPs, yet collect as much

Table 3.1: Performance in simulations

Population	Number of SNPs	Recall Rate	Reduction
AFR	8.5×10^6	$> 99.9\%$	88.2%
AMR	6.7×10^6	$> 99.9\%$	92.4%
ASN	6.1×10^6	$> 99.9\%$	92.8%
EUR	6.6×10^6	$> 99.9\%$	92.6%

The average performance of GRAT in 500 simulated GWASs using 1000 Genomes Project data in four populations. GRAT identified practically all significant associations by only testing 10% of the SNPs.

information as possible on the remaining SNPs.

We mimic a two-stage association testing approach using a traditional tag-SNP selection method and compare its performance to GRAT. In the first stage, we test all the tag-SNPs and use a p-value threshold, α_{tag}, to choose which of the tag-SNPs to follow. If the p-value of a tag-SNP is more stronger than the threshold, the remainder SNPs tagged by this tag-SNP are tested.

We simulated association studies using the 10 HapMap ENCODE regions, which are densely genotyped for four HapMap populations[The04]. In each simulation study, we used the ENCODE regions to generate null regions that harbor no causal SNPs and alternate regions each harboring a causal SNP with 50% statistical power at the genome-wide significance level of $\alpha = 10^{-8}$. Following this approach, we generated 500 association studies in each population.

In each region and in each population, we identified the tag-SNPs using the widely utilized tag-SNP selection method Tagger[BYP05]. Given a set of SNPs and information on their minor allele frequencies and pairwise correlation coefficients, Tagger

selects the minimum number of tag-SNPs such that each of the remaining SNPs correlates to a tag-SNP with a minimum r^2 pairwise correlation value. In our evaluations, we have used the default value of $r^2 = 0.8$. In order to perform a comparison, we also applied GRAT to identify the proxy SNPs and the statistic threshold rules for testing the remainder SNPs to achieve 99% target recall rate on the significant associations.

In Table 3.2 the performance of GRAT is compared to Tagger in four HapMap populations using various p-value threshold values, $\alpha_{\text{tag}} = \{10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}\}$. In each population, GRAT achieved more than 99% recall rate, while testing approximately 10% of all SNPs. Among all the p-value threshold values used, the traditional tag-SNPs led to testing more than twice the number of SNPs tested by GRAT and only achieved the target recall rate in all populations when the p-value threshold value was $\alpha_{\text{tag}} = 10^{-5}$. Unfortunately, Tagger –unlike GRAT– does not guarantee a recall rate so it is not clear how to set the threshold and be certain that no associations are missed.

3.15 Discussion

In the genome-wide association study (GWAS), information on single-nucleotide polymorphisms (SNPs) across the genome is collected from thousands of case and control individuals. Typically, each SNP is tested individually for disease association and the significant SNPs provide insight into the genetics of the disease. Association studies attempt to collect information on as many SNPs as possible to cover the whole genome. However, as the number of collected SNPs increases so does the computational burden to identify the significant associations

We introduced a novel method, GRAT, for genome-wide rapid association testing to identify all significant associations by testing a small subset of the SNPs. Due to the correlation, or linkage disequilibrium (LD), testing a SNP provides information about

Table 3.2: Performance of GRAT and Tagger in ENCODE simulations

Method	CEU			CHB		
	Recall	Reduction	Speedup	Recall	Reduction	Speedup
GRAT	99.89%	89.7%	9.7×	99.73%	89.6%	9.6×
Tagger α_{tag}=1e-8	86.25%	78.9%	4.7×	87.78%	79.7%	4.9×
Tagger α_{tag}=1e-7	95.74%	78.6%	4.7×	97.70%	79.4%	4.8×
Tagger α_{tag}=1e-6	98.40%	78.3%	4.5×	99.62%	79.0%	4.8×
Tagger α_{tag}=1e-5	99.30%	77.8%	4.5×	99.97%	78.4%	4.6×
Method	JPT			YRI		
	Recall	Reduction	Speedup	Recall	Reduction	Speedup
GRAT	99.63%	90.2%	10.2×	99.72%	88.4%	8.6×
Tagger α_{tag}=1e-8	88.53%	80.5%	5.1×	87.62%	65.3%	2.9×
Tagger α_{tag}=1e-7	98.10%	80.1%	5.0×	97.55%	65.3%	2.9×
Tagger α_{tag}=1e-6	99.52%	79.6%	4.9×	99.39%	65.1%	2.9×
Tagger α_{tag}=1e-5	99.92%	79.1%	4.8×	99.94%	65.0%	2.9×

In each HapMap population, the average performance of GRAT and Tagger in 500 simulated GWASs are shown. GRAT guarantees to achieve the 99% target recall rate, while reducing the number of tests by 90%. Using Tagger, we test the remainder SNPs that are tagged by the tag-SNPs that exceed a p-value cut-off threshold, α_{tag}. GRAT outperforms the traditional tag-SNPs in all populations.

the associations of its neighboring SNPs. Using this intuition, the procedure first tests a subset of the SNPs, referred to as the proxy SNPs, across the genome to locate the regions that may contain the significant associations. Once located, additional SNPs are tested from those regions to identify the significant SNPs. Each unobserved, or remainder, SNP is paired with its most strongly correlated proxy SNP, termed best-

proxy, and a threshold value is used for the best-proxy's statistic to decide whether or not to test the unobserved SNP. We introduced a novel approach to choose the proxy SNPs and determine the threshold values for each best-proxy SNP. Through simulations and real GWAS data we showed that the proposed approach can identify more than 99% of the significant SNPs by reducing the number of tests by a factor of 10. Furthermore, GRAT can also be applied to association studies that utilize linear mixed models, where the speed-up is cumulative with recent efforts that decrease the computational burden of computing the actual association statistic.

CHAPTER 4

Improving the Accuracy and Efficiency of Partitioning Heritability into the Contributions of Genomic Regions

4.1 Motivation

Partitioning heritability into the contributions of genomic regions is a general problem with many applications. These include providing insights into the genetic architecture of a trait[YMP11], quantifying the amount of population structure in a study cohort[YMP11, LDR12], quantifying the effect of genes with a certain functional annotation compared to other genes[LDR12], quantifying the phenotypic variation explained by genic versus intergenic regions[YMP11], and estimating the amount of variation corresponding within a previously identified QTL loci compared to the rest of the genome[SFH12]. The current approach, implemented in the widely used method, GCTA[YLG11], estimates the heritability contributions of all regions jointly. However, this approach becomes computationally intractable when the number of regions is large and the regions themselves are smaller than chromosomes.

In this chapter, we present an alternative approach for estimating the contributions of an arbitrary number of regions to the total heritability. Using a linear mixed model approach, we estimate the heritability contribution of each region separately. For each region, we partition the total heritability into the contribution of the region and its genomic complement and repeat this procedure for all regions similar to the approach

presented in Hayes et al.[HPC10]. An advantage of our approach is that in addition to performing the computations in parallel, we also take advantage of spectral decomposition to efficiently estimate the variance components[PT71, KZW08]. Overall, our approach is more efficient especially when the number of regions increases.

Using simulations, we demonstrate that our proposed approach is more accurate than the current approach, when the number of regions is over 100. In our simulations, we partition the genome into a large number of regions and consider different scenarios of heritability contributions from these regions. We further apply the proposed and current approaches to a large-scale GWAS dataset on human height[SSH08]. Both approaches estimate the same genome-wide heritability for height, 62%, but estimate different heritability contributions from genomic regions. We also use our approach to estimate the proportion of the observed heritability accounted for due to population structure.

4.2 Preliminaries

Using a linear mixed model, we attribute the phenotypic variation to additive-genetic effects and the environment,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_f + \sum_{i \in \mathcal{G}} \beta_i \boldsymbol{x}_i + \boldsymbol{e}, \tag{4.1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of the observed phenotypic values from n individuals, $\boldsymbol{X}$ is the $n \times p$ covariate matrix including the intercept, and $\boldsymbol{\beta}_f$ is the $p \times 1$ vector of the (unknown) fixed effect parameters including the population mean. Let $\mathcal{G}$ denote the set of the observed SNPs across the genome. $\boldsymbol{x}_i$, $i \in \mathcal{G}$, denotes the $n \times 1$ normalized genotype vector of SNP i, with effect β_i, where the components of $\boldsymbol{x}_i$ are encoded as $\{\frac{2-2p_i}{\sqrt{2p_i(1-p_i)}}, \frac{1-2p_i}{\sqrt{2p_i(1-p_i)}}, \frac{-2p_i}{\sqrt{2p_i(1-p_i)}}\}$ for minor allele homozygous, heterozygous and major allele homozygous, where p_i is the observed minor allele frequency. We follow

the standard approach and assume that the SNP effects are independent and normally distributed, where $\beta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Finally, $\boldsymbol{e}$ denotes the $n \times 1$ vector of environmental effects in the trait, where $\boldsymbol{e} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I})$ and $\boldsymbol{I}$ is the identity matrix. This model can be succinctly expressed as,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_f + \boldsymbol{g} + \boldsymbol{e}, \tag{4.2}$$

where $\boldsymbol{g}$ is the $n \times 1$ vector of genetic effect such that $\boldsymbol{g} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{K})$ and $\boldsymbol{K} = \sum_{i \in \mathcal{G}} \boldsymbol{x}_i \boldsymbol{x}_i^T$. The narrow-sense heritability accounted by the genetic effect is defined and can be estimated as,

$$\begin{aligned} h_g^2 &= \frac{\mathrm{Var}(\boldsymbol{g})}{\mathrm{Var}(\boldsymbol{g} + \boldsymbol{e})}, \\ \hat{h}_g^2 &= \mathrm{E}[h_g^2] \approx \frac{\hat{\sigma}^2 \,\mathrm{Tr}(\boldsymbol{P}\boldsymbol{K})}{\hat{\sigma}^2 \,\mathrm{Tr}(\boldsymbol{P}\boldsymbol{K}) + n\hat{\sigma}_e^2}, \end{aligned} \tag{4.3}$$

where $\boldsymbol{P} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$, and $\mathrm{Tr}(.)$ denotes the matrix trace.

Given a genomic region defined by the set of SNPs $\mathcal{R}$, $\mathcal{R} \subset \mathcal{G}$, we estimate the contribution of the region to the heritability using the following model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_f + \sum_{r \in \mathcal{R}} \beta_r \boldsymbol{x}_r + \sum_{b \in \mathcal{G} \setminus \mathcal{R}} \beta_b \boldsymbol{x}_b + \boldsymbol{e}, \tag{4.4}$$

where $\beta_r \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_r^2)$ and $\beta_b \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2)$. Equivalently,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_f + \boldsymbol{r} + \boldsymbol{b} + \boldsymbol{e}, \tag{4.5}$$

where $\boldsymbol{r}$ and $\boldsymbol{b}$ are $n \times 1$ vectors denoting the genetic effects due to the region and its genomic background. Both genetic effects follow a multivariate normal distribution, $\boldsymbol{r} \sim \mathcal{N}(0, \sigma_r^2 \boldsymbol{K}_r)$ and $\boldsymbol{b} \sim \mathcal{N}(0, \sigma_b^2 \boldsymbol{K}_b)$, where $\boldsymbol{K}_r = \sum_{\boldsymbol{x}_r \in \mathcal{R}} \boldsymbol{x}_r \boldsymbol{x}_r^T$ and $\boldsymbol{K}_b = \sum_{\boldsymbol{x}_b \in \mathcal{G} \setminus \mathcal{R}} \boldsymbol{x}_b \boldsymbol{x}_b^T$ are the realized genetic relationships matrices (GRMs) among the individuals for the region and the genetic background calculated from SNP data, respectively. We note that inherent to our model is the assumption that $\mathrm{Cov}(\boldsymbol{r}, \boldsymbol{b}) = 0$,

i.e., $\sum_{\forall \boldsymbol{x}_r, \forall \boldsymbol{x}_b} \boldsymbol{x}_r \operatorname{Cov}(\beta_r, \beta_b) \boldsymbol{x}_b^T = 0$, since $\operatorname{Cov}(\beta_r, \beta_b) = 0$. We discuss this assumption in more detail in the discussion. Finally, the phenotype vector follows a multivariate normal distribution,

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}_f\,,\, \sigma_r^2 \boldsymbol{K}_r + \sigma_b^2 \boldsymbol{K}_b + \sigma_e^2 \boldsymbol{I}\right). \tag{4.6}$$

The total covariance due to additive genetics can be expressed as, $\sigma_r^2 \boldsymbol{K}_r + \sigma_b^2 \boldsymbol{K}_b = \sigma_g^2 \boldsymbol{K}_g$, where $\boldsymbol{K}_g = \omega \boldsymbol{K}_r + (1-\omega)\boldsymbol{K}_b$. We determine the unknown variance scalars $\sigma_r^2 = \omega\sigma_g^2$, $\sigma_b^2 = (1-\omega)\sigma_g^2$ and σ_e^2 by using the following approach that iterates over the parameter ω. For given ω, we transform the model to a coordinate system where the covariance matrix $\sigma_g^2 \boldsymbol{K}_g + \sigma_e^2 \boldsymbol{I}$ is diagonal, which lets us find the maximum likelihood parameters efficiently by speeding up the computationally expensive matrix inversion. We use the spectral transformation, $\boldsymbol{Q}^T$, where $\boldsymbol{K}_g = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$ is the eigendecomposition of the covariance structure of the genetic effect. The spectrally transformed model follows,

$$\tilde{\boldsymbol{y}} \sim \mathcal{N}\left(\tilde{\boldsymbol{X}}\boldsymbol{\beta}_f,\ \sigma_g^2 \boldsymbol{\Lambda} + \sigma_e^2 \boldsymbol{I}\right), \tag{4.7}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{Q}^T\boldsymbol{y}$ and $\tilde{\boldsymbol{X}} = \boldsymbol{Q}^T\boldsymbol{X}$. We estimate the unknown variance parameters using the restricted log-likelihood that takes into account the loss in degrees of freedom which results from estimating the fixed effect parameters[Har74, Har77],

$$\begin{aligned}\mathcal{LL}_R\left(\sigma_g^2, \sigma_e^2 \middle| \omega\right) \;= -\frac{1}{2}\Bigg(&(n-p)\log(2\pi) - \log\left(|\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}|\right) + \log\left(|\tilde{\boldsymbol{X}}^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{X}}|\right) \\ &+ \log\left(|\boldsymbol{V}|\right) + \left(\boldsymbol{y} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_f\right)^T \boldsymbol{V}^{-1}\left(\boldsymbol{y} - \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}_f\right)\Bigg),\end{aligned} \tag{4.8}$$

where $\boldsymbol{V} = \sigma_g^2\boldsymbol{\Lambda} + \sigma_e^2\boldsymbol{I}$ and $\hat{\boldsymbol{\beta}}_f = \left(\tilde{\boldsymbol{X}}^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{y}}$. In particular, the restricted log-likelihood can be expressed as an analytic function of $\delta = \sigma_e^2/\sigma_g^2$ and solved using Brent's method to determine the global maximum likelihood[PT71, LW98, KZW08, LLL11].

Finally, the heritability contribution of the region is calculated as,

$$\hat{h}_r^2 = \frac{\hat{\sigma}_r^2 \operatorname{Tr}(\boldsymbol{P}\boldsymbol{K}_r)}{\hat{\sigma}_g^2 \operatorname{Tr}(\boldsymbol{P}\boldsymbol{K}_g) + n\hat{\sigma}_e^2}. \tag{4.9}$$

4.3 Normalization of the heritability contributions

One of the difficulties of partitioning the heritability into the contributions of genomic regions is the linkage disequilibrium (LD) structure of the genome. When HEIDI estimates the heritability contribution of a region, the background model can inadvertently capture a portion of the heritability due to the inclusion of markers in the background genetic relationship matrix which are in LD with markers in the region.

We utilize the following normalization procedure to improve the accuracy of the estimates obtained from HEIDI, which mitigates the effect of LD. First, we estimate the total heritability, followed by estimating the contributions of the autosomes and scaling their contributions such that their sum equals to the total heritability. The advantage of these estimates is that they are not affected by LD. We then we estimate the heritability contributions of the regions and in each chromosome, normalize the regions' contributions such that their sum equals to the normalized chromosomal contribution.

4.4 Simulation model

In our simulations, we partitioned each of the 22 autosomes into 5 regions of equal number of SNPs and use the following model to generate phenotype values,

$$\boldsymbol{y} = \mathbf{1}\mu + \sum_{i=1}^{110} \boldsymbol{r_i} + \boldsymbol{e}, \tag{4.10}$$

where $\boldsymbol{r_i} \sim \mathcal{N}(0, \sigma_i^2 \boldsymbol{K}_i)$, $i \in \{1, \ldots, 110\}$, are the variance components accounting for the chromosome regions, each with the covariance matrix $\boldsymbol{K}_i$, and $\boldsymbol{e} \sim \mathcal{N}(0, \sigma_e^2 \boldsymbol{I})$ is the error term. The total heritability and the contributions of the regions in the total heritability is determined by choosing the variance scalars, σ_i^2's and σ_e^2, accordingly. Finally, phenotype values are generated by sampling from the corresponding multivariate normal distribution,

$$\boldsymbol{y}_{\text{sim}} \sim \mathcal{N}\left(\mathbf{1}\mu\,, \sum_{i=1}^{110} \sigma_i^2 \boldsymbol{K}_i + \sigma_e^2 \boldsymbol{I}\right). \tag{4.11}$$

4.5 Summary of the Real GWAS Data

We used the height measurements and the genotype data available from the Northern Finland Birth Cohort 1966 (NFBC66)[SSH08], from $5,319$ unrelated individuals that were phenotyped for height at age 31. We adjusted the height measurements for sex. The dataset contains $331,450$ autosomal SNPs after applying the exclusion criteria of Hardy-Weinberg equilibrium ($P < 10^{-4}$), genotyping completeness ($< 95\%$) and minor allele frequency ($< 1\%$).

4.6 HEIDI is more accurate than the current approach

We compare HEIDI to the widely used method, GCTA[YLG11], which partitions the heritability into the contributions of genomic regions. Using a linear mixed model, GCTA assigns a variance component to each region and jointly estimates their heritability contributions.

Our approach estimates these contributions separately for each region in two stages. First, we partition the phenotypic variation to that attributable to the total additive genetic effect of the whole-genome and the environment. This allows us to estimate

the genome-wide heritability, as well as predicting the total genetic effect. Next, we estimate the heritability contribution of a region by breaking the total genetic effect into the effects of the region and the rest of the genome. We repeat this step for each region separately, which can be performed in parallel to improve the computational efficiency. Finally, the contributions of the chromosomes are normalized to that of the total heritability and the contributions of regions in each chromosome to that of the normalized contribution of the chromosome.

We performed simulations using the genotype data available from the Northern Finland Birth Cohort[SSH08]. The data consists of $331,450$ common SNPs that passed various exclusion criteria on $5,319$ unrelated individuals (see Methods). We partitioned each of the 22 autosome into 5 regions, where in each chromosome the regions contain an equal number of SNPs. We utilize the same number of markers per region to minimize any possible bias due to a different number of markers being used to compute the genetic relationship matrices (GRMs) in each chromosome. We use the same approach as in the GCTA software to estimate the GRMs. For our simulations, we generate phenotypes by sampling from a multivariate normal distribution with a covariance matrix which is the environmental noise plus the sum of the GRMs wighted by their region contributions in the heritability. We generated three panels of simulations, each with 100 replicates. In these simulations we set the genome-wide heritability to 50%.

In the first panel, each region has the same the heritability contribution. In the second panel, chromosomes contribute to the heritability proportional to their sizes with contributions ranging from 6.96% to 0.32% of the total heritability, and in each chromosome, the contributions of the 5 regions range from 32% to 8% of the chromosome's contribution itself. In this scenario, there is a wide range of heritabilities for different regions. In the third panel, we simulate a scenario where chromosome regions 2 and 4

do not contribute to the heritability, but regions 1,3 and 5 have equal contributions.

We compare the accuracy of the two methods across the simulations with respect to their mean absolute error (MAE). In a region, the absolute error is the magnitude of the difference between the estimated heritability contribution of the region and its true value. In each method, we obtain the absolute errors of the regions in each simulation replicate. MAE is calculated per region by averaging the absolute errors over the replicates. Note that, the unit of MAE is heritability. In the first simulation panel, where the regions have the same contribution, the average MAE values are 0.33 for HEIDI and 0.37 for GCTA. In the second panel, where the contributions change with chromosome size, the average MAE values are 0.32 for HEIDI and 0.34 for GCTA. In the third panel, where the contributions are sparse, the average MAE values are 0.32 for HEIDI and 0.34 for GCTA. Our results suggest that on average HEIDI is over 5% more accurate than GCTA in partitioning the total heritability. Figures 4.1, 4.2 and 4.3 show the MAE of each method in each region. In terms of computation, HEIDI can estimate the contribution of each region in parallel, which significantly facilitates estimating the heritability contributions of many regions.

4.7 Partitioning the heritability of human height and estimating the contribution of population structure

We applied both approaches to the height phenotype collected from the Northern Finland Birth Cohort GWAS data[SSH08]. We estimated the total heritability, the contributions of the autosomes and their partitions accounted for by common variants in the genome. We use the same approach to estimate the GRMs as in the GCTA software.

For the unpartitioned genome, both approaches estimated 62.4% heritability, which is expected since they use the same underlying model. We partitioned the genome-

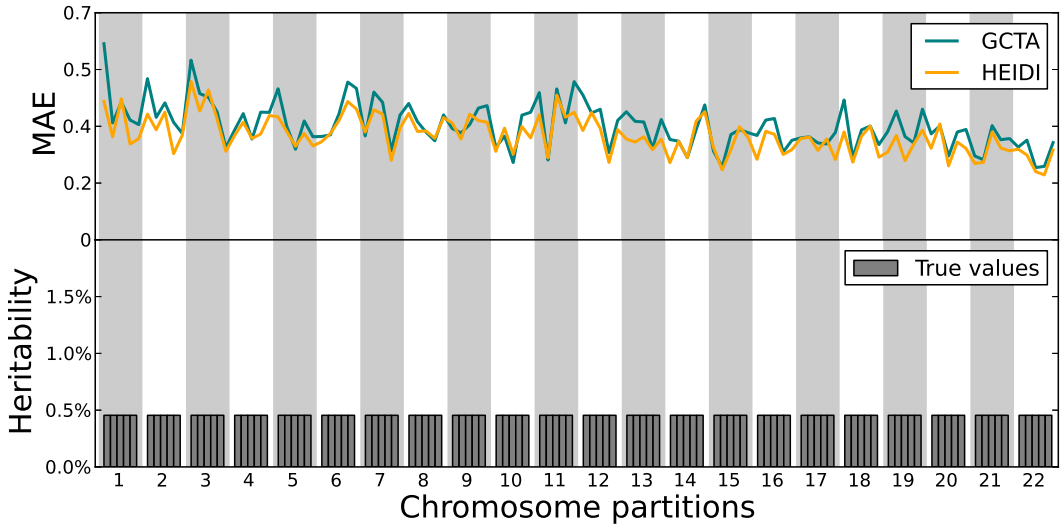


Figure 4.1: Mean absolute error values obtained by HEIDI and GCTA are shown in each region in the simulations where the total heritability is 50% and each region has the same heritability contribution. In this scenario, the accuracy of HEIDI is 8.76% higher than the accuracy of GCTA.

wide heritability into the contributions of the 22 autosomes, shown in Figure 4.4. Both methods estimated similar contributions with high concordance. However, when partitioning the heritability into contributions from smaller regions, differences between GCTA and HEIDI emerged. Figure 4.5 shows the heritability contributions of each autosome split into 5 regions of equal number of SNPs.

One of the main applications of partitioning the heritability into contributions from genomic regions is that it allows us to estimate the contribution of population structure to the observed heritability. Both population structure and cryptic relatedness are known to increase estimates of heritability[BB11, YMP11]. We compare HEIDI and GCTA in performing the analysis described in Yang et al.[YMP11], where they partitioned the heritability into chromosomes and compared these estimates to the estimates of the heritability using only one variance component corresponding to the

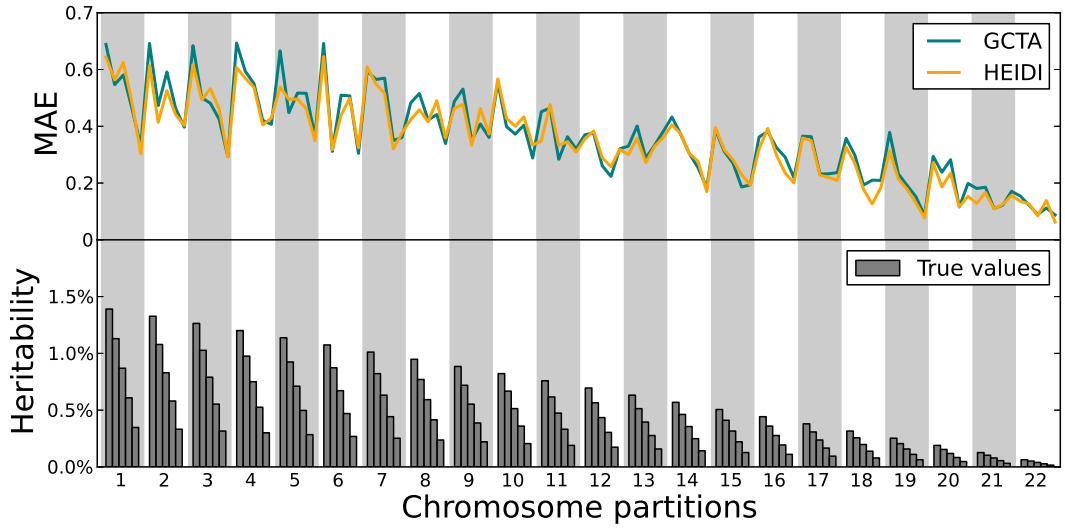


Figure 4.2: Mean absolute error values obtained by HEIDI and GCTA are shown in each region in the simulations where the total heritability is 50% and the regions have heritability contributions which vary across the genome. In this scenario, the accuracy of HEIDI is 3.64% higher than the accuracy of GCTA.

genetic relationship matrix from the chromosome. The idea is that in the presence of population structure or cryptic relatedness, an estimate from a single chromosome will capture part of the heritability from the rest of the genome. The length of the regions is then regressed on the difference between these estimates as shown in Figure 4.6. Yang et al.[YMP11] interpret the intercept corresponding to a measure of the cryptic relatedness and the slope corresponding to a measure of the population structure. The reasoning behind the hypothesized linear relationship between the length of the region and the difference in heritability estimates is due to the number of ancestral informative markers present in a region which can account for the difference is proportional to the length of the region assuming that the markers are evenly spaced. HEIDI estimates the slope and intercept of the graph at 7.51×10^{-5} and 0.0116 while GCTA estimates the slope and intercept at 6.84×10^{-5} and 0.0124. Utilizing the approach

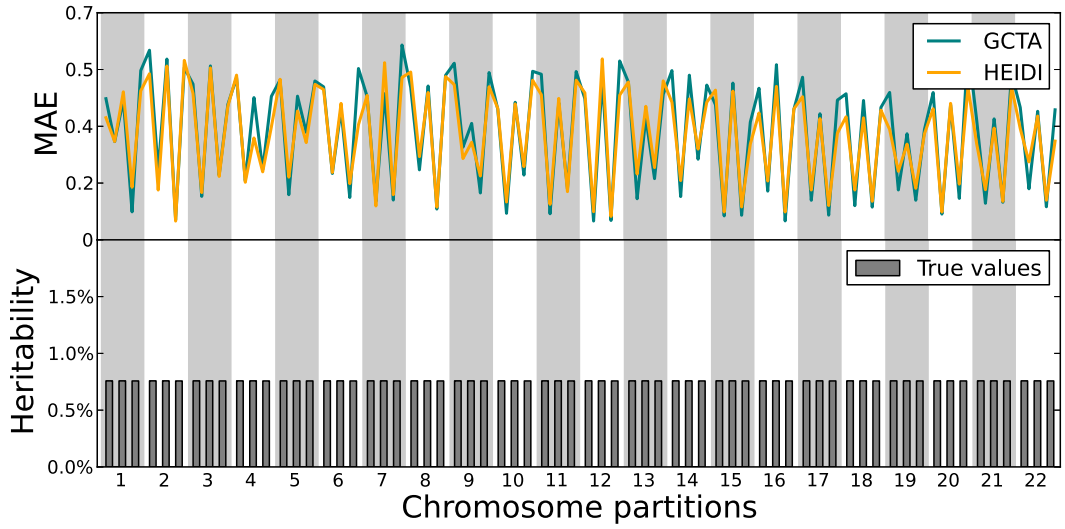


Figure 4.3: Mean absolute error values obtained by HEIDI and GCTA are shown in each region in the simulations where the total heritability is 50% and in each chromosome, the second and fourth regions do not contribute to the heritability while the first, third and fifth regions have equal contributions. In this scenario, the accuracy of HEIDI is 3.36% higher than the accuracy of GCTA.

described in Yang et al.[YMP11], these correspond to an estimate of the contribution of population structure to the heritability of 0.99% and 0.91% for HEIDI and GCTA respectively. We can compare the performance of HEIDI to GCTA in these estimates by considering how well the regression line fits the data in Figure 4.6 by measuring the residual sum of squares (RSS). Intuitively, the better the fit, the better the method is capturing the population structure signal from ancestry informative markers. HEIDI outperforms GCTA with RSS = 5.30×10^{-4} compared to RSS = 5.54×10^{-4}.

We also estimated the contribution of population structure using regions smaller than chromosomes, and the corresponding the regression is shown in Figure 4.7. Although the formula for estimating the population structure contributions in Yang et al.,[YMP11] does not immediately apply to estimates from shorter regions, the heri-

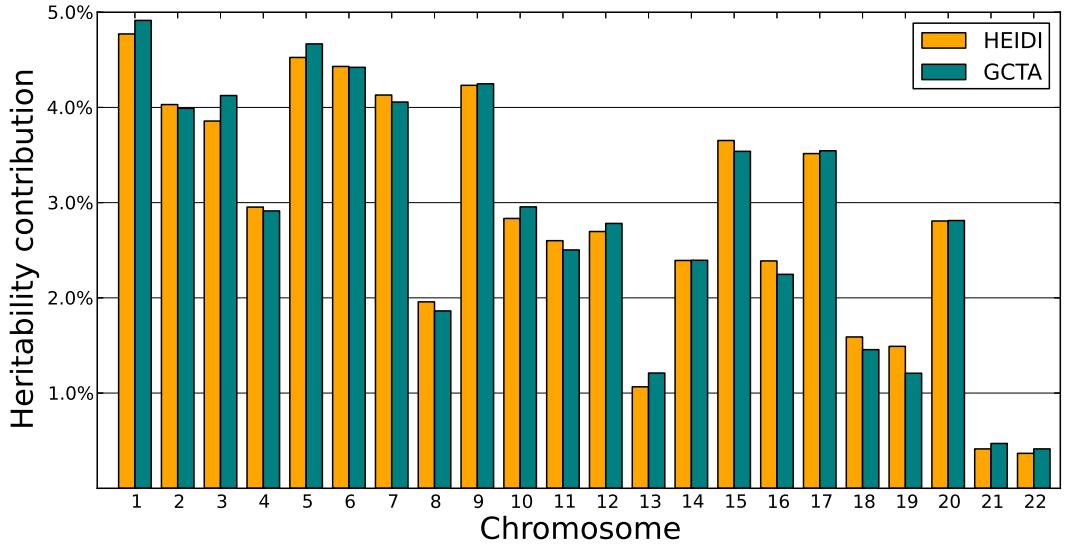


Figure 4.4: The heritability contributions of the 22 autosomal chromosomes to height are shown.

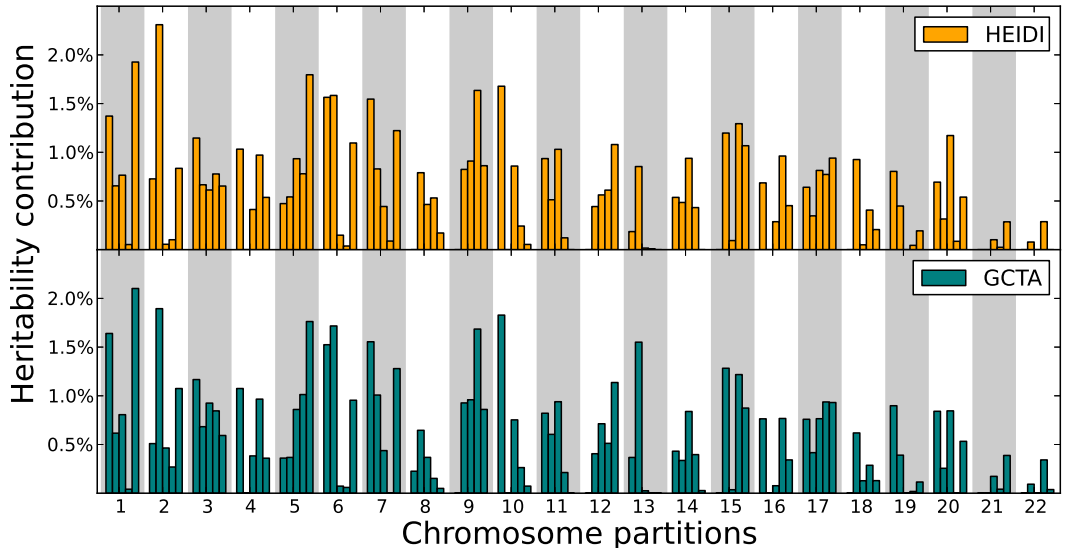


Figure 4.5: The heritability of height is partitioned into the contributions of chromosomal regions. For many regions GCTA estimates no heritability contribution.

tability estimates of HEIDI (RSS = 1.03×10^{-3}) fit the linear model better than GCTA (RSS = 1.12×10^{-3}).

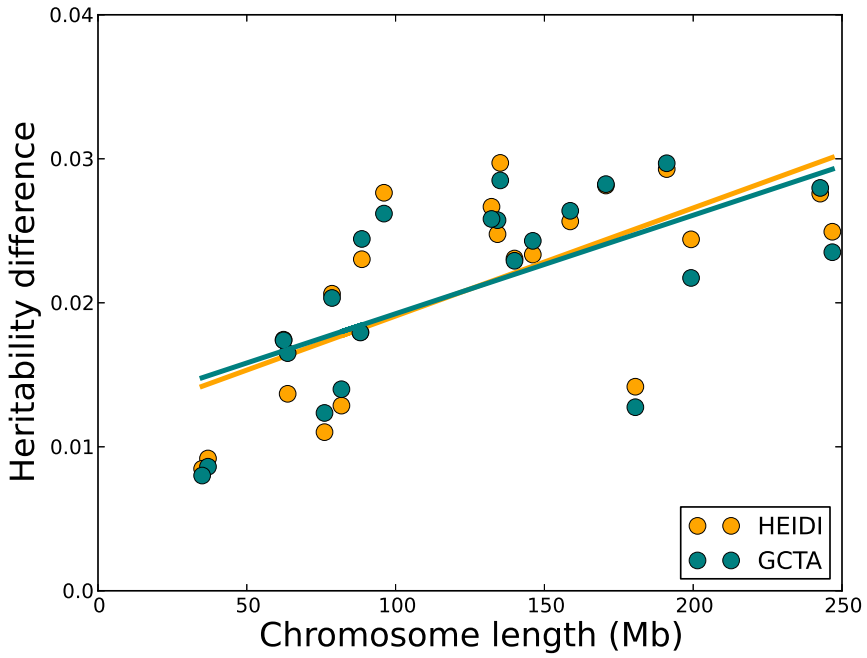


Figure 4.6: The difference between the heritability contribution of each chromosome estimated from partitioning the heritability and estimated independent of the rest of the genome; regressed against the length of the chromosome.

4.8 Discussion

A fundamental question in genetics is to understand the influence of genetic variation in complex traits. In tackling this question, traditional studies used related individuals to estimate the influence of genetics, relative to the environment. Recently, there has been growing interest in estimating heritabilities from genome-wide association study (GWAS) datasets containing large numbers of unrelated individuals. The total heritability can be partitioned into the contributions of the autosomes by including a variance component representing the contribution of each autosome and jointly estimating their contributions in the trait as implemented in the widely used method GCTA. However, as the number of regions increase and the regions themselves be-

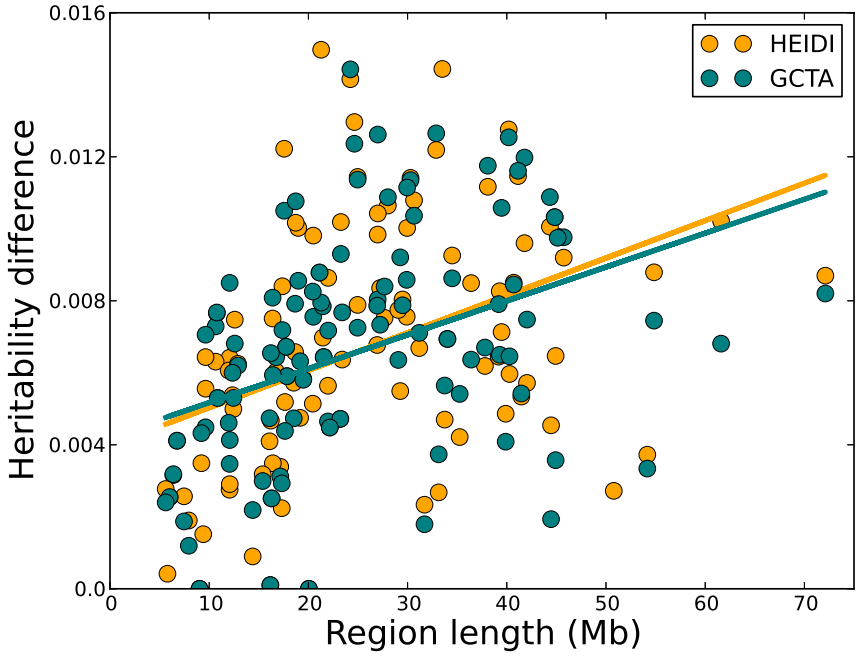


Figure 4.7: The difference between the heritability contribution of each of the 110 regions estimated from partitioning the heritability and estimated independent of the rest of the genome; regressed against the length of the genomic region.

come smaller, this approach becomes computationally intractable. We proposed an alternative method, HEIDI, that partitions the heritability into the contributions of an arbitrary number of regions. For each region, HEIDI decomposes the heritability as the sum of the contributions due to the region and the remainder of the genome. Finally, the heritability contributions are normalized such that their sum equals to the genome-wide heritability. Using simulations, we compared the performance of HEIDI to GCTA. On average, HEIDI performs over 5% more accurately when estimating heritability contributions than GCTA and can estimate these contributions in parallel.

One explanation for the improved accuracy of our approach is that the search space of the partitioned heritability has over 100 dimensions when we consider 5 regions per

chromosome and this space contains many local maxima. While GCTA searches the entire space, our approach takes advantage of spectral decomposition to find the maximum likelihood solutions in a restricted version of the space. Even though the maximum in the space searched by GCTA is higher than HEIDI's, HEIDI finds the maximum likelihood of the restricted space efficiently, which may end up being a better solution than GCTA's. We compared GCTA to HEIDI when partitioning the heritability into only two regions and as expected, the two methods find very similar estimates which is consistent with our explanation since the search space when considering only two regions is relatively small.

Consistent with current approaches to estimate heritability from unrelated individuals, our approach assumes an additive-genetic model, where the covariance between any two regions and by extension any region and it's corresponding genomic background is zero. The presence of gene-gene interactions, or epistatis, will violate this assumption and cause inaccuracies in our estimates of regional heritability contributions, similar to their effect on total heritability estimates[ZHS12, BEL13]. In addition, linkage disequilibrium between markers in neighboring regions will cause this assumption to be violated and result in the estimate of the heritability contribution of a region to be spread among the region along with the neighboring regions. However, this phenomenon is equivalent to how a causal variant causes neighboring variants in linkage disequilibrium to also have elevated statistics in a GWAS and inherent because of the linkage disequilibrium structure of the human genome. HEIDI's normalization of the region contribution estimates mitigates the effect of LD on the estimates.

Gene by gene interactions are not the only reason for observing deviations from the additive model. The presence of population structure in the sample may lead to inflation of the estimates of heritability of each region when estimated individually and cause the sum of the estimates from the regions differ from the estimate of the

total heritability[YBM10, YMP11]. HEIDI reports both the normalized estimates of heritability as well as the unnormalized estimates and can also be used to estimate the contribution of each region independently by omitting the background model. The differences between these estimates provides evidence of deviation from an additive model which can be explained by population structure or interaction effects.

CHAPTER 5

Conclusion

The tools that help us in observing the unknown are at the core of empirical sciences. In this respect, genomic technologies provide an invaluable view of the complex biological landscape within organisms. As these technologies continue to advance and become less expensive, life sciences and medicine will see an exponential influx of genomic data over the next decade.

How to utilize this data effectively to provide reliable clinical solutions for complex diseases tailored to each person's genetic data remains a major challenge. With this goal in mind, we are currently studying disease mechanisms by cataloguing all disease associated genetic polymorphisms. The genome-wide association study (GWAS) approach[DR95, RM96, Int05, HS09], has been an effective tool for this mission, especially when there is no prior knowledge about disease genetics. By genotyping an informative subset of all single-nucleotide polymorphisms (SNPs), referred to as tag SNPs, across the genome, from thousands of case and control individuals, the GWAS approach has led to the discovery of many novel genomic regions underlying complex traits[HSJ09].

In a GWAS, each tag SNP is statistically tested for its marginal association to disease status and regions harboring significant tag SNPs are believed to contain additional disease associated polymorphisms including the causal variants themselves. However, such regions are often large and contain many more SNPs, or candidate SNPs, that originally were not genotyped in the GWAS. The traditional approach per-

forms a follow-up study by genotyping all of these SNPs to uncover and catalogue the associations. In this dissertation, I showed that there are two main problems with the traditional approach. First, by genotyping all the SNPs in a significant region, the traditional approach is not a cost-effective strategy to perform a follow-up study. Second, the traditional approach is under powered by only following up the regions with significant associations because regions with strong associations, yet not significant, may also contain additional disease associations. I introduced a method to design powerful and cost-effective follow-up studies to identify all disease associated SNPs. This method (RFSS) takes into account the observed association statistics of tag SNPs, the correlation structure among all SNPs across the genome and the uncertainty on each candidate SNP begin a causal variant. The candidate SNPs are then ranked with respect to their likelihood of being significantly associated to disease. Clearly, we cannot know the exact correlation among all SNPs (tags and candidates) before genotyping all the candidates. This problem is addressed by leveraging public reference datasets such as HapMap[Int05] and 1000 Genomes Project[The10]. To my knowledge, RFSS is the first method proposed to improve power in GWAS.

After I introduced RFSS[KLE11], alternative approaches have been proposed to improve power in GWAS, such as performing extremely low coverage sequencing (for cost purposes) and using imputation to complete the missed SNPs[PRM12]. This approach has merits but lacks some of the key advantages of RFSS. The RFSS approach does not require the genotype data of SNPs but rather their association statistics, which are publicly available. In addition, RFSS explicitly models the data-generating model and integrates over all uncertainties on model parameters such as the assumed causal SNP structure. One interesting application would be to include prior probability distributions on different classes of genetic variation for being causal, such as missense mutations, SNPs within inter/intra genic regions, epigenetic factors and functional information from previous literature. The model also allows the computation of the

posterior probabilities for model parameters such as the probability of a SNP being causal in disease.

The RFSS model, in my opinion, is an important contribution as it lays out the fundamental mathematical structure to integrate knowledge in disease research from different biological sources. That is, more holistic studies can be achieved by using data on gene expression levels, metabolites and protein levels in addition to SNP data.

The second problem I addressed is the computational burden of analyzing very large scale gene expression association studies. Over the past few years, the GWAS approach has been applied to identify regions of the genome, which harbor genetic variation that affects gene expression levels, referred to as expression quantitative trait loci (eQTL)[BYC02, BK05, KFT07]. In a typical eQTL study, the GWAS approach is applied to tens of thousands of gene expression levels using millions of SNPs, resulting in billions of association statistics to be computed. Besides the computational burden arising from the number of tests performed, in advanced statistical models computing the actual association statistic may also be computationally very expensive. For instance, linear mixed model (LMM) has recently been of great interest to control for confounding due to population structure[KSS10, LLL11, ZS12], however the complexity of computing the association statistic is cubic in the number of individuals.

In analyzing a GWAS dataset, the current paradigm is to compute an association statistic of each SNP and repeating this process for all SNPs. I introduced a different association testing paradigm for high-throughput applications that identifies virtually all significant associations without testing all SNPs[KE13a]. This method (GRAT) utilizes the correlation structure between the SNPs and divides the association testing process into two stages. In the first-stage, GRAT tests a small subset of all SNPs, termed as proxy SNPs, and based on their association statistics chooses which of the remainder SNPs to test (second stage of testing) with a guaranteed expected recall rate

on the number of significant associations. The decision to whether or not to test a remainder SNP is achieved by assigning a threshold statistic to the remainder SNP's most strongly correlated proxy SNP. The objective of GRAT is to minimize the total number of SNPs tested by determining the proxy SNPs and the threshold values. The total number of tests directly corresponds to the time it takes to analyze an eQTL study.

I expressed the problem of finding the threshold values for given proxy SNPs as a constraint optimization problem where the expected number of tests is minimized while satisfying a target expected recall rate. I proved that this problem has a convex formulation, which led me to develop an efficient solver to attain the optimal threshold values. However, determining both the proxy SNPs and the threshold values is non-convex and I believe it can easily be shown to be NP-Hard with reduction from the vertex-cover problem[Kar72]. In order to solve the generalized problem, I introduced a greedy-heuristic approach by iteratively expanding the set of proxy SNPs by checking the value of the objective function had a remainder SNP been used as a proxy SNP. I improved the speed of this algorithm by caching the gradient of the objective function attained from the set of proxy SNPs used in the previous iteration and using this gradient while testing which remainder SNPs to choose as a proxy SNP. Once the proxy SNP is chosen, I then update the cached gradient to be used in the next iteration. This approach reduced the complexity of the greedy algorithm from quadratic to linear in the number of proxy SNPs. For additional speed improvements, I recommend determining the proxy SNPs on a reference dataset (HapMap, 1000G) and using them in GRAT by re-computing the threshold values for a particular GWAS. This only requires solving the convex optimization problem once, which is extremely fast.

The GRAT approach, in my view, has great potential in other genomic applications or domains that involve high-throughput hypothesis testing with correlated random variables where computational speed is a concern. One immediate genomic applica-

tion is on testing SNP-SNP interactions, which is becoming an active area of research. In two-level interactions, the naive approach could be to construct N^2 "interaction SNPs" from the N genotyped SNPs in GWAS and use the GRAT approach. This approach can be extended to higher-level interactions as well. A different application domain could be in high-frequency trading where securities are algorithmically traded at rapid speeds to satisfy positive expected return. Each security trade has an associated transaction cost and the securities have temporally correlated returns.

The third method I introduced is on partitioning the heritability–phenotypic variance attributable to additive genetic factors–into the contributions of genomic regions[KE13b]. While working on an initial version this problem where the regions were autosomes themselves, Yang et al. published their paper[YMP11] addressing the same problem. Their model includes a separate variance component for each region and simultaneously estimates their variance parameters. From the beginning, I aimed at a simpler model to leverage the additive genetics for computational efficiency; using two genetic variance components, one for the genomic region of interest and one for its genetic background. I showed that the traditional approach becomes computationally intractable as the number of regions increases, whereas the proposed method (HEIDI) is immune to this problem and also distribute computations for additional gain in speed.

One issue that required attention in HEIDI was the linkage disequilibrium (LD) among neighboring regions, which inflates the cumulative heritability obtained from all regions. I proposed a simple heritability normalization scheme using the heritabilities obtained from all SNPs, termed genome-wide heritability, and from the SNPs in each chromosome. The heritabilities of chromosomes are normalized such that their sum equals to the genome-wide heritability. Similarly, the heritabilities of regions in each chromosome are normalized such that their sum equals to the normalized heritability of the chromosome. The basis for this scheme is that recombination events

occur independently between different chromosomes, hence in a population we do not expect LD among them. The normalization scheme worked well in simulations and HEIDI performed more accurately than the traditional approach.

I showed that an additional utility of HEIDI is in estimating the proportion of heritability that is due to population structure. Assuming ancestry informative markers (AIMs) are uniformly distributed across the genome, one can regress the differences between a region's heritability estimates with and without the genomic background to region's size, and estimate the contribution of population stratification in heritability[YMP11]. Using HEIDI, I observed less residual error, however when I increased the number of regions the estimate for the contribution of population stratification changed. In my view, improving this population stratification model would make an interesting research direction.

The HEIDI approach can be used to ask additional questions to understand the genetic basis of complex diseases, where the marginal effect sizes of SNPs are small, yet in aggregate they explain a large portion of the phenotypic variation. A good research direction is to perform association testing for regions or known pathways by collapsing the SNPs within the pathway genes. In addition, the joint distribution of the marginal association statistics of genomic regions can be modeled, which lets one to extend RFSS and GRAT methods to be used for genomic regions in addition to SNPs. A second important application of the HEIDI approach would be on phenotype and disease-risk prediction, where due to their small effect sizes using SNPs individually has had limited success.

Furthermore, a common assumption in association studies and heritability calculations is that the contribution of genetics to a trait is the same in different environmental conditions. When this assumption is not met, there is a statistical interaction between genetics and environment, where the environment influences the contribution of genet-

ics to the trait. Such additional source of variation, when unaccounted for, may lead to spurious associations or loss in power. The methods presented in this dissertation provide useful basis for analyzing the interaction phenomenon.

Finally, the methods I introduced model the association of multiple SNPs to a single phenotype. In my opinion, it will be an important contribution to extend the approaches I presented to include multiple phenotypes.

REFERENCES

[Bak12] M. Baker. "Biorepositories: Building better biobanks." *Nature*, **486**(7401):141–146, Jun 2012.

[BB11] S. R. Browning and B. L. Browning. "Population structure can inflate SNP-based heritability estimates." *Am J Hum Genet*, **89**(1):191–193, Jul 2011.

[BEL13] J. S. Bloom, I. M. Ehrenreich, W. Loo, T. V. Lite, and L. Kruglyak. "Finding the sources of missing heritability in a yeast cross." *Nature*, **In Press**, Aug 2013.

[BK05] R. B. Brem and L. Kruglyak. "The landscape of genetic complexity across 5,700 gene expression traits in yeast." *Proc Natl Acad Sci U S A*, **102**(5):1572–1577, Feb 2005.

[Boc03] B. R. Bochner. "Innovations: New technologies to assess genotype-phenotype relationships." *Nat Rev Genet*, **4**(4):309–314, Apr 2003.

[BWD05] L. Bystrykh, E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher, T. Wiltshire, A. I. Su, E. Vellenga, J. Wang, K. F. Manly, L. Lu, E. J. Chesler, R. Alberts, R. C. Jansen, R. W. Williams, M. P. Cooke, and G. de Haan. "Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'." *Nat Genet*, **37**(3):225–232, Mar 2005.

[BYC02] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak. "Genetic dissection of transcriptional regulation in budding yeast." *Science*, **296**(5568):752–755, Apr 2002.

[BYP05] P. I. W. de Bakker, R. Yelensky, I. Pe'er, S. B. Gabriel, M. J. Daly, and D. Altshuler. "Efficiency and power in genetic association studies." *Nat Genet*, **37**(11):1217–1223, Oct 2005.

[CB01] L. R. Cardon and J. I. Bell. "Association study designs for complex diseases." *Nat Rev Genet*, **2**(2):91–99, Feb 2001.

[CDG06] E. Cousin, J. F. Deleuze, and E. Genin. "Selection of SNP subsets for association studies in candidate genes: comparison of the power of different strategies to detect single disease susceptibility locus effects." *BMC Genetics*, **7**:20–29, Apr 2006.

[CER04] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson. "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium." *Am J Hum Genet*, **74**(1):106–120, Jan 2004.

[CGM03] E. Cousin, E. Genin, S. Mace, S. Ricard, C. Chansac, M. del Zompo, and J. F. Deleuze. "Association studies in candidate genes: strategies to select SNPs to be tested." *Hum Hered*, **56**(4):151–159, Mar 2003.

[CLA09] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop. "Mapping complex disease traits with global gene expression." *Nat Rev Genet*, **10**(3):184–194, Mar 2009.

[CLS05] E. J. Chesler, L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. C. Hsu, J. D. Mountz, N. E. Baldwin, M. A. Langston, D. W. Threadgill, K. F. Manly, and R. W. Williams. "Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function." *Nat Genet*, **37**(3):233–242, Feb 2005.

[Col92] F. S. Collins. "Positional cloning: Let's not call it reverse anymore." *Nat Genet*, **1**(1):3–6, Apr 1992.

[Col95] F. S. Collins. "Positional cloning moves from perditional to traditional." *Nat Genet*, **9**(4):347–350, Apr 1995.

[Con04] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome." *Nature*, **431**(7011):931–945, Oct 2004.

[CSE05] V. G. Cheung, R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley, and J. T. Burdick. "Mapping determinants of human gene expression by regional and genome-wide association." *Nature*, **437**(7063):1365–1369, Oct 2005.

[DR95] B. Devlin and N. Risch. "A comparison of linkage disequilibrium measures for fine-scale mapping." *Genomics*, **29**(2):311–222, Sep 1995.

[DSD12] J. van Dongen, P. E. Slagboom, H. H. M. Draisma, N. G. Martin, and D. I. Boomsma. "The continuing value of twin studies in the omics era." *Nat Rev Genet*, **13**(9):640–653, Aug 2012.

[ETZ08] V. Emilsson, G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G. B. Walters, S. Gunnarsdottir, M. Mouy, V. Steinthorsdottir, G. H. Eiriksdottir, G. Bjornsdottir, I. Reynisdottir, D. Gudbjartsson, A. Helgadottir, A. Jonasdottir, A. Jonasdottir, U. Styrkarsdottir, S. Gretarsdottir, K. P. Magnusson, H. Stefansson, R. Fossdal, K. Kristjansson, H. G. Gislason, T. Stefansson, B. G. Leifsson, U. Thorsteinsdottir, J. R. Lamb, J. R. Gulcher, M. L. Reitman, A. Kong, E. E. Schadt, and K. Stefansson. "Genetics of gene expression and its effect on disease." *Nature*, **452**(7186):423–428, Mar 2008.

[GNA02] A. M. Glazier, J. H. Nadeau, and T. J. Aitman. "Finding genes that underlie complex traits." *Science*, **298**(5602):2345–2349, Dec 2002.

[Har74] D. A. Harville. "Bayesian inference for variance components using only error contrasts." *Biometrika*, **61**(2):383–385, 1974.

[Har77] D. A. Harville. "Maximum likelihood approaches to variance component estimation and to related problems." *JASA*, **72**(358):320–338, 1977.

[HKE09] B. Han, H. M. Kang, and E. Eleazar. "Rapid and accurate multiple testing correction and power estimation for millions of correlated markers." *PLoS Genet*, **5**(4):e1000456, Apr 2009.

[HKS05] E. Halperin, G. Kimmel, and R. Shamir. "Tag SNP selection in genotype data for maximizing SNP prediction accuracy." *Bioinformatics*, **21**:195–203, 2005.

[HLN90] J. M. Hall, M. K. Lee, B. Newman, J. E. Morrow, L. A. Anderson, B. Huey, and M. C. King. "Linkage of early-onset familial breast cancer to chromosome 17q21." *Science*, **250**(4988):1684–1689, Dec 1990.

[HPC10] B. J. Hayes, J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. "Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits." *PLoS Genet*, **6**(9):e1001139, Sep 2010.

[HS09] J. Hardy and A. Singleton. "Genomewide association studies and human disease." *N Engl J Med*, **360**(17):1759–1768, Apr 2009.

[HSJ09] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." *Proc Natl Acad Sci U S A*, **106**(23):9362–9367, Jun 2009.

[Int05] International HapMap Consortium. "A haplotype map of the human genome." *Nature*, **437**(7063):1299–1320, Oct 2005.

[Kar72] R. M. Karp. "Reducibility Among Combinatorial Problems." In *Complexity of Computer Computations*, pp. 85–103, 1972.

[KE13a] E. Kostem and E. Eskin. "Efficiently identifying significant associations in genome-wide association studies." *J Comput Biol*, In press 2013.

[KE13b] E. Kostem and E. Eskin. "Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions." *Am J Hum Genet*, In press 2013.

[KFT07] J. J. B. Keurentjes, J. Fu, I. R. Terpstra, J. M. Garcia, G. van den Ackerveken, L. B. Snoek, A. J. M. Peeters, D. Vreugdenhil, M. Koornneef, and R. C. Jansen. "Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci." *Proc Natl Acad Sci U S A*, **104**(5):1708–1713, Jan 2007.

[KLE11] E. Kostem, J. A. Lozano, and E. Eskin. "Increasing power of genome-wide association studies by collecting additional single-nucleotide polymorphisms." *Genetics*, **188**(2):449–460, Jun 2011.

[KSS10] H. M. Kang, J. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. "Variance component model to account for sample structure in genome-wide association studies." *Nat Genet*, **42**(4):348–354, Apr 2010.

[KZW08] H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. "Efficient control of population structure in model organism association mapping." *Genetics*, **178**(3):1709–1723, Mar 2008.

[LA04] Z. Lin and R. B. Altman. "Finding haplotype tagging SNPs by use of principal components analysis." *Am J Hum Genet*, **75**(5):850–861, Nov 2004.

[LDR12] S. H. Lee, T. R. Decandia, S. Ripke, J. Yang, P. F. Sullivan, M. E. Goddard, M. C. Keller, P. M. Visscher, and N. R. Wray. "Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs." *Nat Genet*, **44**(3):247–250, Feb 2012.

[LLL11] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman. "FaST linear mixed models for genome-wide association studies." *Nat Methods*, **8**(10):833–835, Sep 2011.

[LS94] E. S. Lander and N. J. Schork. "Genetic dissection of complex traits." *Science*, **265**(5181):2037–2048, Sep 1994.

[LW98] M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, 1998.

[LWD10] Y. Li, C. J. Willer, J. Ding, P. Scheet, and G. Abecasis. "MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes." *Genet Epidemiol*, **34**(8):816–834, Dec 2010.

[Mar11] E. R. Mardis. "A decade's perspective on DNA sequencing technology." *Nature*, **470**(7333):198–203, Feb 2011.

[Met10] M. L. Metzker. "Sequencing technologies - the next generation." *Nat Rev Genet*, **11**(1):31–46, Jan 2010.

[MP11] J. Majewski and T. Pastinen. "The study of eQTL variations by RNA-seq: from SNPs to phenotypes." *Trends Genet*, **27**(2):72–79, Nov 2011.

[PBG91] M. A. Pericak-Vance, J. L. Bebout, P. C. Gaskell, L. H. Yamaoka, W. Y. Hung, M. J. Alberts, A. P. Walker, R. J. Bartlett, C. A. Haynes, and K. A. Welsh. "Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage." *Am J Hum Genet*, **48**(6):1034–1050, Jun 1991.

[PLW05] F. Pardi, C. M. Lewis, and J. C. Whittaker. "SNP selection for association studies: Maximizing power across SNP choice and study size." *Ann Hum Genet*, **69**(6):733–746, Nov 2005.

[PNT07] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. "PLINK: A tool set for whole-genome association and population-based linkage analyses." *Am J Hum Genet*, **81**(3):559–575, Sep 2007.

[PP01] J. K. Pritchard and M. Przeworski. "Linkage disequilibrium in humans: Models and data." *Am J Hum Genet*, **69**(1):1–14, Jul 2001.

[PR99] J. K. Pritchard and N. A. Rosenberg. "Use of unlinked genetic markers to detect population stratification in association studies." *Am J Hum Genet*, **65**(1):220–228, Jul 1999.

[PRM12] Bogdan Pasaniuc, Nadin Rohland, Paul J. McLaren, Kiran Garimella, Noah Zaitlen, Heng Li, Namrata Gupta, Benjamin M. Neale, Mark J. Daly, Pamela Sklar, Patrick F. Sullivan, Sarah Bergen, Jennifer L. Moran, Christina M. Hultman, Paul Lichtenstein, Patrik Magnusson, Shaun M. Purcell, David W. Haas, Liming Liang, Shamil Sunyaev, Nick Patterson, Paul I. W. de Bakker, David Reich, and Alkes L. Price. "Extremely low-coverage sequencing and imputation increases power for genome-wide association studies." *Nat Genet*, **44**(6):631–635, Jun 2012.

[PSR00] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. "Association mapping in structured populations." *Am J Hum Genet*, **67**(1):170–181, Jul 2000.

[PT71] H. D. Patterson and R. Thompson. "Recovery of inter-block information when block sizes are unequal." *Biometrika*, **58**:545–554, 1971.

[QGA06] Z. S. Qin, S. Gopalakrishnan, and G. R. Abecasis. "An efficient comprehensive search algorithm for tag SNP selection using linkage disequilibrium criteria." *Bioinformatics*, **22**(2):220–225, Jan 2006.

[RG01] D. E. Reich and D. B. Goldstein. "Detecting association in a case-control study while correcting for population stratification." *Genet Epidemiol*, **20**(1):4–16, Jan 2001.

[RK06] M. V. Rockman and L. Kruglyak. "Genetics of global gene expression." *Nat Rev Genet*, **7**(11):862–872, Nov 2006.

[RM96] N. Risch and K. Merikangas. "The future of genetic studies of complex human diseases." *Science*, **273**(5281):1516–1517, Sep 1996.

[SBB07] R. S. Spielman, L. A. Bastone, J. T. Burdick, M. Morley, W. J. Ewens, and V. G. Cheung. "Common genetic variants account for differences in gene expression among ethnic groups." *Nat Genet*, **39**(2):226–231, Feb 2007.

[SFH12] H. Signer-Hasler, C. Flury, B. Haase, D. Burger, H. Simianer, T. Leeb, and S. Rieder. "A genome-wide association study reveals loci influencing height and other conformation traits in horses." *PLoS One*, **7**(5):e37282, May 2012.

[SFY01] G. A. Satten, W. D. Flanders, and Q. Yang. "Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model." *Am J Hum Genet*, **68**(2):466–477, Feb 2001.

[SMD12] B. E. Stranger, S. B. Montgomery, A. S. Dimas, L. Parts, O. Stegle, C. E. Ingle, M. Sekowska, G. D. Smith, D. Evans, M. Gutierrez-Arcelus, A. Price, T. Raj, J. Nisbett, A. C. Nica, C. Beazley, R. Durbin, P. Deloukas, and E. T. Dermitzakis. "Patterns of cis regulatory variation in diverse human populations." *PLoS Genet*, **8**(4):e1002639, Apr 2012.

[SNF07] B. E. Stranger, A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird, C. Beazley, C. E. Ingle, M. Dunning, P. Flicek, D. Koller, S. Montgomery, S. Tavaré, P. Deloukas, and E. T. Dermitzakis. "Population genomics of human gene expression." *Nat Genet*, **39**(10):1217–1224, Sep 2007.

[SRS06] S. F. Saccone, J. P. Rice, and N. L. Saccone. "Power-based, phase-informed selection of single nucleotide polymorphisms for disease association screens." *Genet Epidemiol*, **30**(6):459–470, Sep 2006.

[SSH08] C. Sabatti, S. K. Service, A. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen,

S. Gabriel, P. Elliot, M. I. McCarthy, M. J. Daly, M. Jarvelin, N. B. Freimer, and L. Peltonen. "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population." *Nat Genet*, **41**(1):35–46, Dec 2008.

[SSP03] K. Silventoinen, S. Sammalisto, M. Perola, D. I. Boomsma, B. K. Cornes, C. Davis, L. Dunkel, M. De Lange, J. R. Harris, J. V. B. Hjelmborg, M. Luciano, N. G. Martin, J. Mortensen, L. Nisticò, N. L. Pedersen, A. Skytthe, T. D. Spector, M. A. Stazi, G. Willemsen, and J. Kaprio. "Heritability of adult body height: A comparative study of twin cohorts in eight countries." *Twin Res*, **6**(5):399–408, Oct 2003.

[Ste10] L. D. Stein. "The case for cloud computing in genome informatics." *Genome Biol*, **11**(5):207–214, May 2010.

[Str04] D. O. Stram. "Tag SNP selection for association studies." *Genet Epidemiol*, **27**(4):365–374, Dec 2004.

[Str05] D. O. Stram. "Software for tag single nucleotide polymorphism selection." *Hum Genomics*, **2**(2):144–151, Jun 2005.

[The04] The ENCODE Project Consortium. "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science*, **306**(5696):636–640, Oct 2004.

[The07] The ENCODE Project Consortium. "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." *Nature*, **447**(7146):799–816, Jun 2007.

[The10] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." *Nature*, **467**(7319):1061–1073, Oct 2010.

[The11] The ENCODE Project Consortium. "A user's guide to the encyclopedia of DNA elements (ENCODE)." *PLoS Biol*, **9**(4):e1001046, Apr 2011.

[The12] The ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." *Nature*, **489**(7414):57–74, Sep 2012.

[VHW08] P. M. Visscher, W. G. Hill, and N. R. Wray. "Heritability in the genomics era–concepts and misconceptions." *Nat Rev Genet*, **9**(4):255–266, Apr 2008.

[VP05] B. F. Voight and J. K. Pritchard. "Confounding from cryptic relatedness in case-control association studies." *PLoS Genet*, **1**(3):e32, Sep 2005.

[Wal07] Francis O. Walker. "Huntington's disease." *Lancet*, **369**(9557):218–228, Jan 2007.

[WGS09] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: A revolutionary tool for transcriptomics." *Nat Rev Genet*, **10**(1):57–63, Jan 2009.

[WTC07] WTCCC. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." *Nature*, **447**(7145):661–678, Jun 2007.

[YBM10] J. Yang, B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. "Common SNPs explain a large proportion of the heritability for human height." *Nat Genet*, **42**(7):565–569, Jul 2010.

[YLG11] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher. "GCTA: A tool for genome-wide complex trait analysis." *Am J Hum Genet*, **88**(1):76–82, Jan 2011.

[YMP11] J. Yang, T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, M. de Andrade, B. Feenstra, E. Feingold, M. G. Hayes, W. G. Hill, M. T. Landi, A. Alonso, G. Lettre, P. Lin, H. Ling, W. Lowe, R. A. Mathias, M. Melbye, E. Pugh, M. C. Cornelis, B. S. Weir, M. E. Goddard, and P. M. Visscher. "Genome partitioning of genetic variation for complex traits using common SNPs." *Nat Genet*, **43**(6):519–525, May 2011.

[ZFG06] G. Zheng, B. Freidlin, and J. L. Gastwirth. "Robust genomic control for association studies." *Am J Hum Genet*, **78**(2):350–356, Feb 2006.

[ZHS12] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander. "The mystery of missing heritability: Genetic interactions create phantom heritability." *Proc Natl Acad Sci USA*, **109**(4):1193–1198, Jan 2012.

[ZK12] N. Zaitlen and P. Kraft. "Heritability in the genome-wide association era." *Hum Genet*, **131**(10):1655–1664, Oct 2012.

[ZS12] X. Zhou and M. Stephens. "Genome-wide efficient mixed-model analysis for association studies." *Nat Genet*, **44**(7):821–824, Jun 2012.