UNIVERSITY OF CALIFORNIA, SAN DIEGO

Effective design and analysis of genetic association studies

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Computer Science

by

Buhm Han

Committee in charge:

University of California, San Diego

Professor Vineet Bafna, Chair
Professor Sanjoy Dasgupta
Professor Daniel O'Connor
Professor Pavel Pevzner

University of California, Los Angeles

Professor Eleazar Eskin

2009

The dissertation of Buhm Han is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2009

DEDICATION

To Jee Hee.

EPIGRAPH

If the axe is dull and he does not sharpen its edge,
then he must exert more strength.
Wisdom has the advantage of giving success.
— Ecclesiastes 10:10

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I thank my adviser Eleazar Eskin for his endurance in guiding my research. My writing was so terrible in 2006 but became better little by little thanks to your extensive teaching and guidance. You never looked frustrated while we make corrections on my first paper, the design paper, about 50 times if I correctly remember. You spared urges and pushes, and instead encouraged me with compliments and comforts in spite of my slow learning curve, so that I can slowly develop my knowledge and skills.

I also thank my fantastic colleagues in ZarLab. Especially I thank Hyun Min Kang, my senior of Seoul National University, colleague recipient of Samsung Scholarship, one-year senior of ZarLab, and a member of the same church, for guiding and helping me. Most of all, I thank Hyun for telling my adviser that I am a smart guy so that I can enter the lab easily. I am sure that he was never suspicious of that statement until he accepts me into the lab. I thank Noah Zaitlen for teaching me the association study basics in 1-hour course in the EBU-1 building in UCSD. That was really helpful. I thank Sean O'Rourke for his technical perfection about operating systems and programming languages. I especially thank for his help on the compilation of my method SLIDE using Intel Math Kernel Library. I thank Chun Ye for sharing insights and interesting projects with me. I always feel positive and strengthened when I see him because he is so energetic. I also thank the rest of my current and previous lab colleagues Emrah Kostem, Chris Jones, Olivera Grujic, Eun Yong Kang, Nick Furlotte, Jae-hoon Sul, Dan He, Nils Homer, Michael Sanders, and Ilya Shpitster for sharing scientific insights and spending great time together.

There are many terrific people I would like to offer my gratitude. I would like to thank Vineet Bafna, Banu Dost, and Shaojie Zhang for helping me to write my first conference paper and subsequently a journal paper. That was a really encouraging experience that gave me confidence in this biology-related field which I had to confront without background knowledge. That project started from the class taught by Pavel Pevzner as well, who I would like to thank also for being my thesis committee. I thank Sanjoy Dasgupta for helping me to set up the inter-campus exchanges for every quarter. Due to my slow learning curve, I now start

to understand what you taught in your three machine learning classes I attended. Those classes were very insightful and I expect will help my future researches. I thank Daniel O'Connor for being my committee and guiding our lab to a terrific Italian restaurant in Philadelphia at the last year's American Society of Human Genetics meeting. I also thank Myeong Seong Seo and Brian Hackel for collaborating and helping the projects extensively.

I thank my parents, Yoon-soo Han and Min-hee Kim for their endurance, love, prayers, and encouragements. I also thank my brother, Kye-baek Han, who often called me and encouraged. I thank my parents-in-law, Joong-il Park and Eun-Young Kang for allowing me to marry their beautiful daughter and always loving and providing the best environment for the study. I thank my sister-in-law, Jisu Park, for being a so nice sister-in-law. Without you, it would have been much harder to manage my life since our little baby was born. I also thank my godparents, Jerry Williams and Hee-Ja Seo (Maria Williams), for considering me as a son and not ceasing the care and love. I thank my pretty daughter, Lois Han, for being a so nice girl and being born very healthy. I also thank other family members including my grandmother, Jung-soon Chun, sister-in-law, So-young Lee, grandparents-in-law, and aunts for their love and helps.

Finally, I would like to thank my wife, Jee Hee Park, for waiting so much time in endurance for my graduation. Without your help, I would never be in this state, happy, healthy, full of gratitude, and filled with love. Finally and most importantly, I would like to thank the almighty God and Jesus Christ, my Lord and Savior, for his everlasting love, grace, mercy and kindness.

I was supported by the Samsung Fellowship, National Science Foundation grant no. 0513612, 0713455, and 029049, National Institute of Health grant no. 1K25HL-080029. and the UCLA subcontract of contract N01-ES-45530 from the National Toxicology Program/ National Institute of Environmental Health Sciences to Perlegen Sciences. I thank the National Biomedical Computational Resource (NBCR), the UCSD Fast Wired and Wireless Grid Project (FWGrid), Institute for Digital Research and Education (IDRE) for allowing us to utilize their cluster system.

Chapter 2 was published in PLoS Genetics, 5(4):e1000456, 2009, Buhm Han, Hyun Min Kang, and Eleazar Eskin. "Rapid and accurate multiple testing correction

and power estimation for millions of correlated markers". The dissertation author was the primary investigator and author of this paper.

Chapter 3 was published in Annals of Human Genetics, 72(Pt 6):834-847, Nov 2008, Buhm Han, Hyun Min Kang, Myeong Seong Seo, Noah Zaitlen, and Eleazar Eskin. "Efficient association study design via power-optimized tag SNP selection". The dissertation author was the primary investigator and author of this paper.

Chapter 4 is currently in submission for publication of the material. Buhm Han, Brian M. Hackel, and Eleazar Eskin, "Finding new associations and detecting spurious associations at post-association stage". The dissertation author is the primary investigator and author of this paper.

VITA

2004	Bachelor of Science in Electrical Engineering, Seoul National University, Seoul, Korea
2007	Master of Science in Computer Science, University of California, San Diego
1999-2001	Republic of Korea Army
2004-2005	Teaching Assistant, University of California, San Diego
2005-2006	Research Assistant, University of California, San Diego
2007-2009	Research Assistant, University of California, Los Angeles
2009	Doctor of Philosophy in Computer Science, University of California, San Diego

PUBLICATIONS

Hyun Min Kang, Noah A. Zaitlen, Buhm Han, and Eleazar Eskin, "An adaptive and memory efficient algorithm for genotype imputation", *In Proceedings of the Thirteenth Annual Conference on Research in Computational Biology (RECOMB-2009).* Tuscon, Arizona: May 18th-21st, 2009

Buhm Han, Hyun Min Kang, and Eleazar Eskin, "Rapid and accurate multiple testing correction and power estimation for millions of correlated markers", PLoS Genetics, 5:e1000456, 2009

Noah Zaitlen, Arthur Choi, Buhm Han, Knot Pipatsrisawat, Adnan Darwiche, and Eleazar Eskin,"Efficient Genome Wide Tagging by Reduction to SAT", *In Proceedings of the 8th Workshop on Algorithms in Bioinformatics (WABI-2008).* University Karlsruhe, Germany: September 15-17, 2008

Buhm Han, Hyun Min Kang, Myeong Seong Seo, Noah A. Zaitlen, and Eleazar Eskin, "Efficient association study via power-optimized tag SNP selection", Annals of Human Genetics, 72:834-47, 2008

Buhm Han, Banu Dost, Vineet Bafna, and Shaojie Zhang, "Structural Alignment of Pseudoknotted RNA", Journal of Computational Biology, 15(5): 489-504, 2007

Banu Dost, Buhm Han, Shaojie Zhang, and Vineet Bafna, "Structural Alignment of Pseudoknotted RNA", *In Proceedings of the 10th Annual Conference on Research in Computational Biology (RECOMB-2006).* Venice Lido, Italy: April 205, 2006

FIELDS OF STUDY

Major Field: Computer Science
Professor Eleazar Eskin

ABSTRACT OF THE DISSERTATION

Effective design and analysis of genetic association studies

by

Buhm Han
Doctor of Philosophy in Computer Science
University of California San Diego, 2009
Professor Vineet Bafna, Chair

Genetic association studies are an effective means of discovering associations between genetic variants and diseases. The procedure of association studies can be summarized into four stages of design, sample collection, analysis, and follow-up. There exist many statistical and computational challenges in the design and analysis stages of these studies. These challenges are closely related to exploring the correlation structure of genetic variations in the genome called linkage disequilibrium (LD). In this dissertation, I address some of these challenges and propose solutions which effectively leverage the information in LD patterns.

Multiple hypothesis testing correction is the major challenge in the analysis stage. It is difficult to assess the statistical significance of associations in association studies because a large number of correlated tests are simultaneously performed. Previous approaches are either inaccurate or prohibitively inefficient. I propose a novel multiple testing correction method which takes advantage of the local LD patterns by using a sliding-window approach. My method is highly accurate and efficient, effectively replacing the current approaches.

Estimating statistical power of a study design is a necessary procedure in the design stage to avoid under- or over-powered study. Current approaches are either inefficient or too conservative because they ignore the correlation between tests. I propose a method which takes into account the LD patterns to estimate statistical power of a study design efficiently and accurately.

Tag SNP selection problem is a widely-known challenge in the design stage. I propose a power-based tag SNP selection algorithm which greedily chooses SNPs to maximize the study power. My method outperforms other correlation only-based methods, because I take advantage of the relation between LD and power by accounting for allele frequencies.

In the analysis stage, detecting spurious associations is a challenging problem. I propose a novel method which detects spurious associations at the post-association stage using the LD information. Moreover, I extend this framework to propose a new study scheme which "rescues" associations at markers that are excluded by quality controls. My method is applied to the WTCCC dataset to identify a novel association which is recently replicated.

Chapter 1

Introduction

Human genetic variations such as single nucleotide polymorphisms (SNP) are considered to be the major sources of human phenotypic variations such as heights, eye colors, and susceptibilities to genetic diseases [19, 61, 13]. For many diseases having genetic risk factors, there have been a tremendous amount of efforts to identify the variations increasing the risk of being affected by the disease [74, 2]. A successful identification will lead us to the knowledge of how the disease phenotypes are developed and ultimately to the development of effective medicines.

The *genetic association study* is the the most recent and popular study framework for finding these disease-causing variants. Compared to the previous *linkage study* framework [61], genetic association studies have advantages such as the large sample size allowing high statistical power and high resolution to locate the causal variants. The significant price drop of high-throughput genotyping technology platforms now allows more than hundreds of association studies to be performed a year to report many candidate causal variants, which will be subsequently verified and ultimately utilized to beneficial applications [74, 4, 72].

The advent of genetic association studies is closely related to the correlation structure of genetic variations in human genome called linkage disequilibrium (LD) [58]. Since genetic variations of ancestors are passed to descendants through a limited number of recombination events, genetic markers which are proximal to each other are often correlated in the population level (Figure 1.1). As a result, each of the current human populations now shows a unique correlation structure of the genetic

variations which can be estimated from the reference data such as the HapMap [1, 35]. The idea behind genetic association studies is that we can virtually detect causal variants at any place of the whole genome because uncollected variants are correlated to the collected variants in the form of LD patterns.

I propose that many current computational and statistical challenges in genetic association studies can be solved by fully utilizing this LD information. Among the four major stages of association studies (Figure 1.2), which are design, sample collection, analysis, and follow-up, I focus on the major problems in the design and analysis stages. In particular, I address five different but related problems: the multiple hypothesis testing correction, power estimation, tag SNP selection, spurious association detection, and rescuing associations discarded by quality controls. I show that all these problems can be effectively solved by leveraging the full potential of LD information.

One of the major challenges at the analysis stage is the multiple hypothesis testing correction problem. The decision whether an observation is statistically significant or not is made based on the probability that an equal or more significant observation will be observed under the null hypothesis of no association [73] (called p-value). For example, a p-value of 0.01 means that there is only 1% of chance to observe this much significance under the null. However, assessing p-values is a hard problem for association studies with a large number of correlated tests. If multiple tests are simultaneously performed, the p-value obtained for each individual test (called pointwise p-value) is no longer equal to the probability that we will observe that much significance under the null (called corrected p-value) [67, 14]. Correcting the pointwise p-values to the corrected p-values is the popular multiple testing problem. The traditional solutions for this problem are either too conservative (e.g. the Bonferroni correction) or prohibitively time consuming (e.g. the permutation test) [14].

In Chapter 2, I present a novel method called SLIDE (Sliding-window approach for Locally Inter-correlated markers with asymptotic Distribution Errors cor-

rected), which efficiently and accurately solves the multiple hypothesis testing problem utilizing the LD information [30]. My method is based on two ideas. First, a sliding-window approach can effectively take into account the correlation structure between tests because LD patterns are localized, which considerably increases the efficiency of the procedure. Second, for very significant p-values of genome-wide association studies, the distribution of statistics are not well approximated by the continuous asymptotic distributions such as the Gaussian or χ^2 distribution. I tweak the continuous asymptotic distributions by scaling them to fit to the true discontinuous distributions. This idea considerably increases the accuracy of the procedure. As a result, my method is not only accurate but also very efficient taking only several CPU hours for genome-wide datasets, thereby practically replacing the current gold standard, permutation test.

Estimating power of a study design is an important challenge in the design stage to avoid a under- or over-powered study. For example, if we know that the design is under-powered, we can increase the sample size to meet the desired power. De Bakker *et al.* [18] suggested a *in silico* sampling procedure for estimating power based on the reference dataset [1, 35]. However, this procedure is prohibitively inefficient if we want power estimates for a wide range of possible disease models. On the other hand, the analytical power calculation ignoring the correlation structure is efficient but too conservative, leading to an over-powered study [42, 37].

In Chapter 2, I introduce a novel power estimation method called SLIP (Sliding-window approach with Locally Inter-correlated markers for Power estimation) which takes advantage of the LD information [30]. SLIP is accurate because it takes into account the local correlation structure using the sliding-window approach similar to SLIDE. SLIP is very efficient that it can obtain accurate power estimates of a study design for hundreds of different disease models in several CPU minutes.

Tag SNP selection is a widely-known challenge in the design stage. The objective is to select a subset of the SNPs which will be genotyped so that the statistical power can be maximized. Although the use of commercial genotyping platforms using a fixed set of SNPs are prevalent, tag SNP selection is still of major interest in the area of follow-up candidate gene studies and the design of genotyping platforms. Previous approaches make use of the LD information [10, 60, 18], but does

not fully utilize the information because they ignore the allele frequencies which also affect the power.

In Chapter 3, I introduce a tag SNP selection method which greedily picks SNPs to maximize the statistical power [29]. Since my method directly uses the statistical power itself as a selection criterion, my method outperforms other correlation only-based methods. My method can be thought of as utilizing the full LD information because the information of statistical power is extracted from the LD by taking into account the allele frequencies.

Spurious association detection is an important problem in the analysis stage. Spurious associations can occur due to many causes such as genotyping errors. If we fail to detect them, the overall false positive rate is no longer controlled as desired. The current quality control (QC) approach removes potentially problematic markers based on the genotype data to avoid spurious associations. However, spurious associations can still occur because QC cannot detect a marker with a small number of errors which can also cause a spurious association if the errors occur unequally between cases and controls. For this reason, quality refinement at the post-association stage can provide an added protection against spurious associations. What is typically performed at the post-association stage is to detect spurious associations using the LD information, based on the observation that the nearby markers in LD should show comparably significant p-values if the association is caused by true genetic effect. However, this procedure is only performed manually and there is no existing statistical framework for this procedure [74].

In Chapter 4, I propose a new approach which detects spurious associations based on the LD information called RESQUE (REsults QUality Enhancement), which is a formalization of what is typically performed in studies using intuition. Given a statistically significant association, RESQUE computes likelihood of the observations at nearby markers under the two models, say, true and spurious association models. To the best of my knowledge, RESQUE is the first formal statistical framework for quality refinement at the post-association stage. RESQUE is applied to the Wellcome Trust Case Control Consortium (WTCCC) dataset [74] to find 6 putative spurious associations for type 1 diabetes in the MHC region of chromosome 6.

Finally, I address a potential drawback of the current QC approach that QC often can be too conservative. For example, markers with high missing rate ($> 5\%$) are typically removed by QC. However, if the missings occur independently of the alleles or case/control status, this is just equivalent to having a smaller sample. Therefore, removing these markers may take away the chance to detect true genetic associations at these markers. Given that many genetic association studies find only a handful of associations, the result of this conservative QC approach can be devastating.

In Chapter 4, I propose a new study scheme which "rescues" the possible true associations among the markers excluded by QC. The markers excluded by QC are not removed but tested for associations, and "rescued" if the strong evidence of true associations are detected based on the LD information. The evidence of true associations can be found by extending the framework of RESQUE, because RESQUE is based on the likelihood ratio between two models. To the best of my knowledge, this is the first attempt to find meaningful results from the markers which otherwise would be discarded. My method is applied to the WTCCC data and find 101 new associations among the markers excluded by the WTCCC QC. Interestingly, one association for type 1 diabetes at PGM1 gene is recently found as significant by a meta-analysis performed 2 years after the WTCCC analysis [4], showing that my method can actually increase the study power.

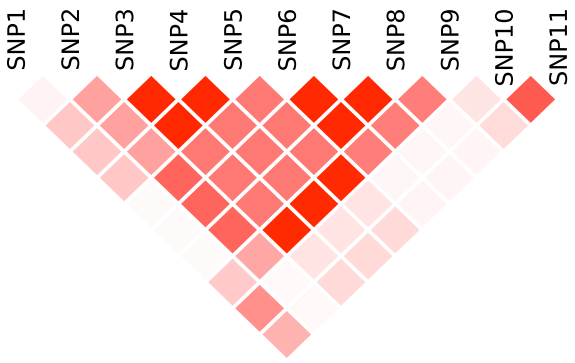


Figure 1.1: Example LD patterns of 10 SNPs. Red colors denote high correlations (r^2 close to 1) and white colors denote low correlations (r^2 close to 0).

1. Design
    - Budget and study plan
    - Sample design
    - Choosing technologies
2. Data collection
    - Sample collection
    - Genotyping
    - Quality controls
3. Analysis
    - Statistical tests
    - Significance assessment
    - Multiple testing correction
4. follow-up
    - Replication study
    - Biological verification
    - Applications

Figure 1.2: Four major stages of genetic association studies.

Chapter 2

Rapid and Accurate Multiple Testing Correction and Power Estimation for Millions of Correlated Markers

2.1 Motivation

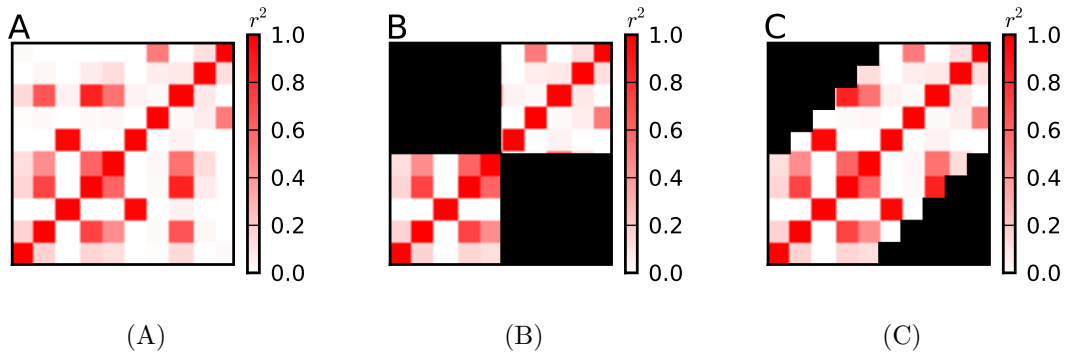
Association studies have emerged as a powerful tool for discovering the genetic basis of human diseases [19, 61, 13]. With the development of sequencing and high-throughput genotyping technologies, the number of single nucleotide polymorphism (SNP) markers genotyped by current association studies is dramatically increasing. The large number of correlated markers brings to the forefront the multiple hypothesis testing correction problem and has motivated much recent activity to address it [14, 22, 52].

There are two common versions of the multiple testing correction problem: per-marker threshold estimation and p-value correction. In a typical study which collects M markers, at each marker, we perform a statistical test and obtain a p-value which we refer to as a *pointwise p-value*. We would like to know how significant a pointwise p-value needs to be in order to obtain a significant result given that we are observing M markers. The *per-marker threshold* can be defined as the threshold

for pointwise p-values which controls the probability of one or more false positives [52]. Similarly, we would like to quantitatively measure the significance of a pointwise p-value taking into account that we are observing M markers. For each pointwise p-value, the *corrected p-value* can be defined as the probability that, under the null hypothesis, a p-value equal to or smaller than the pointwise p-value will be observed at any marker [75]. For example, the Bonferroni correction corrects a pointwise p-value p to pM, or estimates the per-marker threshold as α/M given a significance threshold α.

While the Bonferroni (or Šidák) correction provides the simplest way to correct for multiple testing by assuming independence between markers, permutation testing is widely considered the gold standard for accurately correcting for multiple testing [75]. However, permutation is often computationally intensive for large data sets [14]. For example, running 1 million permutations for a dataset of 500,000 SNPs over 5,000 samples takes up to 4 CPU years using widely used software such as PLINK [59] (See Results). On the other hand, the Bonferroni (or Šidák) correction ignores correlation between markers and leads to an overly conservative correction, which is exacerbated as the marker density increases.

In this paper, we correct for multiple testing using the framework of the multivariate normal distribution (MVN). For many widely used statistical tests, the statistics over multiple markers asymptotically follow a MVN [67, 47]. Using this observation, several recent studies [67, 47, 14] proposed efficient alternative approaches to the permutation test, and showed that they are as accurate as the permutation test for small regions at the size of candidate gene studies (with $< 1\%$ average error in corrected p-values) [14]. However, when applied to genome-wide datasets, they are not as accurate. In our analysis of the Wellcome Trust Case Control Consortium (WTCCC) data [74], these methods eliminate only two-thirds of the error in the corrected p-values relative to the Bonferroni correction. There are two main reasons why these methods do not eliminate all of the error. First, the previous MVN-based methods can be extended to genome-wide analyses only by partitioning the genome into small linkage disequilibrium (LD) blocks and assuming markers in different blocks are independent, because they can handle only up to hundreds of markers in practice [67, 14]. This block-wise strategy leads to conservative estimates



(A) (B) (C)

Figure 2.1: Block-wise strategy and sliding-window approach. (A) Correlations between 10 markers are depicted. (B) Correlations taken into account by a block-wise strategy with a block size of 5. The ignored correlations are shown as black. (C) Correlations taken into account by a sliding-window approach with a window size of 5. The ignored correlations are shown as black.

because inter-block correlations are ignored (Figure 2.1B). Second, these methods do not account for the previously unrecognized phenomenon that the true null distribution of a test statistic often fails to follow the asymptotic distribution at the extreme tails of the distribution, even with thousands of samples.

We propose a method for multiple testing correction called SLIDE (a **S**liding-window approach for **L**ocally **I**nter-correlated markers with asymptotic **D**istribution **E**rrors corrected), which differs from previous methods in two aspects. First, SLIDE uses a sliding-window approach instead of the block-wise strategy. SLIDE approximates the correlation matrix as a band matrix (a matrix with non-zero elements along the diagonal band), which can effectively characterize the overall correlation structure between markers given a sufficiently large bandwidth. Then SLIDE uses a sliding-window Monte-Carlo approach which samples a statistic at each marker by conditioning on the statistics at previous markers within the window, accounting for entire correlation in the band matrix (Figure 2.1C).

Second, SLIDE takes into account the phenomenon that the true null distribution of a test statistic often fails to follow the asymptotic distribution at the tails of the distribution. It is well known that if the sample size is small, the true distribution

and the asymptotic distribution show a discrepancy [79, 77]. However, to the best of our knowledge, the effect of this discrepancy in the context of association studies has not been recognized, since thousands of samples are typically not considered a small sample. We observe that this discrepancy often appears in genome-wide association studies, even with thousands of samples, because of the extremely small genome-wide per-marker threshold (or pointwise p-value). The error caused by this discrepancy is more serious for datasets with a large number of rare variants, highlighting the importance of this problem for association studies based on next-generation sequencing technologies (See Materials and Methods). SLIDE corrects for this error by scaling the asymptotic distribution to fit to the true distribution.

With these two advances, SLIDE is as accurate as the permutation test. In our simulation using the WTCCC dataset [74], the error rate of SLIDE's corrected p-values is more than 20 times smaller than the error rate of previous MVN-based methods' corrected p-values, and 80 times smaller than the error rate of the Bonferroni-corrected p-values. Our simulation using the 2.7 million HapMap SNPs [1] shows that SLIDE is accurate for higher-density marker datasets as well. In contrast, the error rates of previous MVN-based methods increase with the marker density, since the dataset will include more rare variants. Computationally, our simulation shows that SLIDE is orders of magnitude faster than the permutation test and faster than other competing methods.

The MVN framework for multiple testing correction is very general, allowing it to be applied to many different contexts such as quantitative trait mapping or multiple disease models [14]. We show that the MVN framework can also correct for multiple testing for the weighted haplotype test [80, 53] and the test for imputed genotypes based on the posterior probabilities [51].

In addition to multiple testing correction, we extend the MVN framework to solve the problem of estimating the statistical power of an association study with correlated markers. There are two traditional approaches to this problem: a simulation approach constructing case/control panels from the reference dataset [14, 47, 18, 51], which is widely considered the standard but is computationally intensive; and the best-tag Bonferroni method [37, 42, 29], which is an efficient approximation but is often inaccurate.

The power estimation problem can be solved within the MVN framework because the test statistic under the alternative hypothesis follows a MVN centered at the non-centrality parameters (NCP). The vector of the NCPs turns out to be approximately proportional to the vector of correlation coefficients (r) between the causal SNP and the markers. This is a multi-marker generalization of the Pritchard and Preworzki [58] single-marker derivation of the NCP proportional to r. Our method SLIP (**S**liding-window approach for **L**ocally **I**nter-correlated markers for **P**ower estimation) efficiently estimates a study's power using the MVN framework.

Seaman and Müller-Myhsok [67] and Lin [47] pioneered the use of the MVN for multiple testing correction. Seaman and Müller-Myhsok described the direct simulation approach (DSA) method. Conneely and Boehnke [14] increased its efficiency by adapting an available software package called `mvtnorm` [25, 26]. Both studies primarily focused on datasets used in candidate gene studies and suggested the block-wise strategy as a possible approach for genome-wide studies.

Another approach for multiple testing correction is to estimate the effective number of tests from eigenvalues of the correlation matrix [54, 12, 46]. Recently, Moskvina and Schmidt [52] and Pe'er *et al.* [57] showed that the effective number of tests varies by the p-value levels, demonstrating that a method estimating a constant effective number can be inaccurate. Moskvina and Schmidt [52] proposed a pairwise correlation-based method called Keffective, which estimates the effective number taking into account the significance level. Keffective is a sliding-window approach similar to SLIDE, but it differs because within each window it uses the pairwise correlation to the most correlated marker, while SLIDE uses the conditional distribution given all markers. Fitting the minimum p-value distribution by a beta distribution [21] has been shown often to be inaccurate [52]. Kimmel and Shamir [41] developed an importance sampling procedure called rapid association test (RAT). RAT is efficient for correcting very significant p-values, but requires phased haplotype data.

Connecting the multiple testing correction and power estimation problems leads to the insight that the per-marker threshold estimated from the reference dataset for estimating power can be used as a precomputed approximation to the true per-marker threshold for the collected samples. In simulations using the WTCCC control data, we show that the per-marker threshold estimated from the HapMap

CEU population data approximately controls the false positive rate.

Our methods SLIP and SLIDE require only summary statistics such as the correlation between markers within the window size, allele frequencies, and the number of individuals. Therefore unlike the permutation test, our method can still be applied even if the actual genotype data is not accessible. Our methods are available at `http://slide.cs.ucla.edu`.

2.2 Materials and Methods

2.2.1 Multiple testing correction

Multivariate normal approximation

For many widely used statistical tests, the vector of statistics over multiple markers asymptotically follows a MVN [47, 67]. The covariance matrix of the MVN can be derived for many popular statistical tests such as Armitage's trend test in the context of the general score test [14, 67]. We perform this derivation at the haplotype level using the properties of the hypergeometric distribution in the context of the χ^2 test in order to highlight the connection between the multiple testing correction and the power estimation problems. In Text S1 of Han *et al.* [30], we also derive the covariance for the weighted haplotype test [80, 53] and the test for imputed genotypes [51, 64, 49]. All of the results presented here for balanced case/control studies can be extended to unbalanced studies. We will interchangeably use the terms 'covariance matrix' and 'correlation matrix', because the variances are 1.

Assume we permute N case haplotypes and N control haplotypes. Let p_i be the minor allele frequency (MAF) at marker m_i estimated from the sample. Let $\hat{p}_i^+$ and $\hat{p}_i^-$ be the observed MAFs in the permuted case and control haplotypes. Although p_i itself is an observed value from the sample, we will consider it as a constant because it is invariant over random permutations. The minor allele count in the permuted case haplotypes, $N\hat{p}_i^+$, follows a hypergeometric distribution. If N is large, the test statistic at m_i

$$S_i = \sqrt{\frac{2N-1}{4}} \frac{\hat{p}_i^+ - \hat{p}_i^-}{\sqrt{p_i(1-p_i)}} \sim \mathcal{N}(0,1) \; .$$

The squared statistic differs from the Pearson's χ^2 statistic by a constant $\frac{2N-1}{2N}$.

Let S_i and S_j be the statistics at marker m_i and m_j. Let $p_{ij}, p_{Ij}, p_{iJ}, p_{IJ}$ be the sample frequencies of the four haplotypes with minor and major alleles at m_i and m_j respectively. A random permutation is equivalent to selecting N case haplotypes from 4 bins of different haplotypes. Thus, the haplotype count in the permuted case haplotypes, $(N\hat{p}^+_{ij}, N\hat{p}^+_{Ij}, N\hat{p}^+_{iJ}, N\hat{p}^+_{IJ})$, follows a multivariate hypergeometric distribution. By the properties of the hypergeometric distribution,

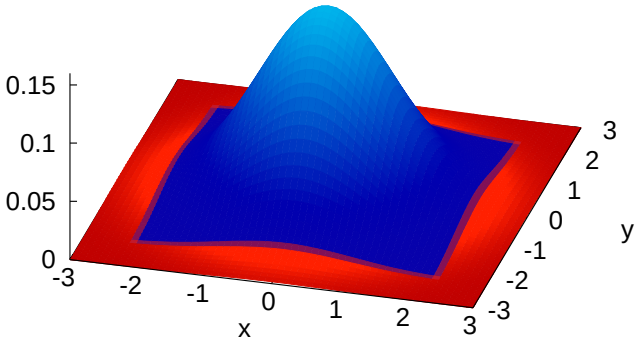
$$\begin{aligned}
\mathrm{Cov}\left(\hat{p}^+_i, \hat{p}^+_j\right) &= \mathrm{Cov}\left(\hat{p}^+_{ij} + \hat{p}^+_{iJ}, \hat{p}^+_{ij} + \hat{p}^+_{Ij}\right) \\
&= \frac{1}{2N-1}\Big(p_{ij}(1-p_{ij}) - p_{iJ}p_{ij} - p_{ij}p_{Ij} - p_{iJ}p_{Ij}\Big) \\
&= \frac{1}{2N-1}(p_{ij} - p_i p_j) \qquad (2.1) \\
\mathrm{Cov}\left(\hat{p}^+_i - \hat{p}^-_i, \hat{p}^+_j - \hat{p}^-_j\right) &= \mathrm{Cov}\left(\hat{p}^+_i - (2p_i - \hat{p}^+_i), \hat{p}^+_j - (2p_i - \hat{p}^+_j)\right) \\
&= 4\mathrm{Cov}\left(\hat{p}^+_i, \hat{p}^+_j\right) \qquad (2.2) \\
\mathrm{Cov}\left(S_i, S_j\right) &= \frac{p_{ij} - p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} = r_{ij} \qquad (2.3)
\end{aligned}$$

where r_{ij} is the correlation coefficient between m_i and m_j measured in the sample.

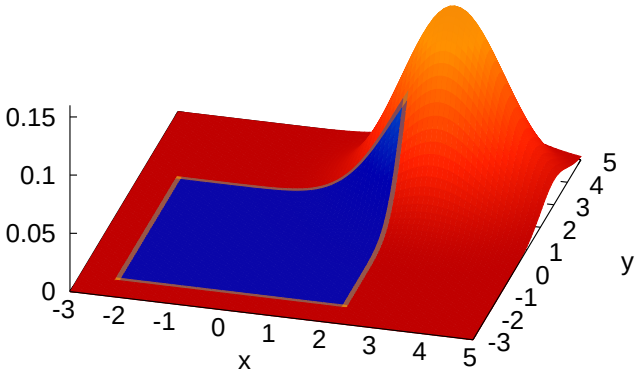
Let $\Sigma = \{r_{ij}\}$ be the $M \times M$ covariance matrix between M markers. By the multivariate central limit theorem [73], if N is large, the vector of statistics $S = (S_1, ..., S_M)$ asymptotically follows a MVN with mean zero and variance Σ. Given a pointwise p-value u, let $R(u)$ be the M-dimensional rectangle with corners $\Phi^{-1}(u/2)\mathbf{1}_M$ and $\Phi^{-1}(1-u/2)\mathbf{1}_M$ where Φ is the cumulative density function (c.d.f.) of the standard normal distribution and $\mathbf{1}_M$ is the vector of M ones. The corrected p-value u' is approximated as the outside-rectangle probability,

$$u' = 1 - \frac{1}{(2\pi)^{\frac{M}{2}}|\Sigma|^{\frac{1}{2}}} \int_{R(u)} e^{-\frac{1}{2}X^T\Sigma^{-1}X} dX \ , \qquad (2.4)$$

as shown in Figure 2.2A. Similarly, given a significance threshold α, the per-marker threshold α_m is approximated by searching for a pointwise p-value whose corrected p-value is α.



(A)



(B)

Figure 2.2: Probability density function of a bivariate MVN at two markers. The area outside the rectangle is the critical region. (A) Under the null hypothesis, the MVN is centered at zero. The outside-rectangle probability is the corrected p-value (or the significance level). (B) Under the alternative hypothesis, the MVN is shifted by the non-centrality parameter. The outside-rectangle probability is power.

Discrepancy between asymptotic and true distributions

If the asymptotic MVN closely approximates the true distribution of the statistic, then Formula (2.4) will provide an accurate multiple testing correction; this has been shown to be true for small regions such as those tested in candidate gene studies [14]. One may expect that the discrepancy between the asymptotic and true distributions would be negligible in current association studies, given their thousands of samples.

However, we observe that this discrepancy can appear in genome-wide association studies, in spite of the large sample size, because of the extremely small per-marker threshold (or pointwise p-value) caused by the large number of tests. At its extreme tails, the asymptotic distribution is typically thicker than the true distribution.

This phenomenon can be illustrated with a single-SNP experiment using the χ^2 test. For a threshold t, the asymptotically approximated p-value (asymptotic p-value) is $p_{\text{asym}} = 2\Phi(-\sqrt{t})$. Assume 1,000 case and 1,000 control haplotypes. Given a fixed number of minor alleles, we can list every possible 2×2 table. The true p-value p_{true} is the sum of the probabilities of the tables whose statistic is $\geq t$. If the asymptotic approximation is accurate, then $p_{\text{asym}} = p_{\text{true}}$. We compare these two p-values for many different thresholds and plot the ratio in Figure 2.3. We repeat the experiments for various MAFs and sample sizes.

Figure 2.3 shows that even with thousands of samples, at the genome-wide significance level, the asymptotic p-value is highly inflated compared to the true p-value. The inflation is more dramatic for SNPs with small MAFs. We observe the similar phenomenon using genotypes and the trend test (data not shown).

One may argue that this phenomenon is not important because it mostly occurs at rare SNPs (MAF $\leq 5\%$) where current studies already have low power to detect associations. However, an incorrect approximation of the distributions at some SNPs affects the corrected p-values of all SNPs. This is because the corrected p-value depends on the distributions of the statistics at all of the SNPs, as it is defined as the probability observing significant results at any marker. For example, suppose we

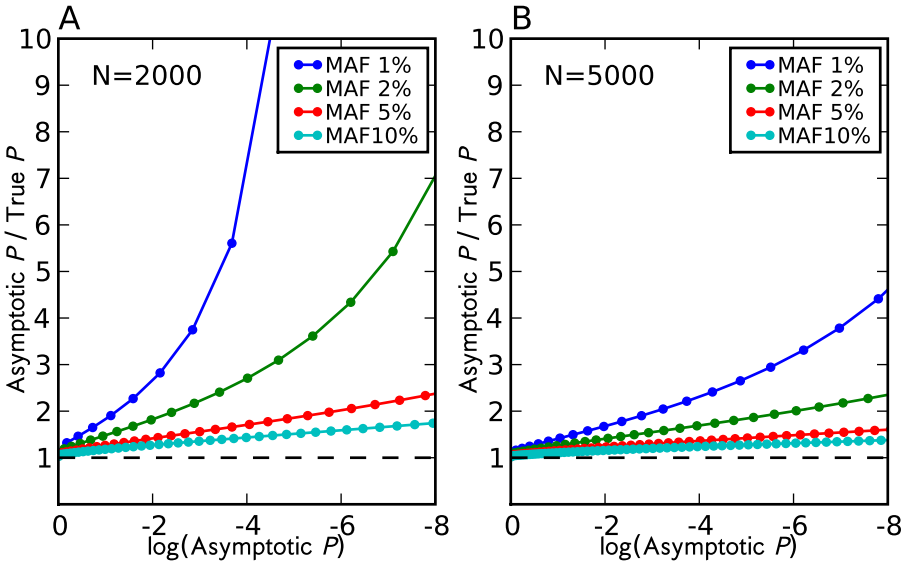


Figure 2.3: Discrepancy between asymptotic p-value and true p-value in a single SNP experiment. Given a χ^2 threshold t, the asymptotic p-value is $2\Phi(-\sqrt{t})$. The true p-value is obtained by listing all possible contingency tables. The number of individuals (N) denotes the number of haplotypes, half control and half case.

approximate 10 independent normal distributions at 10 independent SNPs. Assume that we correctly approximate 9 distributions, but for one distribution we think that the tails are thicker than the true distribution by a factor of 100. For any given pointwise p-value p, the true corrected p-value is $1-(1-p)^{10} \approx 10p$ by the Šidák correction. However, we will estimate the corrected p-value as $1-(1-p)^9(1-100p) \approx 101p$ by integrating over the MVN. This shows that incorrectly approximating the distributions at rare SNPs can adversely affect the corrected p-values of all SNPs, including common SNPs.

One can avoid this type of error in corrected p-values by using a method not dependent on the asymptotic approximation, such as the permutation test, or by eliminating rare SNPs in the analysis. It may be sensible to remove rare SNPs with a few or tens of minor allele counts, if the power is very low or if the SNPs are error-prone in their calling. However, Figure 2.3 shows that the error caused by using the asymptotic approximation happens even at SNPs with minor allele counts in the hundreds. Therefore removing all of them will decrease our power to detect

associations.

SLIDE

SLIDE corrects for multiple testing by using a sliding-window approach to approximate the MVN and then scaling the MVN to approximate the true distribution of the statistic. There are two underlying intuitions. First, a sliding window approach takes into account most of the correlations in the data due to the local LD structure. Second, even though the asymptotic MVN shows a departure from the true distribution at the tail, the scaled MVN will closely approximate the true distribution because the covariance between the statistics is identical in both the true distribution and the MVN. (The covariance derivation does not involve the central limit theorem.)

Step 1 SLIDE first approximates the MVN by using a sliding-window Monte-Carlo approach. Given M markers, let $(S_1, \cdots, S_M)$ be the vector of statistics which asymptotically follows a MVN under the null hypothesis. Let $f(S_1, S_2, \cdots, S_M)$ be the joint probability density function (p.d.f.) of the statistics. Our goal is to generate a large number of samples, $(\widehat{S}_1, \widehat{S}_2, \cdots, \widehat{S}_M)$, to approximate the MVN. If M is very large, the standard sampling approach using the Cholesky decomposition [27] is impractical unless we split the region into small blocks.

Under the local LD assumption, the statistics at distant markers are uncorrelated. Thus, given a window size w, we can assume that S_i is conditionally independent of $S_1, S_2, \cdots, S_{i-w-1}$ given $S_{i-w}, S_{i-w+1}, \cdots, S_{i-1}$. Then by the chain rule,

$$f(S_1, S_2, \cdots, S_M) = f(S_1)f(S_2|S_1)f(S_3|S_1, S_2)\cdots f(S_M|S_{M-w}, \cdots, S_{M-1}) \ .$$

Thus, $\widehat{S}_i$ can be sampled given $\widehat{S}_{i-w}, \cdots, \widehat{S}_{i-1}$, based on the conditional distribution $f(S_i|S_{i-w}, \cdots, S_{i-1})$. The conditional distributions are given by the standard formula for the MVN. Thus we can efficiently generate a large number of samples. The procedure is described in detail in Text S2 of Han *et al.* [30].

Step 2 We scale the approximated MVN to fit to the true distribution of the statistic (Figure 2.4). The rationale for this step is that, if we only consider the marginal distribution at each marker, it is possible to analytically compute the

true distribution by listing all possible 2×2 or 2×3 contingency tables [76]. This allows us to directly compare the asymptotic distribution and the true distribution, and to compute how much we should scale the asymptotic distribution to fit to the true distribution.

The level of discrepancy between the asymptotic and true distributions is large at the tails of the distribution compared to the center. Thus, in order to scale the asymptotic distribution to fit to the true distribution, we cannot multiply the entire distribution by a single scaling factor, but must instead compute the scaling factor for each different threshold.

Given a χ^2 threshold t, we compute the scaling factor as follows. The asymptotic p-value is $2\left(1-\Phi(\sqrt{t})\right)$. Let X be a random variable following the true discrete distribution of the χ^2 statistic. The exact true p-value is $p_{\text{true}} = \Pr(X \geq t)$. The scaling factor is computed as $\sqrt{t}/\Phi^{-1}(1-p_{\text{true}}/2)$, because if we scale the standard normal distribution by this factor, the asymptotic p-value for the scaled distribution becomes exactly p_{true} at the threshold t. In practice, we find that using the so-called mid p-value $p_{\text{mid}} = \Pr(X > t) + \frac{1}{2}\Pr(X = t)$ [76] instead of p_{true} provides a better approximation to the true distribution.

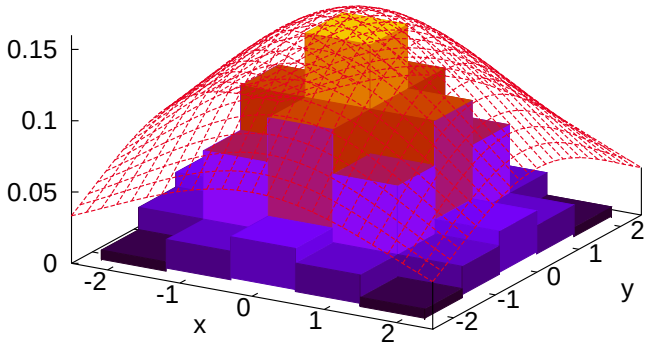
Note that, for unbalanced case/control studies, the level of discrepancy is not symmetric at the upper and lower tails of the normal distribution. Thus, we should compute the scaling factor for each tail of the normal distribution separately.

Step 3 Given the scaled MVN, p-values are corrected by integrating over the outside of the rectangle as in Formula (2.4).

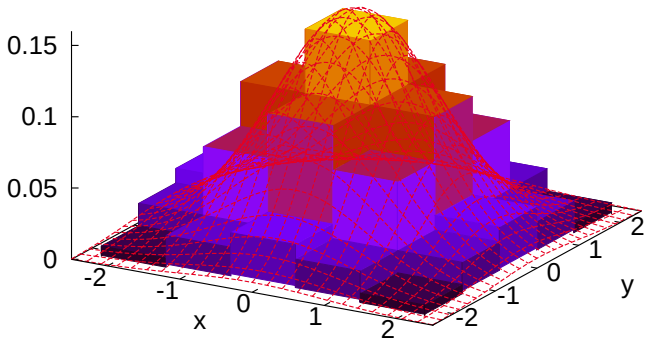
2.2.2 Power estimation

Assumptions

A discussion of association study power depends on many arbitrary assumptions. Though our framework can be extended to other assumptions, in this paper, we adopt those used in De Bakker *et al.* [18]: (1) The disease status is affected by a single SNP. (2) The allele effect is multiplicative. (3) The relative risk is known. (4) The phased reference dataset represents the population. (5) All marker SNPs are in



(A)



(B)

Figure 2.4: SLIDE's scaling procedure. The probability density function of the asymptotic bivariate MVN is depicted as a grid. The probability mass function of the true distribution is depicted as a histogram. (A) The asymptotic distribution often shows a discrepancy from the true distribution. (The discrepancy is exaggerated in this figure.) (B) After scaling down the asymptotic distribution, the discrepancy is removed.

the reference dataset. (6) All possible causal SNPs are in the reference dataset. (7) Each possible causal SNP is equally likely to be causal.

For complex diseases, assumption (1) can still be applied if each causal SNP marginally contributes to the risk. Assumptions (4) and (5) can lead to an overestimation of power, especially if the markers are chosen using the reference dataset [6]. Instead of assumption (7), a non-uniform distribution can also be used [24].

Finally, we assume that the investigator has determined the number of individuals in the study and the significance threshold.

Multivariate normal approximation

We extend the MVN framework to the power estimation problem. Consider a study design which defines markers and plans to collect $N/2$ case and $N/2$ control diploid individuals. Let p_i^h be the population MAF at marker m_i estimated from the reference dataset ('h' denoting the HapMap [1]). Let p_i^+ and p_i^- be the MAFs in the case and control populations.

Single marker If marker m_i is causal for a disease of prevalence F with relative risk γ, under the multiplicative model,

$$p_i^+ = \gamma p_i^h/((\gamma-1)p_i^h+1) \quad \text{and} \quad p_i^- = (p_i^h - Fp_i^+)/(1-F)\ . \tag{2.5}$$

The case/control study can be thought of as a procedure which draws N chromosomes from the case population and N chromosomes from the control population. Let $\hat{p}_i^+$ and $\hat{p}_i^-$ be random variables denoting the observed MAFs in the collected cases and controls. Let $\hat{p}_i = (\hat{p}_i^+ + \hat{p}_i^-)/2$ and $p_i = (p_i^+ + p_i^-)/2$. Then, since each of $N\hat{p}_i^+$ and $N\hat{p}_i^-$ follows a binomial distribution, if N is large, the test statistic at marker m_i

$$S_i = \frac{\hat{p}_i^+ - \hat{p}_i^-}{\sqrt{2/N}\sqrt{\hat{p}_i(1-\hat{p}_i)}} \ \sim\ \mathcal{N}(\lambda_i\sqrt{N}, 1)\ ,$$

where

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

is the non-centrality parameter.

If the marker and the causal SNP are distinct (a condition called *indirect association*), the NCP derivation changes. Suppose a SNP s_c is causal but we collect

marker m_i. If we put an imaginary marker m_c at SNP s_c, we can compute the NCP at marker m_c ($\lambda_c\sqrt{N}$), and compute the correlation coefficient between m_i and m_c from the reference dataset (r^h_{ic}). Pritchard and Preworzki [58] show that the NCP at marker i is approximately $r^h_{ic}\lambda_c\sqrt{N}$.

Multiple markers We examine the covariance between the statistic S_i at marker m_i and S_j at marker m_j given that SNP s_c is causal. Let $p^h_{ij}, p^h_{Ij}, p^h_{iJ}, p^h_{IJ}$ be the haplotype frequencies with minor and major alleles at m_i and m_j respectively, in the overall population. Let $p^+_{ij}, p^+_{Ij}, p^+_{iJ}, p^+_{IJ}$ and $p^-_{ij}, p^-_{Ij}, p^-_{iJ}, p^-_{IJ}$ be the frequencies in the case and control populations.

Collecting cases (or controls) is equivalent to drawing N chromosomes from four possible haplotypes. Thus, the haplotype count in cases,

$$(N\hat{p}^+_{ij}, N\hat{p}^+_{Ij}, N\hat{p}^+_{iJ}, N\hat{p}^+_{IJ}) \ ,$$

follows a multinomial distribution. By the properties of the multinomial distribution,

$$\begin{aligned}
\mathrm{Cov}\left(\hat{p}^+_i, \hat{p}^+_j\right) &= \mathrm{Cov}\left(\hat{p}^+_{ij} + \hat{p}^+_{iJ}, \hat{p}^+_{ij} + \hat{p}^+_{Ij}\right) \\
&= \frac{1}{N}\left(p^+_{ij}(1-p^+_{ij}) - p^+_{iJ}p^+_{ij} - p^+_{ij}p^+_{Ij} - p^+_{iJ}p^+_{Ij}\right) \\
&= \frac{1}{N}(p^+_{ij} - p^+_i p^+_j) \\
\mathrm{Cov}\left(\hat{p}^+_i - \hat{p}^-_i, \hat{p}^+_j - \hat{p}^-_j\right) &= \mathrm{Cov}\left(\hat{p}^+_i, \hat{p}^+_j\right) + \mathrm{Cov}\left(\hat{p}^-_i, \hat{p}^-_j\right) \\
&= \frac{1}{N}(p^+_{ij} - p^+_i p^+_j) + \frac{1}{N}(p^-_{ij} - p^-_i p^-_j) \\
\mathrm{Cov}\left(S_i, S_j\right) &= \frac{(p^+_{ij} - p^+_i p^+_j) + (p^-_{ij} - p^-_i p^-_j)}{2\sqrt{p_i(1-p_i)p_j(1-p_j)}} \qquad (2.6) \\
&\approx \frac{p^h_{ij} - p^h_i p^h_j}{\sqrt{p^h_i(1-p^h_i)p^h_j(1-p^h_j)}} = r^h_{ij} \ , \qquad (2.7)
\end{aligned}$$

where r^h_{ij} is the correlation coefficient between m_i and m_j estimated from the reference dataset.

In practice, approximation in Formula (2.7) usually leads to an accurate power estimate. However, if the relative risk is very large, the Formula (2.6) can be computed exactly and used as follows. By Formula (2.5), we can calculate p^+_c and p^-_c, the MAFs of the causal SNP s_c in the case and control populations. We can then estimate $p_{i|c}$ or $p_{i|C}$, the conditional probability that we will observe the minor allele

at m_i given we observe the minor or major allele at s_c. Note that these conditional probabilities are exactly, not approximately, invariant between cases and controls (See Text S3 of Han *et al.* [30]). Therefore $p_i^+ = p_{i|c}p_c^+ + p_{i|C}(1-p_c^+)$. We can similarly estimate p_i^- and the haplotype frequencies (p_{ij}^+ and p_{ij}^-), which allows us to compute Formula (2.6).

Let $\Sigma^h = \{\text{Cov}\,(S_i, S_j)\}$ be the $M \times M$ covariance matrix between M markers. Let

$$\Lambda_c\sqrt{N} = (\lambda_1, \lambda_2, ..., \lambda_M)\sqrt{N} = \lambda_c\sqrt{N}\left(r_{c1}^h, r_{c2}^h, ..., r_{cM}^h\right) \tag{2.8}$$

be the vector of NCPs induced by the causal SNP s_c. By the multivariate central limit theorem [73], if N is large, the vector of statistics $(S_1, ..., S_M)$ asymptotically follows a MVN with mean $\Lambda_c\sqrt{N}$ and variance Σ^h.

Power depends on the per-marker threshold α_m^h. Given a significance threshold α, α_m^h is set to a level which controls the outside-rectangle probability of the null MVN at α such that

$$\alpha = 1 - \frac{1}{(2\pi)^{\frac{M}{2}}\left|\Sigma^h\right|^{\frac{1}{2}}}\int_{R(\alpha_m^h)} e^{-\frac{1}{2}X^T(\Sigma^h)^{-1}X}dX\ . \tag{2.9}$$

Given α_m^h, the per-causal-SNP power with respect to a causal SNP s_c is the outside-rectangle probability of the alternative MVN,

$$\text{Power}(s_c) = 1 - \frac{1}{(2\pi)^{\frac{M}{2}}\left|\Sigma^h\right|^{\frac{1}{2}}}\int_{R(\alpha_m^h)} e^{-\frac{1}{2}(X-\Lambda_c\sqrt{N})^T(\Sigma^h)^{-1}(X-\Lambda_c\sqrt{N})}dX\ , \tag{2.10}$$

as shown in Figure 2.2B. The average power is obtained by averaging per-causal-SNP powers over all putative causal SNPs.

SLIP

Our method SLIP estimates the power of a study design using the MVN framework. First, like SLIDE, SLIP estimates the per-marker threshold in Formula (2.9) using a sliding window approach. Then SLIP samples causal SNPs, approximates the alternative MVN to estimate the per-causal-SNP power, and averages per-causal-SNP powers over sampled causal SNPs.

Since power is typically larger (e.g. 80%) than a p-value (e.g. .01), a small error in the per-marker threshold barely affects the estimate. Thus, the error caused

by using the asymptotic approximation is negligible. Also, given a causal SNP, we can assume that nearby markers (e.g. those within ±1Mb) can capture most of the statistical power due to local LD. Thus, we can set a window size and only use the markers within that window to estimate the alternative MVN, which will be a n-dimensional marginal MVN if we use n markers.

The computation becomes very efficient if we use approximation (2.7). Since approximation (2.7) states that the covariance is the same for the null and alternative MVNs, we can re-use the null MVN constructed for estimating the per-marker threshold, by shifting it by the NCP to get the alternative MVN. If we re-use the random samples this way, the constructed random samples will be not completely random, as they depend on each other. However, we observe that the inaccuracy caused by this dependency is negligible if we generate a large number of samples for the null MVN. If we re-use the samples, then with almost no additional computational cost, SLIP can generate power estimates for multiple relative risks or study sample sizes, since these only change the NCP.

2.2.3 Multiple testing correction using reference dataset

Multiple testing correction is generally performed using the collected data and not the reference data. Recall that the difference between the per-marker threshold for multiple testing correction (α_m) and the per-marker threshold for power estimation (α_m^h) is that the former is estimated from the collected data, the latter from the reference data. We suggest that multiple testing can be approximately corrected using the reference data, by using α_m^h as a substitute of α_m. The advantage is that we can obtain an idea of the per-marker threshold even before the samples are collected. In Results, we show the accuracy of this approximation using the HapMap data and the WTCCC data.

2.2.4 Genotype data

We downloaded the HapMap genotype data (release 23a, NCBI build 36) from the HapMap project web site [35, 1] and phased the data into haplotypes using HAP [81], which can handle the trio information. We downloaded the case/control

genotype data from the Wellcome Trust Case Control Consortium web site [74] and phased it into haplotypes using Beagle [8].

2.2.5 Web Resources

The URL for methods presented herein is as follows:
`http://slide.cs.ucla.edu`

2.3 Results

2.3.1 Multiple testing correction

P-value correction in Chromosome 22 of WTCCC data

In order to compare how accurately and efficiently different methods correct multiple testing, we simulate a study using the WTCCC data [74]. We use the chromosome 22 data (5,563 SNPs) of the Type 2 diabetes (T2D) case/control study (4,862 individuals). Since not every method can be applied to unphased genotype data, we use haplotype data using the allelic χ^2 test and permutation by chromosomes. We first remove any existing associations by randomly dividing the chromosomes into half cases and half controls. Removing associations is necessary because to correct a pointwise p-value, RAT currently requires an actual SNP with that pointwise p-value to be implanted in the dataset as the most significant SNP.

First, we perform 10M permutations to correct ten different pointwise p-values from 10^{-4} to 10^{-7}, whose corrected p-values are from .04 to .0004. We will consider the corrected p-values by the permutation test as the gold standard, and call them *permutation p-values.* We will assume a method is accurate if its corrected p-values are close to the permutation p-values.

We use SLIDE, DSA, `mvtnorm`, RAT, and Keffective to correct p-values. DSA and `mvtnorm` are MVN-based methods using the block-wise strategy. We use a constant block size (window size) of 100 markers for all methods. Since RAT defines the window size in terms of physical distance, we use 600kb, the average distance of 100 markers in the dataset. We use `-X -e2` option for RAT for an

exact computation of the importance sampling procedure as suggested by Kimmel and Shamir [41]. For every method, we use a large number (>1M) of sampling iterations, which allows 95% confidence interval within $\pm.01p$ for $p = .04$ and $\pm.1p$ for $p = .0004$. Keffective corrects p-values by estimating the effective number of tests for a significance threshold and dividing the pointwise p-values by that number. We use $\alpha = .05$ and window size of 100 for Keffective.

Figure 2.5 shows the ratios between the ten corrected p-values and the permutation p-values. An accurate method will yield a ratio of 1 for all ten different thresholds. The dashed lines denote the area where an accurate method's estimate will be found more than 95% of the time. As expected, the Bonferroni correction is very conservative, overestimating the p-values by 64% on average.

DSA is conservative with an average error of 19%. This is equivalent to reducing the error by only about two thirds relative to the Bonferroni correction. The reasons for the errors include the block-wise strategy ignoring inter-block correlations, and not correcting for the error caused by using the asymptotic approximation. In addition to these errors, `mvtnorm` suffers from an anti-conservative bias which grows as the p-value becomes more significant. This is because the p-value in each block is too small for `mvtnorm` to accurately estimate. Our simulation shows that this anti-conservative bias increases with the number of sampling iterations (data not shown).

Keffective is more accurate and faster than DSA and `mvtnorm`. The average error of Keffective is 10.6%. Note that Keffective is optimized to provide an efficient approximation for the effective number of tests within ∼10% of error. Thus, Keffective is achieving its goal.

Both RAT and SLIDE show accurate estimates with the same average error of 0.8%. Thus, the error rate of SLIDE's corrected p-values is more than 10 times smaller than the error rate of Keffective's corrected p-values, more than 20 times smaller than the the error rate of DSA's corrected p-values, and 80 times smaller than the error rate of the Bonferroni-corrected p-values.

We now explore how each source of error in MVN-based methods – the block-wise strategy and the use of the asymptotic approximation without correction –

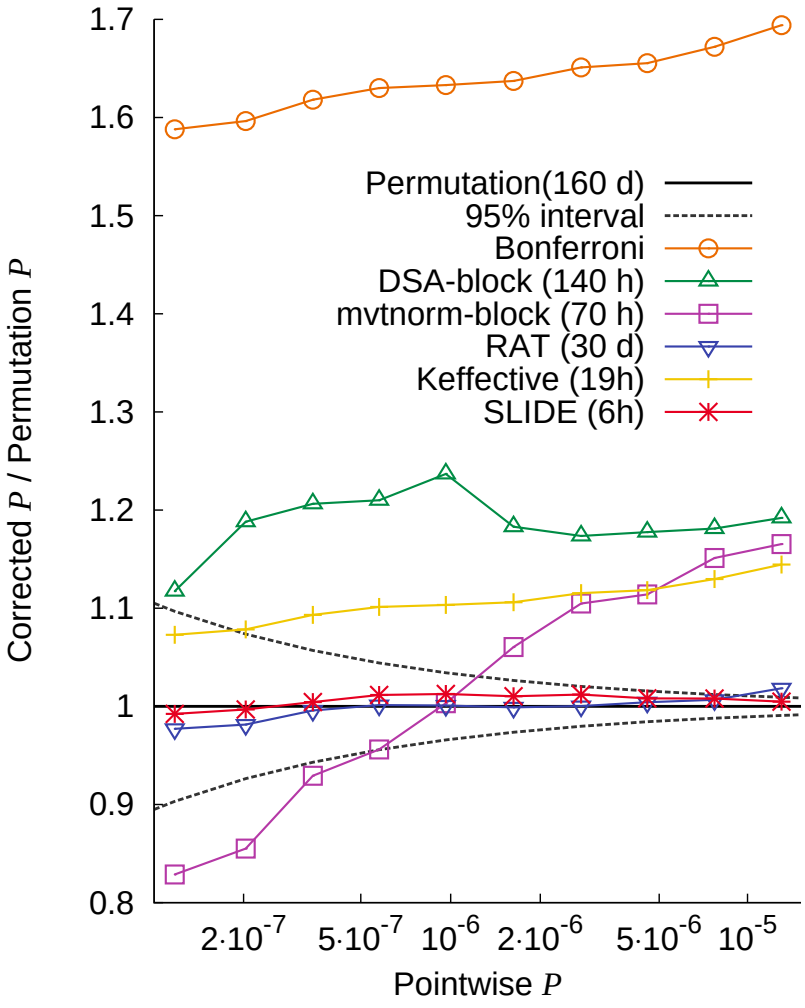


Figure 2.5: Ratios between corrected p-values and permutation p-values for ten different p-value thresholds. We use the WTCCC T2D case/control chromosome 22 data. Approximated time is for correcting 10 p-values with respect to 500K SNPs assuming 100K permutations. The dashed lines denote the interval where an accurate methods' estimate will be found more than 95% of the time.

affects the error rate. We remove 1,048 rare SNPs (MAF $< .05$) and perform multiple testing correction with respect to the remaining 4,515 common SNPs. When considering only common SNPs, the error caused by using the asymptotic approximation will be much smaller (See Materials and Methods). Figure S1 of Han *et al.* [30] shows that the average error of DSA is reduced from 19% to 3.5%, showing that a considerable amount of the error is due to using the asymptotic approximation without correction. The error of Keffective is also reduced from 10.6% to 6.5%. The error of `mvtnorm` is increased from 9.4% to 12.9% because the conservative error caused by using the asymptotic approximation no longer compensates for its anti-conservative bias. SLIDE and RAT are consistently accurate regardless of the exclusion of rare SNPs. Although many methods look relatively accurate when considering only common SNPs, they are inaccurate when considering all SNPs.

Table 2.1 shows the extrapolated running time of each method for correcting p-values with 500K SNPs tested over 5,000 individuals. The running times of RAT, DSA, and `mvtnorm` increases linearly with the number of p-values we correct, since they are currently implemented to correct one p-value at a time (though this may change in future versions). Since Keffective is not a sampling approach, its running time is independent of the number of samples. Given a window size of 100, our time estimate for Keffective (19 h) is similar to the estimate ($\sim$20 h) in Moskvina and Schmidt [52].

In many settings, SLIDE is 500 times faster than the permutation test and considerably faster than the other methods. The running time of SLIDE, Keffective, DSA, and `mvtnorm` is approximately independent of the study sample size, whereas the time of the permutation test is linearly dependent on it. Thus, the efficiency gain of these methods relative to the permutation test will increase as the study size increases. We summarize the accuracy and efficiency of the tested methods in Figure 2.6.

Here we describe a few details of our running time measurements. We used our own C implementation for the permutation test. However, we expect that the

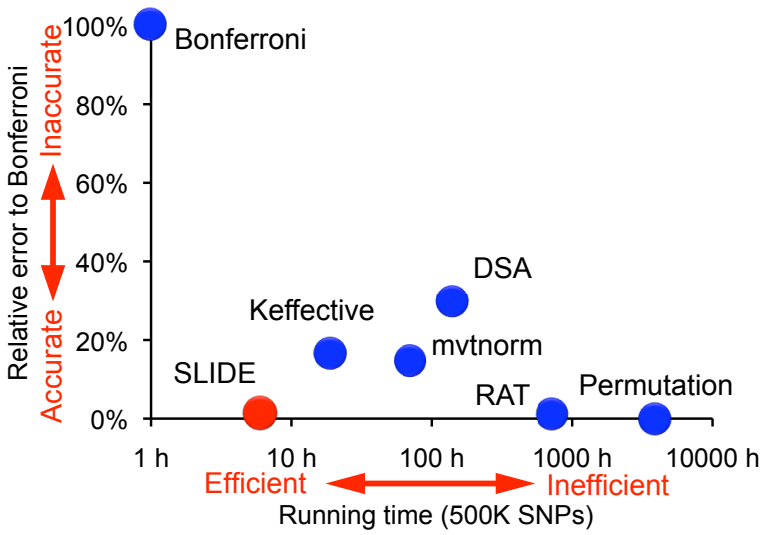


Figure 2.6: SLIDE's accuracy and efficiency compared to other methods. We use the WTCCC T2D case/control chromosome 22 data. The vertical axis is the average error in corrected p-values relative to the Bonferroni correction. The horizontal axis is the approximated time for correcting 10 genome-wide p-values for 500K SNPs assuming 100K permutations.

measured time will be similar to that for commonly used software such as PLINK [59], based on the claimed running time of PLINK on its website (1 CPU-day for 50k permutations over 100K SNPs of 350 samples). Note that PLINK's default "adaptive permutation" is a single SNP permutation to estimate the pointwise p-value, thus its `max(T)` permutation is required for multiple testing correction. Measuring the running time of `mvtnorm` has some subtleties since it has two parameters, the number of samples (`maxpts`) and the absolute error (`abseps`). The procedure is terminated if either the maximum number of samples is reached or the specified error is obtained. Therefore, we set `abseps` to a very small level (10^{-20}) so that the specified number of samples will always be sampled. RAT also has some subtleties involving accuracy and efficiency. If we drop the `-X -e2` parameters for an approximated importance sampling, RAT becomes much faster, but the resulting p-values are underestimated by a factor of up to 5 (data not shown). We assumed a corrected p-value of 10^{-4} to

calculate the number of iterations for RAT using the formula presented in Kimmel and Shamir [41]. Since the formula is conservative, the running time of RAT may be overestimated. The constant window size of 100 may be too large for Keffective, since its purpose is to efficiently approximate the estimate. With a window size of 10, Keffective takes only 2 hours for 500K SNPs. However, if we reduce the window size, the time for other methods including SLIDE will also be reduced.

Using the same WTCCC chromosome 22 dataset, we perform an additional experiment for the unphased genotype data using the trend test, assuming unbalanced case/controls. We find SLIDE achieves similar accuracy (See Text S4 and Figure S2 of Han *et al.* [30]).

Per-marker threshold estimation using all SNPs in HapMap

In this experiment, we assume that a single threshold is being estimated to decide which findings to follow up, instead of correcting each pointwise p-value. We estimate the per-marker threshold corresponding to a significance threshold of .05. We use the 2.7 million polymorphic SNPs in the HapMap CEU data over the whole genome, instead of a single chromosome.

We generate a simulated dataset using the phased haplotype data of 60 HapMap CEU parental individuals. Specifically, we create a new haplotype by randomly shuffling the 120 chromosomes so that the average length of a haplotype segment is approximately 1Mb. We mutate (flip) each SNP with probability 10^{-5}. We create 2,000 cases and 2,000 controls by randomly pairing 8,000 such haplotypes. Although this model is arbitrary, it suffices to compare different methods. The results of the relative comparison between methods do not greatly vary using different parameters, such as a different average haplotype segment length (data not shown).

We compare the permutation test, Keffective, and SLIDE. RAT is not efficient for this setting because it is optimized for very significant p-values, much smaller than .05. We expect that the results of DSA or `mvtnorm` will be similar to or worse than those of Keffective, as in the previous experiment.

We perform 10K permutations for this experiment. We run SLIDE with 10K samplings and window size 100. We run Keffective with window sizes 100 and 10. Figure 2.7 shows the "effective number of tests" estimated by each method,

which is simply the significance threshold (.05) divided by the estimated per-marker threshold. The permutation test estimates the effective number of tests as 1,068,746 out of 2,721,223 tests. Thus, the Bonferroni correction is conservative by 155%. Note that in the previous experiment with a less-dense SNP set, the Bonferroni correction was conservative by 64%. The Bonferroni correction's error will continue to increase with the marker density.

The dashed lines denote the interval where an accurate methods' estimate will be found more than 95% of the time. SLIDE estimates the effective number as 1,038,888 (2.8% error), which is within the 95% interval. This small anti-conservative error is only due to the stochastic error and not an inherent bias, since the result becomes highly accurate as 1,068,445 (0.03% error) if we increase the number of samples to 100K.

Keffective estimates the effective number as 1,409,811 (32% error) with window size 10 and as 1,252,986 (17% error) with window size 100. Unlike the previous experiment, for this higher-density marker dataset, Keffective no longer keeps the error within 10%. We do not expect that a larger window size will increase the accuracy of Keffective, because the error does not seem to be due to the missing long range correlations, since SLIDE is accurate with the same window size of 100.

The running time is 260 hours for permutation, 10 hours for SLIDE, 10 hours for Keffective with window size 10, and 90 hours for Keffective with window size 100.

Window size

Since SLIDE takes into account only correlations within the window size, here we investigate the effect of window size on performance. A reasonable choice for the window size will be the number of markers whose average distance is the average or maximum LD distance in the data. For our experiments, we use the WTCCC T2D case/control chromosome 22 dataset. A large number (10M) of permutations allows us to find that a pointwise p-value 1.53×10^{-5} corresponds to the corrected p-value .05. We correct this pointwise p-value using SLIDE with various window sizes, and see if the corrected p-values are close to .05.

Figure 2.8 shows the ratio between the corrected p-value and the permutation

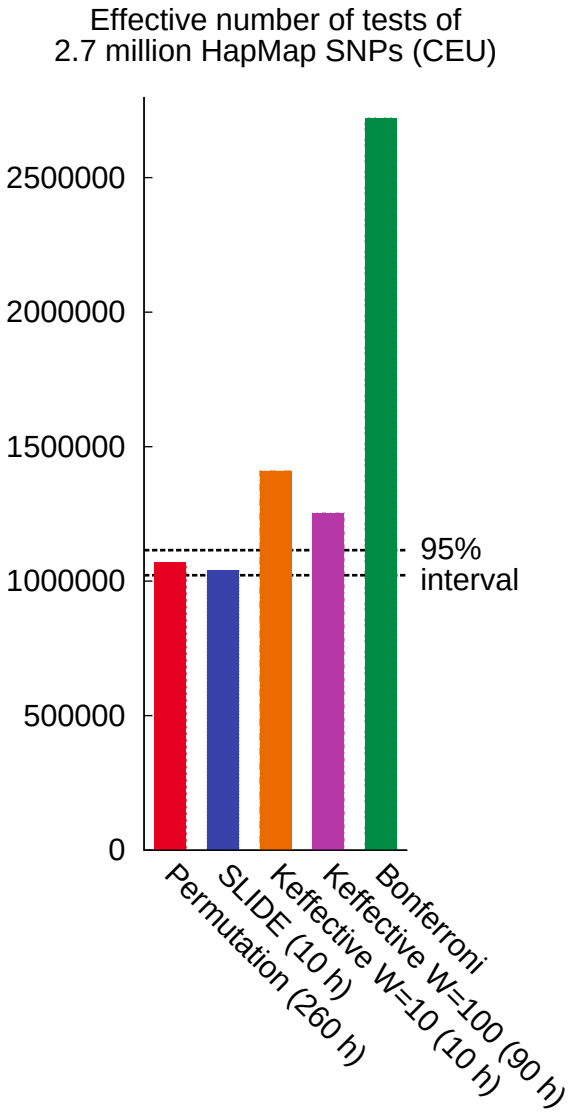


Figure 2.7: Effective number of tests of the 2.7 million HapMap SNPs for a simulated dataset. A dataset of 2,000 cases and 2,000 controls is generated from the HapMap CEU data. Using each method, we estimate the per-marker threshold corresponding to a significance level of .05. The effective number of test is simply .05 divided by the per-marker threshold. The dashed lines denote the interval where an accurate methods' estimate will be found more than 95% of the time.

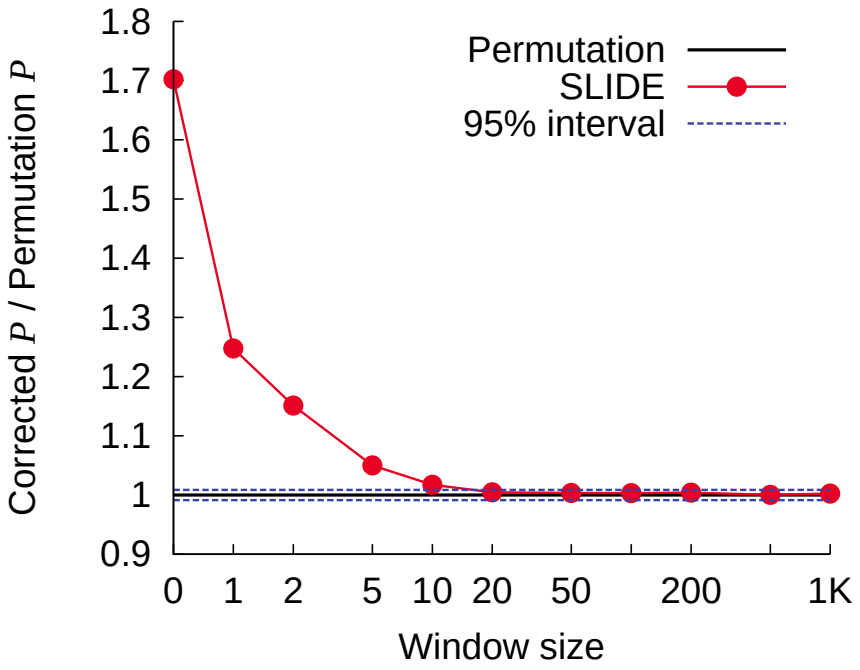


Figure 2.8: Effect of window size on SLIDE's performance. Using the WTCCC T2D case/control chromosome 22 data, we plot the ratios between the corrected p-value and the permutation p-value for varying window sizes for SLIDE. We use the pointwise p-value corresponding to the permutation p-value .05. The window size zero denotes the Bonferroni correction. The dashed lines denote the interval where an accurate methods' estimate will be found more than 95% of the time.

p-value (.05) for various window sizes. Window size zero denotes the Bonferroni correction. The estimate is within the 95% interval for window sizes greater than 20, showing that this is the minimum choice of the window size for this dataset. In this dataset, the average distance between 20, 50, and 100 markers are approximately 100Kb, 300Kb, and 600Kb.

Multiple testing correction using reference dataset

We now examine whether the per-marker threshold estimated from the reference dataset can approximate the true per-marker threshold for a study which may have a different sample correlation structure from the reference dataset. The marker

set we use is the SNPs in the Affymetrix 500K chip over the whole genome.

First, we apply SLIDE to the HapMap data using window size 100, to obtain the per-marker threshold 2.19×10^{-7} corresponding to the significance threshold .05. Then, we permute the WTCCC data to estimate the false positive rate given this per-marker threshold. We use the WTCCC 1958 British birth cohort control data, which consists of 1,504 individuals. We randomly permute the dataset 100K times. We estimate the false positive rate, as the proportion of permutations showing significance given the per-marker threshold, to be .0508. Thus, in this experiment, the per-marker threshold estimated from the reference data controls the false positive rate with only 1.6% relative error. This result shows that, even if the reference population and the target population are slightly different (one from the Utah, U.S.A., and the other from the Great Britain), the per-marker threshold estimated from the reference data is a reasonable approximation.

2.3.2 Power estimation

We compare four different methods for estimating genome-wide power: standard simulation, null/alternative panel construction, best-tag Bonferroni, and SLIP. We assume a multiplicative disease model with a relative risk of 1.2 and a disease prevalence of .01, and a significance threshold of .05. We use the CEU population data in the HapMap as the reference dataset. We use the genome-wide markers in the Affymetrix 500K chip and assume a uniform distribution of causal SNPs over all common SNPs (MAF $\geq$ .05) in the HapMap.

We first perform the standard simulation, which we will consider as the gold standard. We construct a number of genome-wide 'alternative' panels from the HapMap data by randomly assigning a causal SNP for each panel. We permute each panel 1,000 times to estimate the panel-specific per-marker threshold. The power is estimated as the proportion of panels showing significance given its per-marker threshold. Conneely and Boehnke [14] used this procedure for power estimation.

Another panel construction-based approach is the null/alternative panel construction method. Instead of permuting each of alternative panels, this method constructs another set of 'null' panels under the null hypothesis. The null panel

gives us a 'global' per-marker threshold that can be applied to all alternative panels. Since this method is as accurate as the standard simulation but is more efficient, it is widely used [18, 51, 29].

We apply SLIP and re-use the samples for the null MVN for estimating the alternative MVNs. Lastly, we apply the analytical best-tag Bonferroni method [37, 42, 29] which uses the Bonferroni correction for the per-marker threshold and estimates power for each causal SNP by using the most correlated marker (best tag SNP). This method can also be accelerated by sampling the causal SNPs and setting a window size.

For the standard simulation, we use 10K alternative panels. For the null/ alternative panel construction method, we use 10K alternative panels and 10k null panels. For SLIP, we use 10K sampling points. For the best-tag Bonferroni method, we use 10K samples for causal SNPs. For SLIP, we use a window size of 100 markers. For all other methods, we use a window size of 1Mb.

Figure 2.9 shows that both SLIP and the null/alternative panel construction method are as accurate as the standard simulation. The best-tag Bonferroni method is inaccurate, underestimating power by up to 5%.

Table 2.2 shows the running time of each method for estimating genome-wide power. As shown, SLIP is very efficient. Since SLIP uses the correlation structure, the running time is approximately independent of the study sample size, whereas the running time of the standard simulation or the null/alternative panel construction method is linearly dependent on the sample size.

2.4 Discussion

SLIDE and SLIP provide efficient and accurate multiple testing correction and power estimation in the MVN framework. SLIDE shows a near identical accuracy to the permutation test by using a sliding-window approach to account for local correlations, and by correcting for the error caused by using the asymptotic approximation. SLIDE can be applied to datasets of millions of markers with many

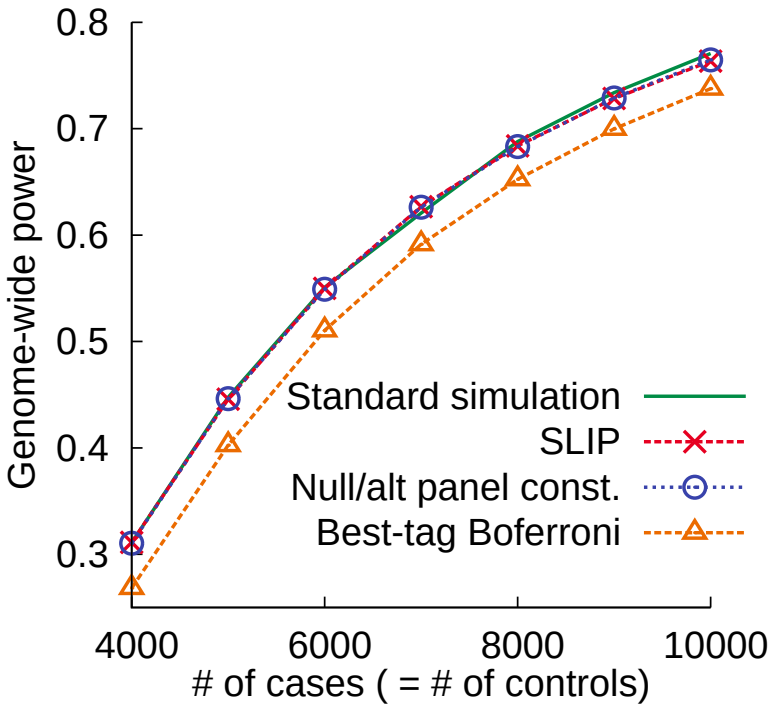


Figure 2.9: Genome-wide power of the Affymetrix 500k chip estimated by different methods. We use the HapMap CEU reference data. We assume a multiplicative disease model with relative risk 1.2, disease prevalence .01, and a uniform distribution of causal SNPs over common SNPs (MAF $\geq$ .05). We use the significance threshold of .05.

rare SNPs, while other MVN-based methods become inaccurate as more rare SNPs are included. To the best of our knowledge, SLIP is the first MVN-based power estimation method.

Throughout this paper, we considered the classical multiple testing correction controlling family-wise error rate (FWER) [75], the probability of observing one or more false positives. SLIDE can be extended to control false discovery rate [5, 69] as well, using a similar approach to Lin [47]. In Text S1 of Han *et al.* [30], we show that the MVN framework can be extended to the weighted haplotype test [80, 53] and the test for imputed genotypes [51]. SLIDE can be use for any multiple testing correction problem with a local correlation structure, as long as the covariance between statistics can be derived.

We considered the permutation test as the gold standard for multiple testing correction. The permutation test can be performed in two different ways: at each permutation, we can either assess the maximum statistics among the markers (max-T permutation), or assess the minimum pointwise p-value among the markers by performing another permutation for each marker (min-P permutation) [75, 69]. We used the former approach because the latter approach is computationally very intensive.

In Text S5 and Figure S3 of Han *et al.* [30], we describe some additional insights obtained through the study. When marker frequencies do not follow the Hardy-Weinberg proportions (HWP), the use of an allelic test (e.g. allelic χ^2 test) for unphased genotype data is not recommended due to the possible bias [63]. However, widely used software [59] often allows the use of an allelic test for genotype data under the reasoning that, as long as the permutation or an exact test is performed, the pointwise p-value will be the same as if we use a genotypic test (e.g. Armitage's trend test). Theoretically, this is due to the fact that the allelic and genotypic test statistics differ only by their variance [20]. However, for assessing corrected p-values, the permutation test does not provide this kind of "protection". Even after a quality control process that excludes SNPs which significantly deviate from the Hardy-Weinberg equilibrium (HWE), still many SNPs may not follow HWP. Therefore, using an allelic test for genotype data for multiple testing correction can result in inaccurate estimates.

Recently, a different view of multiple testing correction has been introduced [57, 22], which suggest that we should correct for the uncollected or unknown markers as well as the collected markers, in order to take into account additional testing burdens such as the possible testings in a follow-up study. Pe'er *et al.* [57] estimates the per-marker threshold by extrapolating from the resequenced ENCODE regions, and Dudbridge *et al.* [22] estimates the per-marker threshold by subsampling the SNPs at an increasing SNP density. Although we employed the classical point of view that corrects for multiple testing only over observed SNPs, our method can also be applied to this alternative view. Our method can be used to estimate the effective number of tests for a representative resequenced region or for the set of subsampled SNPs. Since the SNP density of genotyping technology is dramatically increasing,

we assume that the number of unknown and uncollected SNPs will decrease, causing the two different views to converge.

In our experiments, we used a constant block size for the block-wise strategy. In practice, it will be more reasonable to split the region according to the LD blocks. However, this is not always possible because LD blocks are often ambiguous and some blocks can be larger than the maximum block size of the method. For example, if we collect 10 million SNPs, a block size of 1,000 is required to cover 300kb LD. However, the maximum block size of `mvtnorm` that allows an accurate estimate is currently 300 [14], and DSA with window size 1,000 often requires a prohibitively large memory in our simulations (data not shown). By contrast, SLIDE with window size 1,000 for the WTCCC chromosome 22 data requires $\sim$150 Mb memory and thus is feasible. Nevertheless, it should be noted that the block-wise strategy can always be implemented to have the same block size as SLIDE.

Recently, a method called PRESTO [7] was introduced, which increases the efficiency of the permutation test by applying several optimization techniques. Based on the claimed running time, SLIDE is $\sim$10 times faster than PRESTO, but PRESTO has an advantage that it does not depend on the asymptotic approximation but provides exactly the same result as the permutation test.

We considered the pairwise correlation between SNPs. There can also be so-called higher-order correlations, such as the correlation between a haplotype and a SNP. For example, even though three SNPs are pairwisely independent, the combination of the first two SNPs can be a perfect proxy to the third SNP. However, the multivariate central limit theorem proves that the joint distribution of the test statistics is fully characterized by the matrix of the pairwise correlations. Thus, the effect of the other correlation terms on the joint distribution is asymptotically negligible. Nevertheless, our method is not limited to the SNP test. If our method is applied to the weighted haplotype test [80, 53] as shown in Text S1 of Han *et al.* [30], the pairwise correlation in the correlation matrix can be interpreted as the higher-order correlations between a haplotype and a SNP or between haplotypes.

In summary, SLIP and SLIDE are two useful methods for genome-wide association studies which provide accurate power estimation at the design step and accurate multiple testing correction at the analysis step. The software is available

as a resource for the research community.

Chapter 2 was published in PLoS Genetics, 5(4):e1000456, 2009, Buhm Han, Hyun Min Kang, and Eleazar Eskin. “Rapid and accurate multiple testing correction and power estimation for millions of correlated markers”. The dissertation author was the primary investigator and author of this paper.

Table 2.1: Running time for correcting genome-wide p-values in a study with 500K SNPs over 5,000 individuals. $*$ denotes a method which is often anti-conservative. All values are extrapolated from the chromosome 22 results.

Procedure	# Permutations	Permutation	SLIDE	DSA	`mvtnorm`*	RAT	Keffective
Correcting 1 p-value	10K	16 d	0.6 h	1.4 h	0.7 h	7 h	19 h
Correcting 10 p-values	10K	16 d	0.6 h	14 h	7 h	70 h	19 h
Correcting 1 p-value	100K	160 d	6 h	14 h	7 h	72 h	19 h
Correcting 10 p-values	100K	160 d	6 h	140 h	70 h	30 d	19 h
Correcting 1 p-value	1M	4 years	3 d	6 d	3 d	30 d	19 h
Correcting 10 p-values	1M	4 years	3 d	60 d	30 d	300 d	19 h

Table 2.2: Running time for estimating genome-wide power with 10K samplings.
$*$ denotes a method which is inaccurate (average error is not within 1%).

Procedure	#cases/controls	Best-tag-Bonf.*	SLIP	Null/altern.	Std. simul.
Estimating power	1,000/1,000	0.1 h	0.6 h	36 h	10 d
	5,000/5,000	0.1 h	0.6 h	8 d	50 d
Estimating power	1,000/1,000	0.1 h	0.6 h	8 d	50 d
for 5 different relative risks	5,000/5,000	0.1 h	0.6 h	40 d	250 d

Chapter 3

Efficient Association Study Design Via Power-Optimized Tag SNP Selection

3.1 Motivation

Discovering statistical correlation between causal genetic variation and clinical traits through association studies is an important method for identifying the genetic basis of human disease [61, 2]. Typically, a genetic association study gathers case/control individuals, collects genetic variation information such as genotypes at single nucleotide polymorphisms (SNPs), and tests the significance of association for each SNP using a statistical test such as a χ^2 test. Since fully resequencing a cohort is prohibitively costly, a set of representative SNPs (called tags or tag SNPs) are chosen as proxies for nearby SNPs, utilizing the local correlation structure of SNPs (or linkage disequilibrium) to find associations [58]. While many current association studies are performed using commercially available high-throughput genotyping products that define a set of tag SNPs, selection of these SNPs remains an important problem for both custom follow-up studies as well as designing the high-throughput genotyping products themselves [70, 71, 18, 15, 16, 28, 48, 55, 60, 62, 10].

In the context of association studies, maximizing statistical power is the most relevant goal of tag SNP selection. Since the actual causal SNP is not known, the

statistical power of an association study is defined as the average power over all possible causal SNPs. Recent availability of reference data sets such as the HapMap [32, 31, 1] allows us to empirically measure power of an association study design [18, 56, 43]. A standard method for picking tags is greedily choosing the smallest number of SNPs with a minimum cut-off of correlation (r^2) between tag SNPs and uncollected SNPs [18, 10, 31, 1]. However, choosing tag SNPs based on r^2 alone does not necessarily maximize power, because r^2 does not take into account minor allele frequency (MAF) which also influences power.

In this paper, we present a flexible study design framework that chooses tag SNPs to maximize the statistical power of an association study.

The underlying intuition is that we quickly find the "key tag SNPs" that contribute a considerable amount of power. The power a tag SNP contributes depends on (1) the coverage of a tag SNP (the number of putative causal SNPs a tag SNP can cover), (2) the correlation (the r^2 between a tag SNP and each causal SNP it covers), and (3) the MAF of each causal SNP. We observe that r^2-based methods do not consider (3) and maximize (1) by setting (2) to a fixed threshold. Instead, we use a greedy procedure that evaluates each candidate tag SNP's possible average power increase, and selects the best SNP as a tag SNP at each step. By evaluating the average power increase, we take into account all three aspects of a tag SNP. By not fixing a minimum value of r^2, we allow more flexibility in selecting a tag SNP of maximum power. For example, if a tag SNP has a low r^2 to causal SNPs but covers many common SNPs (bad at (2) but good at (1) and (3)), we can select the SNP based on the power increase unlike the r^2-based methods.

Empirical simulations based on the HapMap ENCODE regions show that our power-optimized method requires 21% fewer tag SNPs on average than widely used r^2-based methods, to achieve equivalent power. When applied to whole genome association mapping, our power-optimized tag sets consistently outperform the r^2-based tag sets across all populations. We compare our designs to the commercial products as well. Our 100k tag set provides equivalent power to the Affymetrix 500k chip for European and Asian populations. In addition, our 300k tag set outperforms the Illumina 550k chip across all three HapMap populations. We apply our method to the custom follow-up study design problem where the goal is to select tag SNPs in

addition to those already present on a commercial product to maximize the statistical power within a region of interest. Our method provides up to twice the power increase using the same number of additional tag SNPs compared to the widely used r^2-based methods.

Since study parameters such as relative risk are generally unknown, a possible pitfall of using statistical power instead of a study-independent measure such as r^2 is "fitting" the design to an incorrect parameter. We show that when the parameters are correct our method performs optimally, and when the parameters are incorrect our method still outperforms or performs similarly to the widely used r^2-based methods, within a wide range of parameters.

During the course of design, our procedure requires us to evaluate the power of candidate tag sets numerous times, thus the use of empirical simulation for measuring power [18] is computationally impractical. We combine the use of an analytical approximation for the power in our tag selection method with an efficient empirical simulation that can accurately measure the power of a tag set. The efficiency of our method allows us to design an association study in one ENCODE region in 3 seconds and a genome wide study in 1.5 CPU hours. The empirical simulation for accurately measuring power is based on a standard technique described in de Bakker *et al.* [18]. To the best of our knowledge, no one has analyzed this standard simulation procedure with respect to its accuracy. We improve the efficiency of this simulation and scale it to the whole genome using a sampling procedure, for which we derive the corresponding confidence intervals. This allows us to determine the number of sampling iterations required for a given level of accuracy. The key insight in this sampling procedure is that the variance of the estimate of the power is independent of the shape of the distribution of the true power over the causal SNPs.

Previous works in tag SNP selection include haplotype-based methods [36, 70, 71, 48, 28], correlation-based methods [10, 60, 1, 18], and power-based methods [9, 55, 15, 16, 62]. The correlation-based methods are power-based methods in that r^2 is closely related to power [58], but here we group power-based methods separately based on whether MAF is taken into account. Among the power-based methods, Byng *et al.* [9] and Pardi *et al.* [55] use the generalized linear model to test the association between the region of interest and the disease. Their approach is different

from ours which considers single SNP association for each SNP in order to detect and locate the association. Cousin *et al.* (2003, 2006) maximizes average power over all possible parameters, specifically over a relative penetrance from 0 to 1 which corresponds to the relative risk from 1 to ∞. Since such a high relative risk is often of little interest in the current association studies, and since sometimes the relative risk can be approximated from the previous studies, our method can be more suitable for those cases by allowing a flexible choice of parameter values or ranges. Saccone *et al.* [62] focus on the observation that the power is affected by the phase of the correlation, whether a tag SNP and the causal SNP are correlated positively or negatively. However, if we use r^2 as a correlation measure instead of D' they use [19], the power is approximately independent of the phase of correlation [58]. Thus, selecting tag SNPs based on the phase may not maximize power.

The implementation of our method is publicly available via web server at `http://design.cs.ucla.edu`. On this web site, we provide power analysis for all popular commercial products as well as candidate gene study designs for every gene in the human genome.

3.2 Results

3.2.1 Performance

We evaluate the performance of our power-optimized method by comparing it to those of widely used r^2-based methods. *Pairwise* r^2 *tagging* [10, 18, 1] is the most common r^2-based method. It greedily selects tags until every SNP is covered with a given minimum r^2 threshold. *Best-N* r^2 is another r^2-based method [18]. It greedily selects a fixed number of tags to cover as many SNPs as possible with a given minimum r^2 threshold. We use the HapMap ENCODE regions which consist of ten 500kb regions that have been widely used to evaluate design methodologies due to their complete ascertainment of common SNPs (MAF $\geq 5\%$).

In this experiment and throughout this paper, we assume a multiplicative disease model with fixed relative risk of 1.2 and disease prevalence of 0.01. We assume a uniform distribution of causal SNPs over all common SNPs defined by a

5% MAF threshold, and use a 5% region-wide significance level (α) for statistical tests. We note that other studies often assume a varying relative risk depending on MAF. For example, Marchini *et al.* [51] and de Bakker *et al.* [18] set a relative risk so that a single SNP can have a 95% of nominal power at a nominal significance level of 1% (ignoring multiple hypothesis testing). This corresponds to a relative risk of 1.21 for a SNP of 50% MAF and a relative risk of 1.48 for a SNP of 5% MAF when 4,000 cases and 4,000 controls are used. In this paper, we assume a uniform relative risk of 1.2, to evaluate the worst case power over all disease models with relative risk of 1.2 or above. This model is often more realistic than the varying relative risk model for the case that the relative risk is estimated from previous studies. These assumptions are used in both analytically designing tag sets and empirically measuring their power.

First, we consider the ENr232 ENCODE region containing 533 (CEU), 596 (CHB), 573 (JPT), and 740 (YRI) common SNPs. The full set of common SNPs achieves the maximum possible power. We will call this the full-SNP-set power. For each population, assuming 4,000 cases and 4,000 controls (= 8,000/8,000 chromosomes), we use our power-optimized method to construct 100 different tag sets of increasing size. The number of tags in each tag set is increased by 1% of total common SNPs.

For comparison, we construct another 100 tag sets of similar size using pairwise r^2 tagging. Since we can only vary the r^2 threshold in pairwise r^2 tagging, we use binary-search over the r^2 threshold with a precision of 0.001, to find a tag set having the desired size as closely as possible. Then we construct another 100 tag sets using best-N r^2. We use a widely used threshold of $r^2 = 0.8$ for best-N r^2. We will use this threshold for every experiment using best-N r^2.

For each tag set, we use the standard empirical simulation for estimating power [18]. We create 100,000 null panels for multiple hypothesis correction and 100,000 alternate panels for estimating power, which gives a 95% confidence interval for a <0.6% error in power. We will use the same number of panels whenever we measure the power of a tag set through the paper.

Figure 3.1 shows the results of the CEU population. The results of the other populations are shown in Supplementary Figure of Han *et al.* [29]. The power-

optimized method reaches the full-SNP-set power (dashed horizontal line) faster than both r^2-based methods. The range of the number of tag SNPs (x-axis) is shown from zero up to the required number of tag SNPs for best-N r^2 to cover every SNP. Thus, at the end of the graph, pairwise r^2 tagging and best-N r^2 become an equivalent procedure, where the threshold of pairwise r^2 tagging happens to be 0.8, and best-N r^2 happens to cover every SNP with $r^2 = 0.8$. To achieve 95% of full-SNP-set power, our power-optimized method requires 37, 71, 57, and 207 SNPs while pairwise r^2 tagging requires 85, 124, 120, and 259 SNPs and best-N r^2 requires 37, 89, 80, and 310 SNPs in the CEU, CHB, JPT, and YRI populations respectively. Pairwise r^2 tagging shows low power with a small number of tags since the r^2 has to be very low to cover every SNP. It has reported that overly lowering r^2 threshold of pairwise r^2 tagging may result in a performance not better than a random tags [18]. Best-N r^2 shows a good performance with a small number of tags, although not better than our power-optimized method, and often shows a worse performance than pairwise r^2 tagging with a large number of tags, as shown in Figure 3.1.

Next, we consider all ten ENCODE regions and obtain similar results (Supplementary Figure of Han *et al.* [29]). We report the fraction of SNPs required to achieve 95% of full-SNP-set power in each region (Supplementary Figure of Han *et al.* [29]). The power-optimized method reduces the required number of tag SNPs by 60.0% compared to pairwise r^2 tagging and 20.9% compared to best-N r^2 on average over all populations and regions.

3.2.2 r^2 and power distribution

We examine the underlying reasons why our power-optimized method achieves equivalent power using fewer tag SNPs than the r^2-based methods. From the previous experiment in the ENr232 ENCODE region of the CEU population, we select three tag sets designed by each of power-optimized method, pairwise r^2 tagging, and best-N r^2. For each method, we select the smallest tag set which achieves the same 99% of the full-SNP-set power. These are 85 tag SNPs designed by power-optimized method, 123 tag SNPs designed by pairwise r^2 tagging, and 138 tag SNPs designed by best-

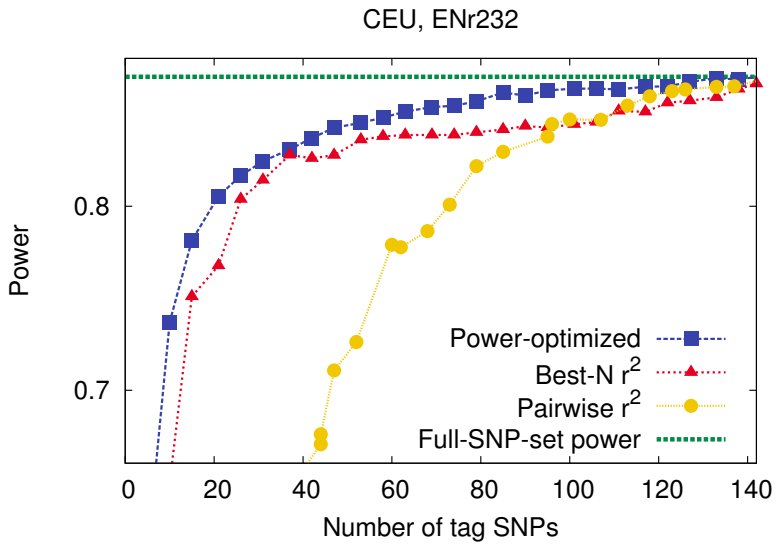


Figure 3.1: Power comparison between our power-optimized tag SNP selection method and a widely used r^2-based methods, pairwise r^2 tagging and best-N r^2, in the ENr232 ENCODE region of the CEU population. We use a 5% region-wide significance level, a 5% MAF threshold for causal SNPs, and assume relative risk of 1.2, disease prevalence of 0.01, and 4,000 cases and 4,000 controls. We use the r^2 threshold of 0.8 for best-N r^2. The x-axis ranges up to the number of tags obtained by best-N r^2 to cover every SNP with $r^2 = 0.8$. The purple horizontal dashed line indicates the full-SNP-set power achievable by genotyping the full set of SNPs.

N r^2. They have almost the same power of 86.2%, 86.3%, and 86.4% respectively. Pairwise r^2 tagging is designed with $r^2 = 0.703$.

In order to analyze the performance of each tag set, we measure the tag set's maximum r^2 to each putative causal SNP and the tag set's power to detect each putative causal SNP (per-causal-SNP power). We group the causal SNPs into three groups based on their MAF: infrequent (5-10%), semi-frequent (10-25%), frequent (25-50%), which contain 74, 150, 299 SNPs respectively. We plot the r^2 and per-causal-SNP power distribution in Figure 3.2. In the infrequent group, the r^2 distribution of our power-optimized method is not very concentrated on the high level compared to the r^2-based methods. In this group, the average r^2 of our power-optimized method is 0.75 while those of pairwise r^2 tagging and best-N r^2 are very high at 0.98 and 0.99 respectively. However, compared to the r^2 difference, the average power of our method is 36% which is not much lower than the 40% of the two r^2-based methods. In the semi-frequent group, the average r^2 of power-optimized method is 0.94 which is slightly lower than 0.95 of the two r^2-based methods, but the average power is 88% which is slightly higher than 87% of the r^2-based methods. In the frequent group, the average r^2 of our method, pairwise r^2 tagging, and best-N r^2 are 0.93, 0.91, and 0.94, and the average power estimates are 98%, 97%, and 98% respectively.

The reason that in the infrequent or semi-frequent group our method achieves comparable or higher power with lower average r^2 is because our method takes into account MAF in selecting tag SNPs. If a causal SNP has a high MAF, the SNP is worth covering with high r^2 because the power will significantly increase. If a causal SNP has a low MAF, the SNP might not be worth covering with high r^2 because the power will still be low. In that case, we can allow the SNP to be covered with low r^2 without much power loss. Within a MAF group, our method strategically covers the SNPs of relatively high MAF with high r^2, thus having high power with low average r^2. This strategy is applied across MAF groups as well. Often, it can be possible to gain more power by spending a tag SNP to cover the SNPs in the frequent group than the SNPs in the infrequent group. Although our method has lower power than r^2-based methods in the infrequent group, our method successfully covers SNPs

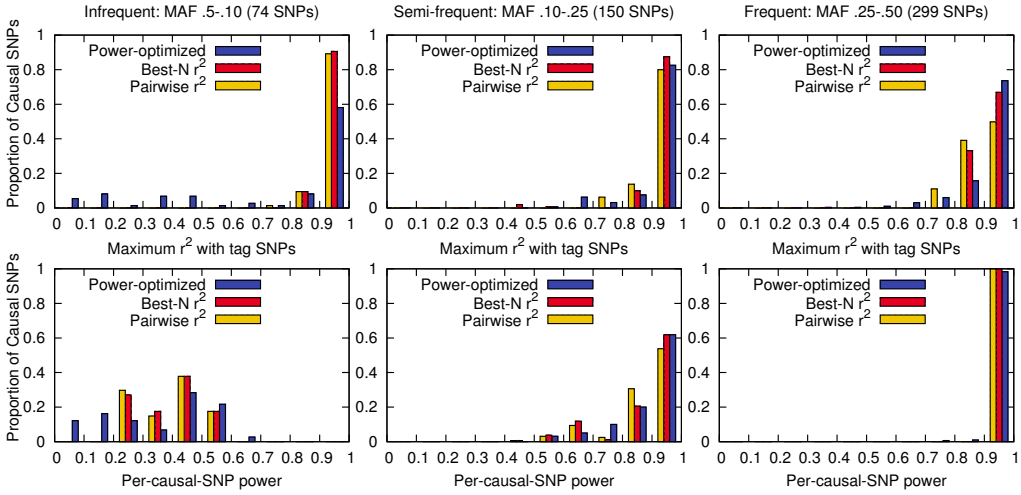


Figure 3.2: Maximum r^2 and per-causal-SNP power distribution over all 533 causal SNPs in the ENr232 ENCODE region of the CEU population. We divide the causal SNPs into three groups by their MAF: infrequent (5%-10%), semi-frequent (10%-25%), frequent (25%-50%), which contain 74, 150, 299 SNPs respectively. Each bar of different colors represents the power-optimized tagging method (85 tag SNPs), pairwise r^2 tagging method (123 tag SNPs), and best-N r^2 tagging method (138 tag SNPs). The three tag sets achieve the same 99% of full-SNP-set power.

in the semi-frequent and frequent groups with high r^2. Consequently, our method achieves the same average power with much fewer tag SNPs than pairwise r^2 tagging and best-N r^2, reducing the tag set size by 31% and 38% compared to those methods respectively.

We note that covering SNPs of high MAF with high r^2 is not the only behaviour of our method. If the per-causal-SNP power is saturated to 100%, then it can be possible to cover the causal SNP with moderate r^2 and still have 100% or very high power. In that case, our method strategically loosens the r^2 for that SNP so that it can spend the tag SNP for other SNPs which would increase power with high r^2. All these decisions are automatically made based on the average power increase.

3.2.3 Robustness

Our power-optimized tag SNP sets depend on the choice of study parameters such as relative risk and number of individuals. One concern with this approach is the potential for a performance drop due to using incorrect parameters. If the true relative risk is higher than expected, then some tag SNPs are wasted on common SNPs that already have very high power. If the true relative risk is lower than expected, then some tag SNPs are wasted on rare SNPs that are too difficult to capture even with higher r^2. We evaluate this performance drop with two experiments, and show that our method still performs better than or similarly to the r^2-based methods in most cases. Both experiments are performed in the ENr232 ENCODE region of the CEU population.

In the first experiment, we design three different tag sets assuming relative risks of 1.1, 1.2, 1.4, and measure their power based on the assumption of a relative risk of 1.2. For each relative risk, we select 100 tag SNPs assuming 4,000 cases and 4,000 controls. For comparison, we design tag sets of the same size using pairwise r^2 tagging and best-N r^2. As shown in Figure 3.1, the two r^2-based methods have similar power at this number of tag SNPs. Figure 3.3 shows the results. As expected, the tag set based on a correct relative risk (1.2) works better than the tag sets based on incorrect relative risks (1.1 and 1.4) at the number of individuals assumed in the design (4,000 cases and 4,000 controls). As the number of individuals decreases, the

tag set based on a lower relative risk (1.1) shows the highest power amongst the tag sets. This is because lowering the number of individuals has the same effect on the test statistic as lowering the relative risk. For the same reason, as the number of individuals increases, the tag set based on a higher relative risk (1.4) obtains the highest power amongst the tag sets.

At the number of individuals assumed in the design, even though r^2 is incorrectly assumed in the design, our method works similarly to the r^2-based methods. If r^2 is correctly assumed in the design (line with diamond), even though the number of individuals varies, our method works similarly to the r^2-based methods. Our method works comparably to the r^2-based methods for a wide range of parameters, except for the extreme case that the bias of two parameters affect the statistic in the same direction, for example a smaller relative risk (1.1) is assumed in the design and a large number of individuals (8,000 cases and 8,000 controls) are used.

In the second experiment, we use the tag set based on a relative risk of 1.2 and the tag sets designed by r^2-based methods from the previous experiment. We measure the power of the tag sets assuming 20 different relative risks from 1.0 to 1.5, and 160 different study sizes from 0 cases and 0 controls to 8,000 cases and 8,000 controls. Figure 3.4 shows the power difference between our method and r^2-based methods over the two-dimensional parameter space (total 3,200 points). As expected, an optimal power gain is obtained when the parameters that the design is based on (γ=1.2 and 4,000 cases and 4,000 controls) or equivalent designs are applied (diagonal red curve). In this experiment, our method performs better than pairwise r^2 tagging when the actual effect size is smaller than assumed (lower left plane), and better than best-N r^2 when the actual effect size is larger than assumed (upper right plane). For both comparisons, our design works better than or similarly to the r^2-based methods within a wide range of parameters.

Varying study parameters such as relative risk, sample size, disease prevalence, and significance level can all be interpreted as varying the effect size, which can be thought of intuitively as the difference in the test statistic between the null

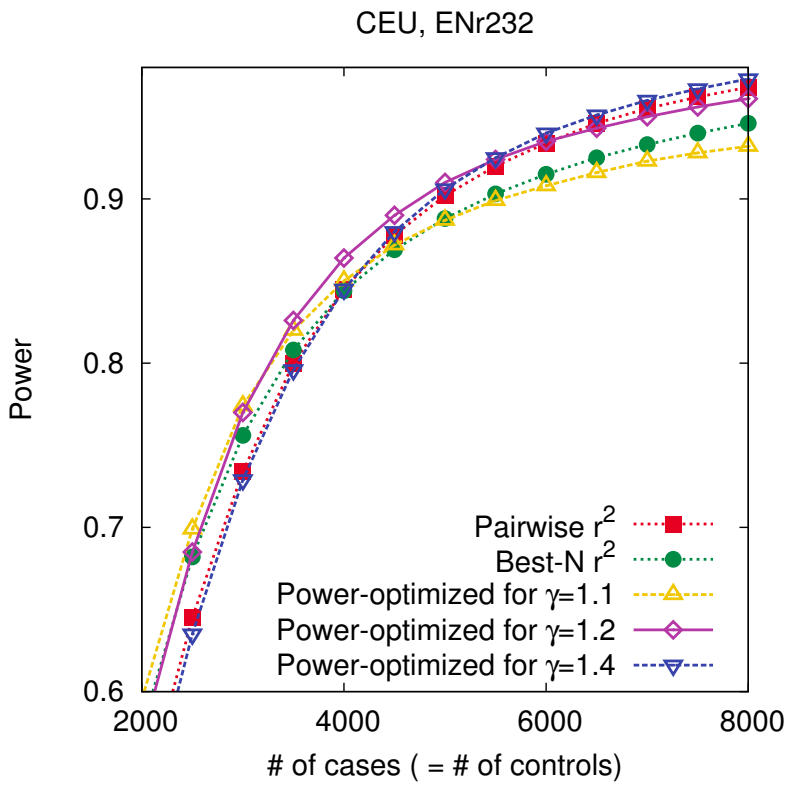
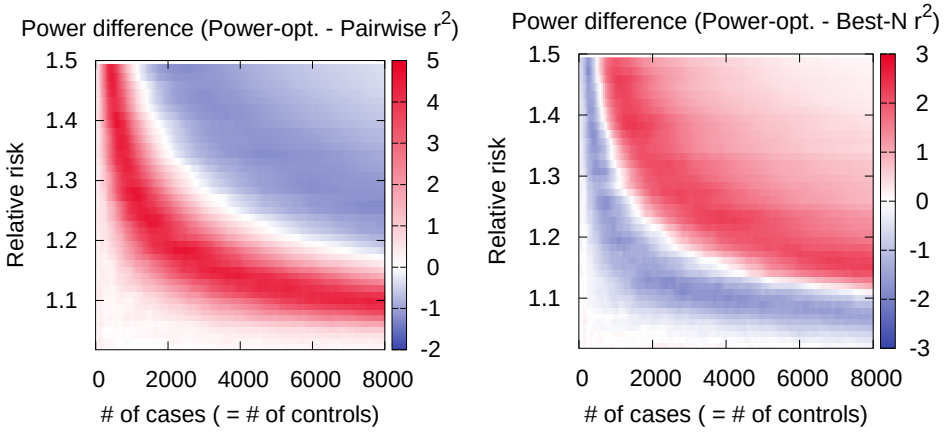


Figure 3.3: Robustness of our power-optimized method to errors of parameter selection. We use our power-optimized method to design three different tag sets of size 100 assuming different relative risks of 1.1, 1.2, 1.4 in the ENr232 ENCODE region of the CEU population. We also design two more tag sets of the same size using pairwise r^2 tagging and best-N r^2. We then measure the power of each tag set based on the assumption of a true relative risk of 1.2. We use a 5% region-wide significance level, a 5% MAF threshold for causal SNPs, and assume a disease prevalence of 0.01 and 4,000 cases and 4,000 controls when designing the tag sets.



(A) (B)

Figure 3.4: Distribution of power gain of our power-optimized method compared to r^2-based methods over a parameter space. We design a tag set of size 100 assuming relative risk of 1.2 and 4,000 cases and 4,000 controls in the ENr232 ENCODE region of the CEU population. We also design tag sets of the same size using pairwise r^2 tagging and best-N r^2. We measure the three tag sets assuming many different parameters, varying the relative risk from 1.0 to 1.5 and the sample size from 0 case and 0 control to 8,000 cases and 8,000 controls. Then we plot the power difference (a) between our power-optimized method and pairwise r^2 tagging and (b) between our power-optimized method and best-N r^2, over the space of these various parameters. We use a 5% region-wide significance level, a 5% MAF threshold for causal SNPs, and assume disease prevalence of 0.01.

and alternative hypothesis. Thus, the results of our experiments on varying the two major factors affecting the effect size (relative risk and sample size) can be straightforwardly generalized to the other parameters as well. Since the performance drop by using incorrect parameters exists, a study-independent method such as r^2-based methods can be an appropriate design choice if the study parameters are completely unknown. But even when only the expected ranges of parameters are known, which we believe to be the case in current association studies, our method can provides robust performance.

3.2.4 Custom follow-up study design

After finding a putative association, a follow-up study verifies the association by replicating the result with independent samples. In many cases in follow-up studies, the samples are already in hand and have already been processed with a commercial product. A practical way to increase power is adding more tag SNPs to a commercial product by designing a custom SNP set.

We simulate a custom follow-up study by adding tag SNPs to the Affymetrix 500k chip in the ten ENCODE regions. For each region, assuming 4,000 cases and 4,000 controls, we incrementally add 5 tag SNPs to the tag set, and construct 100 different tag sets of increasing size. For comparison, we construct another 100 tag sets of similar size using pairwise r^2 tagging and best-N r^2.

Figure 3.5 shows the power increase as we add more SNPs to the Affymetrix 500k chip in the ENr232 region of the YRI population. Adding tag SNPs in this region increases substantial power because the large number of SNPs (1,075) are not relatively well captured by the tag SNPs in the Affymetrix 500k chip (52 tag SNPs). Among the three methods, our method increases the most power. The results of the other populations and regions are similar and shown in Supplementary Figure of Han *et al.* [29]. In the ENr232 region, by adding 1 SNP per 25kb (20 SNPs), our method improves power 6%, 10%, 10%, and 9% in the CEU, CHB, JPT, and YRI populations respectively, while pairwise r^2 tagging improves power 2%, 4%, 4%, and 3% and best-N r^2 tagging improves power 5%, 8%, 5%, and 6% in the same

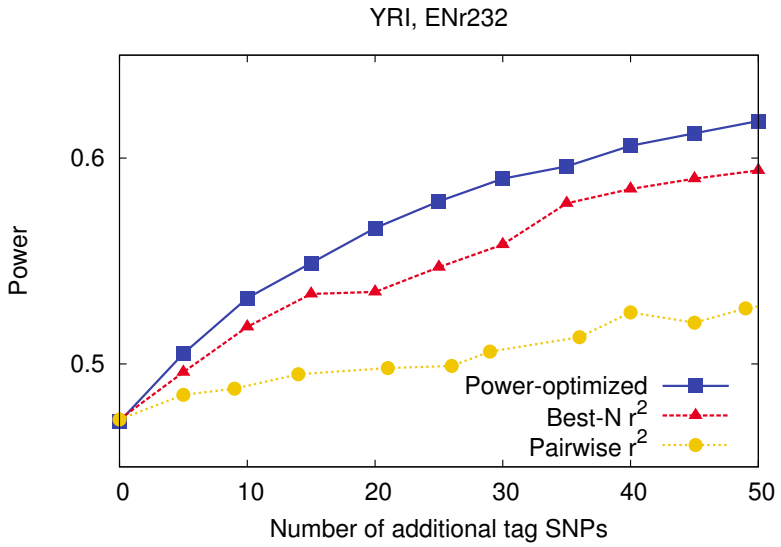


Figure 3.5: Power comparison between our power-optimized method and r^2-based methods with respect to the number of tag SNPs added to a commercial chip. To simulate a custom follow-up study, we use each of our power-optimized method, pairwise r^2 tagging, and best-N r^2, to add tag SNPs to the tag set defined by the Affymetrix 500k chip in the ENr232 ENCODE region of the YRI population. We use a 5% region-wide significance level, a 5% MAF threshold for causal SNPs, and assume relative risk of 1.2, disease prevalence of 0.01, and 4,000 cases and 4,000 controls.

populations.

The power gain by adding more tag SNPs varies between the regions depending on the coverage of the chip. For example, by adding 50 SNPs in the same YRI population, we get a 6% and 7% power increase in the ENm013 and ENm014 regions, while we get a 17% and 15% power increase in the ENm010 and ENr232 regions. Therefore, it is important to examine the coverage of the commercial chip for the region of interest, to see if we will get sufficient power by adding more SNPs. Since our design framework provides efficient empirical simulation for measuring power as well as an efficient tag SNP selection method, we can accurately evaluate power before and after adding tag SNPs, and decide which SNPs to add. Our method can provide optimal performance in custom follow-up study designs because the value of the relative risk can be estimated from the result of the original study.

3.2.5 High-throughput genotyping product design

Since our power-optimized tagging method can scale to the whole genome, we can apply the method to design a whole genome high throughput genotyping product. For each of the HapMap populations, we design 500k, 300k, 100k whole genome tag sets using our power-optimized method assuming 8,000 cases and 8,000 controls. We also design the same size of tag sets using pairwise r^2 tagging and best-N r^2. Figure 3.6 (CEU) and Supplementary Figure of Han *et al.* [29] (all populations) show that our tag sets outperform the r^2-based tag sets.

We also compare our tag sets to commercial products. Figure 3.7 (CEU) and Supplementary Figure of Han *et al.* [29] (all populations) show that our tag sets work better than the commercial products of the same size. Our 100k tag set performs similarly to the Affymetrix 500k chip in the CEU and JPT+CHB populations, but performs worse in the YRI population, because 100k tag set is not large enough to capture the variations in the YRI population. Our 100k tag set also performs similarly to the Illumina 300k chip, except in the CEU population for which the Illumina 300k chip seems to be optimized. Our 300k tag set performs better than or comparable to any commercial product evaluated including the Illumina 550k chip, and our 500k tag set outperforms all products in all populations. For the same

80% genome-wide power level, our 500k tag set requires 26%, 29%, and 33% fewer individuals than the Affymetrix 500k chip and 7%, 11%, and 23% fewer individuals than the Illumina 550k chip in each of the CEU, JPT+CHB, YRI populations.

3.2.6 Efficient power estimation

The analytical approximation for power that we use in design is efficient enough to estimate the whole genome power of a 500k tag set in 3 minutes. This efficiency allows us to design on one ENCODE region in 3 seconds and on the whole genome in 1.5 CPU hours using the adjusted greedy algorithm (See Methods). An underlying reason why we use an analytical approximation instead of a more accurate empirical simulation, other than the computational feasibility, is that we only need a rough estimate of the power to select tag SNPs. The analytical approximation always underestimates power yet in the vast majority of cases preserves the relative ordering of candidate tag sets with respect to their power.

However, an analytical approximation is overly inaccurate for the final estimate of the power of a design, because it applies two assumptions which ignore the correlation structure between SNPs. The Bonferroni assumption ignores the correlation structure between tags by assuming they are independent for multiple-hypothesis correction. The best-tag assumption ignores the correlation structure between a causal SNP and multiple tags by assuming a causal SNP is detected only by its best tags, disregarding the possibility that other tags can also detect the causal SNP. We measure the effects of these assumptions on power. Given a fixed tag set defined as the common SNPs in the Illumina 550k chip, we perform empirical simulations for measuring power with all four combinations of the two assumptions (Bonferroni and best-tag) and compare the results to the analytical approximation. Details of how we incorporate these assumptions into the simulations are described in Methods.

Figure 3.8 (ENr232, CEU) and Supplementary Figure of Han *et al.* [29] (other regions) show that we can underestimate the power by up to 15% using both assump-

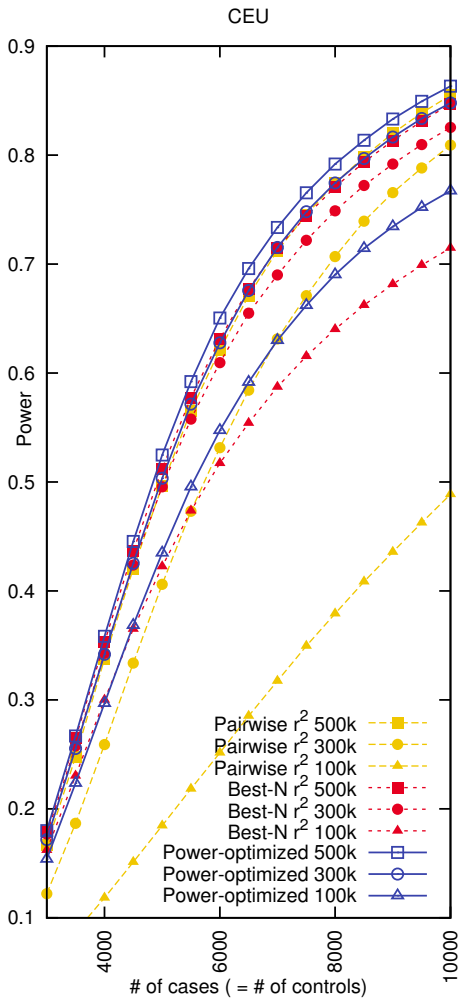


Figure 3.6: Genome-wide power comparison between whole genome tag sets designed by our power-optimized method, pairwise r^2 tagging, and best-N r^2 in the CEU population. We measure the power assuming relative risk of 1.2, disease prevalence of 0.01, a 5% genome-wide significance level, and a 5% MAF threshold for causal SNPs. We use 8,000 cases and 8,000 controls when designing the tag sets.

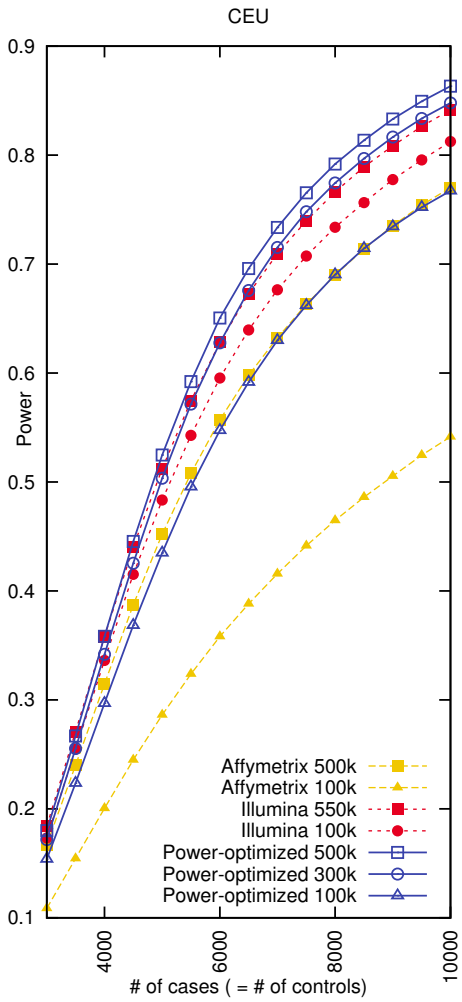


Figure 3.7: Genome-wide power comparison between whole genome tag sets designed by our power-optimized method and the commercial products in the CEU population. We measure the power assuming relative risk of 1.2, disease prevalence of 0.01, a 5% genome-wide significance level, and a 5% MAF threshold for causal SNPs. We use 8,000 cases and 8,000 controls when designing the tag sets.

tions. The effect of the best-tag assumption is shown to be more critical than the effect of the Bonferroni assumption in our results. The difference between the effects of the two assumptions is most significant in the YRI population. The small effect of the Bonferroni assumption implies that the tag SNPs are nearly independent due to the short LD in the YRI population. The significant effect of the best-tag assumption implies that there are many ungenotyped SNPs which are correlated to multiple tag SNPs with moderate r^2. (If a SNP is directly genotyped or highly correlated to a tag SNP, then the effect of the best-tag assumption is small.) From the same reasoning, we can expect that as we collect more and more tag SNPs, the effect of the Bonferroni assumption will increase and the effect of the best-tag assumption will decrease (but not disappear entirely). The empirical simulation with both assumptions (red circles) is almost equivalent to the analytical approximation (black small diamonds) showing that the significant difference in power between the empirical simulation and the analytical approximation is directly due to the assumptions and not the stochastic nature of simulation.

After design, we run empirical simulations for measuring power to avoid the inaccuracy of the analytical approximation. This resampling approach using a reference data set is originally described by de Bakker *et al.* [18]. We improve the efficiency of this procedure and scale it to the whole genome using a random sampling procedure. If we directly apply the standard simulation to the whole genome to measure the power of the Affymetrix 500k chip for 4,000 cases and 4,000 controls in the CEU population, it takes 4,000 CPU hours to construct null and alternate panels. Using our improved simulation procedure, it takes less than 10 CPU hours for the same construction.

3.3 Discussion

We introduced a design framework which provides an efficient tag SNP selection method based on power and a quick empirical simulation procedure that can accurately measure the power of a tag set. The tag SNP selection and the empirical simulation can efficiently scale to the whole genome. Our framework efficiently finds

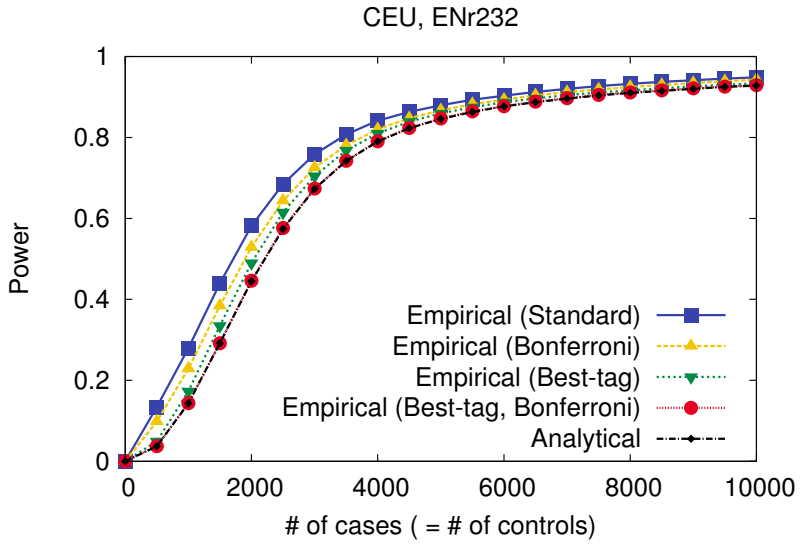


Figure 3.8: Comparison between four different empirical estimates of power and the analytical approximation in the ENr232 ENCODE region of the CEU population. Given a tag set consisting of the common SNPs in the Illumina 550k chip, we perform empirical simulations with all four combinations of the best-tag and Bonferroni assumptions, and compare to the analytical approximation. Details of how we incorporate these assumptions into our simulations are described in the Methods. We use a 5% region-wide significance level, a 5% MAF threshold for causal SNPs, and assume relative risk of 1.2 and disease prevalence of 0.01.

the "key" tag SNPs contributing to power thus providing superior performance to the widely used r^2-based methods in both custom follow-up study design and whole genome tag set design.

We assumed a fixed relative risk of 1.2 for all causal SNPs, since the fixed relative risk assumption is often more realistic than the varying relative risk when we can approximate the relative risk before the study. Our method can maximize power under the varying relative risk assumption as well. We assumed a multiplicative disease model, but the same tag SNP selection technique based on other disease inheritance models can be straightforwardly developed. Furthermore, our method can be optimized over multiple parameters. For example, if we want to design a study optimized for both relative risks of 1.2 and 1.4, we can select tags to maximize the average power over these two disease models. This approach will expand the robustness of our method over a wider range of parameters, at the expense of the peak performance at the single disease model and parameters used in the design. If the study parameters are completely unknown, a study-independent measure such as r^2 can be a suitable choice. However, since many current association studies have at least an expectation of the ranges of parameters, in that case, our method can provide superior performance over the r^2-based methods. A good example is a custom follow-up study, where the relative risk is estimated from the original study.

The results show that our whole genome tag set works significantly better than the commercial products. This comparison is unfair because we designed a tag set for each population while the commercial products are designed for multiple populations. Howie *et al.*[34] propose an r^2-based tag SNP selection method for multiple populations. Our method can also select tag SNPs for multiple populations, by maximizing the sum of the power over multiple populations ($\sum p_i$ where p_i is the power for population i). However, this might bias against populations that have low power such as the YRI population. We can avoid this problem by heuristically adding second-order terms to penalize the bias toward a specific population ($\sum p_i + \sum p_i p_j$). We found that tag sets designed for multiple populations in this way have similar power in each of the populations to a tag set designed for a single population (data not shown).

The computational core of our tagging method is an efficient procedure for

selecting tag SNPs given a fixed number of individuals and a fixed number of tags. Since this core procedure is very efficient we can answer many design questions by repeatedly searching with this core procedure and by using our efficient empirical simulation for accurately measuring power. For example, our method can answer questions such as "How many additional tags do we need to achieve 80% power given a sample size in addition to the Affymetrix 500k chip for a candidate region?", "If we use the individual genotyping for a small region of interest, what is the optimal cost point between the number of individuals and number of tags given a desired power of 80%?" [55] or "How many individuals should we collect for 70% genome-wide power when using the Illumina 550k chip?".

Our experiments for custom follow-up study design are performed in the context of replication analysis of a genomic region of interest without prior knowledge. In addition, our method can leverage the results from previous studies by either explicitly including prioritized tag SNPs or by applying a weighted prior of causal SNPs obtained from previous studies [78, 24]. Our method can also be extended to maximize the power of joint analysis combining the original and the replicated data sets [68]. Furthermore, our method can be easily modified to maximize the minimum power over all causal SNPs instead of the average power.

A recent methodological development in statistical genetics allows us to estimate the probability distribution of ungenotyped SNPs given a tag set and directly compute the test statistic from the distribution. This is called imputation or multi-marker analysis [80, 51]. Since the test statistic is based on estimated information which also has an uncertainty (variance), the multiple hypothesis correction is more subtle. To the best of our knowledge, there is no established tag SNP selection method for this analysis. Our method can be applied to this multi-marker analysis in two different ways. First, we can select tag SNPs to maximize the imputed power at each step. The computational cost of this procedure will be very high. Second, we can design a tag set assuming a single marker analysis, and then apply multi-marker analysis to the resulting tag set. We assume that this latter approach will work reasonably well, since we expect that if a tag set has a good power in a single marker analysis, in most cases it will also have a good power in a multi-marker analysis. We expect that tag SNP selection for imputation analysis will be an active area of

research in the future.

In summary, we present an efficient and accurate power-optimized design framework which also provides flexibility and robustness. The utility of our method ranges from custom follow-up study designs to whole genome high-throughput product design. Our method is publicly available for research purposes via web server at `http://design.cs.ucla.edu`.

3.4 Materials and Methods

3.4.1 Power-optimized tag SNP selection

Our power-optimized tag SNP selection method is a stepwise greedy procedure to maximize power. We assume that we can estimate the relative risk (γ). We determine the MAF threshold and the significance level α. Then, we fix at least one degree among the three degrees of freedom in design which are (1) the number of individuals, (2) number of tags, and (3) desired power. If we fix two of them, our method will give one design. If we fix one of them, our method will iterate and give many designs to choose among. The computational core of this procedure is selecting tags to maximize power given the fixed numbers of individuals and tags. Since this core procedure is very efficient, our method can quickly iterate to find the solution for any choice of fixed parameters. For example, if we fix the desired power, our method will use binary-search over the number of individuals by repeating the core procedure until the resulting tag set meets the desired power. It will iterate this whole process for every number of tags.

For simplicity, we will only consider the core tagging procedure where both the numbers of individuals (N) and the number of tags (n_t) are fixed. Let S be the set of all SNPs in the region. Let $C \subseteq S$ be the set of (common) putative causal SNPs defined by the MAF threshold. Let $I \subseteq S$ and $E \subseteq S$ be the sets of SNPs that we want to include into or exclude from the tag set. Then our tagging procedure is as follows.

1. Initialize the tag set as $T \leftarrow I$

2. For every candidate tag SNP $x \in S - (T \cup E)$, analytically estimate per-causal-SNP power for every causal SNP $c \in C$ using the tag set $T \cup \{x\}$, to get the average power $P(T \cup \{x\})$.

3. Select the best candidate tag SNP x' which maximizes $P(T \cup \{x'\})$.

4. $T \leftarrow T \cup \{x\}$

5. Repeat from step 2 while $|T| < n_t$

A more detailed pseudo-code is shown in Supplementary Figure of Han *et al.* [29]. How we analytically estimate the per-causal-SNP power given a tag set at step 2 will be described below. During the procedure, we measure the average power of a tag set for $O(n_c n_t)$ times where n_c and n_t are the number of causal SNPs and tag SNPs respectively. Since empirically measuring power through simulation for this number of times is computationally impractical, we use an analytical approximation for power.

For genome-wide design, we assume the maximum distance of LD to be 250kb and use the adjusted greedy algorithm (Supplementary Figure of Han *et al.* [29]) to reduce the computational burden. 250kb is not long enough to capture long range LDs, but enough for selecting tag SNPs based on power. We will assume a longer range of maximum LD (10Mb) when we estimate the power of a design using empirical simulation. The adjusted greedy algorithm picks k "independent" SNPs at each round, from the top of the candidate tag SNP list sorted by their power increase. We define two SNPs to be independent if the distance between them is greater than the twice the length of maximum distance of LD. We set k to be 1% of the total number of SNPs. We consider the power between a tag SNP and a causal SNP only if the r^2 between them is ≥ 0.1.

We now describe how to analytically estimate the per-causal-SNP power of a tag set at step 2 of the tagging procedure. We use the framework of Pritchard *et al.* [58], Jorgenson *et al.* [37], Klein *et al.* [42], and Eskin [24]. Given an association study which collects genotypes in $N^+/2$ case and $N^-/2$ control individuals (equivalently N^+ and N^- chromosomes), we assume that a marker A with population minor allele frequency p_A affects the disease with relative risk γ. Let F be the disease prevalence.

The case and control allele frequencies are then

$$p_A^+ = \frac{\gamma p_A}{(\gamma-1)p_A+1} \qquad \text{and} \qquad p_A^- = \frac{p_A - F p_A^+}{1-F} \qquad (\text{or, } p_A^- \approx p_A \text{ if } F \text{ is very small})$$

respectively. We denote the observed case and control frequencies in the collected sample as $\hat{p}_A^+$ and $\hat{p}_A^-$. The association statistic at marker A,

$$S_A = \frac{\hat{p}_A^+ - \hat{p}_A^-}{\sqrt{\hat{p}_A^{\pm}(1-\hat{p}_A^{\pm})}}\sqrt{\frac{N^+N^-}{N^+ + N^-}} \qquad (\text{where } \hat{p}_A^{\pm} = (\hat{p}_A^+ + \hat{p}_A^-)/2)$$

approximately follows a normal distribution with variance 1 and mean (non-centrality parameter)

$$\lambda_A = \frac{p_A^+ - p_A^-}{\sqrt{p_A^{\pm}(1-p_A^{\pm})}}\sqrt{\frac{N^+N^-}{N^+ + N^-}}\ . \qquad (\text{where } p_A^{\pm} = (p_A^+ + p_A^-)/2)$$

If we genotype a marker B correlated with A with a correlation coefficient of r_{AB}, the power that the marker B will be detected as significant is analytically approximated as

$$P_B = 1 - \frac{1}{\sqrt{2\pi}} \int_{\Phi^{-1}(\alpha/2)+\lambda_A\sqrt{r_{AB}^2}}^{\Phi^{-1}(1-\alpha/2)+\lambda_A\sqrt{r_{AB}^2}} e^{-\frac{1}{2}x^2} dx$$

with respect to the significance threshold α, where $\Phi(x)$ is the c.d.f. of the standard normal distribution.

Now we can estimate the single marker power between the causal SNP A and the tag SNP B. To extend this single marker power to multiple markers, we apply two simplifying assumptions. A *best-tag assumption* assumes that each causal SNP is only detected by its best tag, that is, the most correlated tag SNP with the highest r^2. A *Bonferroni assumption* assumes that every SNP is independent allowing us to use α/n_t as a significance level for a single test where α is the region-wide significance level and n_t is the number of tags (the Bonferroni correction). With these two assumptions, the per-causal-SNP power for each causal SNP is efficiently computed as the single marker power at the best tag SNP. Thus, the average power can be analytically estimated by averaging the per-causal-SNP power over every causal SNP.

3.4.2 r^2-based tag SNP selection

Pairwise r^2 tagging [18, 10, 31, 1] is a widely used r^2-based tag SNP selection method that greedily chooses the smallest number of tag SNPs with a minimum r^2

threshold between tag SNPs and uncollected SNPs. The procedure starts with a SNP pool containing every SNP of interest, which is defined by MAF $\geq 5\%$ in our experiments. At each step, the procedure selects a tag SNP which covers the most SNPs in the pool with the r^2 threshold, and removes the tag SNP and the SNPs it covers from the pool. Then the procedure is repeated until the pool becomes empty.

Since the only parameter we can vary in pairwise r^2 tagging is the r^2 threshold, we use binary-search over the threshold when we want to design a specific tag set size, with the precision of 0.001. For the follow-up study design which adds SNPs to a pre-defined tag set, we first remove from the pool the pre-defined tag SNPs and the SNPs they cover, and then resume the normal procedure. For the genome-wide study design, as in our power-optimized method, we assume 250kb as the maximum distance of LD and use the adjusted greedy algorithm. The algorithm picks k "independent" SNPs at each round, from the top of the candidate tag SNP list sorted by the number of SNPs they cover. We define two SNPs to be independent if the distance between them is greater than the twice the length of maximum distance of LD. We set k to be 1% of the total number of SNPs.

Best-N r^2 [18] is another r^2-based method. The procedure is the same as pairwise r^2 tagging except that the tag SNPs are selected until the desired tag set size is obtained, not until the SNP pool becomes empty. Thus, the tag set size can be controlled without varying the r^2 threshold. We use a fixed r^2 threshold of 0.8 in our experiments. For genome-wide design, we use the same assumption of maximum distance of LD and the adjusted greedy algorithm as for pairwise r^2 tagging.

3.4.3 Empirical simulation for power

We empirically measure the final estimate of the power of a tag set after design. Our empirical simulation is based on the standard simulation procedure described by de Bakker *et al.* [18]. This procedure resembles the "bootstrapping" statistical procedure which samples from the data set with replacement to estimate the sampling distribution of an estimator [73, 23]. The major difference is that a typical bootstrapping procedure draws the same number of samples as the data set, while the de Bakker [18] simulation amplifies the number of samples based on a small

reference data set, which is the HapMap. This procedure assumes that although the currently available reference data set is small, the correlation structure between SNPs will be mostly conserved independent of the size of the data set. Since this procedure does not require the conservative assumptions used in the analytical approximation, it is a standard method for measuring power [18, 51].

The procedure consists of creating null panels and alternate panels. Random chromosomes are drawn from the reference data set to create many case/control panels without any causal association (null panels). For each null panel, the best χ^2 statistic among all tag SNPs is obtained. Given a region-wide significance level α, the maximum χ^2 value exceeded in α of null panels is chosen as the threshold to declare a positive result. Next, based on the assumption of a causal SNP which defines the expected allele frequencies in cases and controls, random chromosomes are drawn from the reference data set to create many case/control panels (alternate panel). For each alternate panel, a positive result is recorded if the best χ^2 statistic among all tag SNPs exceeds the χ^2 threshold obtained in the null panels. The power is estimated as the proportion of the positive findings among alternate panels. Previous studies [18, 56, 80] assume a uniform distribution for the causal SNP, and construct an even number of panels per every putative causal SNP.

This standard simulation is not based on the best-tag or Bonferroni assumption, but we can incorporate these assumptions into the simulation for the purpose of comparison (Figure 3.8 and Supplementary Figure of Han *et al.* [29]). The Bonferroni assumption is incorporated by using the Bonferroni correction instead of null panels to assess the per-marker threshold. The best-tag assumption is incorporated by declaring a positive result in an alternate panel only when the causal SNP's most correlated tag shows significance, regardless of other tags.

The computation cost of empirical simulation is a major bottleneck of optimal design of association studies. We improved the efficiency of empirical simulations by taking advantage of having small reference samples. Instead of drawing each simulated case and control from the reference samples, we count how many times each chromosome in reference samples are drawn in cases and controls. Since the number of reference samples are typically much smaller than the number of individuals in the simulation, such an implementation improves the efficiency of simulation studies

by orders of magnitude compared to the straightforward implementation. With r simulation panels, N individuals and t tags, the computational complexity is reduced from O(rNt) to O(r(N+t)) assuming the number of reference panels is a constant smaller than N.

To the best of our knowledge, no one has applied this standard simulation to the whole genome. We introduce two ideas to efficiently scale it to the whole genome. First, we observe that the SNPs very far away from the causal SNP have the same distribution as null panels. Thus, positive results found on those SNPs are likely to be false positives. We use this insight to set a maximum distance L between a casual SNP and a tag SNP, to avoid having to generate alternate panels consisting of an entire chromosome. We conservatively choose L=10Mb not to miss any long range correlation. This idea reduces the computational load of alternate panel construction by more than 5-fold.

Second, we introduce a sampling procedure. The standard strategy of creating an even number (k) of panels per every putative causal SNP (*even-k strategy*) is impractical for the whole genome even when $k = 1$. Instead, we sample the causal SNP from the uniform distribution, and create a number of panels per each sampled causal SNP (*sample-k strategy*). We analyze the variance of the power estimate in this strategy to determine the number of sampling iterations. Using the sampling strategy, the variance of the average power estimate is approximately given as $3p(1-p)/m$ where p is the true average power and m is the number of samplings (See Supplementary Materials of Han *et al.* [29]). Thus, the number of samplings can be estimated given a desired accuracy. The idea of sampling reduces the computational load by more than 20-fold compared to a naive even-k strategy which constructs one panel per causal SNP.

These two ideas for efficient genome-wide simulation increase the efficiency of alternate panel construction but do not help the null panel construction where a causal SNP is not defined. We can reduce the computational load in the null panel construction by adjusting the number of individuals to an appropriate level, based on the fact that the adjusted χ^2 threshold is independent of the number of individuals when the number of individuals is large [14]. This fact allows us to construct null panels once and use each for many different numbers of individuals. We construct

null panels of 1,000 cases and 1,000 controls when we measure the power of designs in our experiments.

3.4.4 Genotype data

We downloaded the HapMap genotype data (build 36) for the whole genome and ENCODE regions from the HapMap project web site [1]. The project collected SNP information from 30 trios in each of the African (YRI) and the European (CEU) populations, and 45 unrelated individuals in each of the Japanese (JPT) and Chinese (CHB) populations. The data includes 2,605,595, 2,471,887, and 2,926,893 polymorphic SNPs in each of the CEU, JPT+CHB, and YRI populations. We phased the data into haplotypes using the HAP software [81].

3.5 Web Resources

The URL for the method presented herein is as follows:
`http://design.cs.ucla.edu`.

Chapter 3 was published in Annals of Human Genetics, 72(Pt 6):834-847, Nov 2008, Buhm Han, Hyun Min Kang, Myeong Seong Seo, Noah Zaitlen, and Eleazar Eskin. “Efficient association study design via power-optimized tag SNP selection”. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Finding New Associations and Detecting Spurious Associations at Post-Association Stage

4.1 Motivation

Genome-wide association studies using high-throughput genotyping technologies are an effective means of discovering associations between genetic variants and diseases [61, 74]. Considerable efforts in these studies are spent to avoid spurious associations in the results. Typically, quality control (QC) approaches are taken to remove potentially problematic markers before the markers are tested for statistical associations. The filters applied in QC step include testing for Hardy-Weinberg equilibrium (HWE) [33, 44], filtering based on the rate of missing genotypes, and filtering based on the ambiguity of genotype clustering plot [74]. Recently, filtering based on the error probability estimated from the hidden Markov model has also been proposed [40, 66].

Even after applying QC filters, systematic bias and errors can still occur causing spurious associations. One of the reasons is that the amount that a marker deviates from the "normal" status such as HWE does not necessarily provide a measure of how much the errors will affect the statistical significance. This is because the statistical significance is a function of the phenotypes (case/control status) as

well as the genotypes. For example, a marker with small number of errors may not significantly deviate from HWE in both cases and controls. However, if one allele is slightly excessive in cases by errors and the other allele is in controls, even small errors turn out to be the cause of spurious association once we take into account case/control status to compute a statistic. For this reason, it is beneficial to perform additional quality refinement based on the association results (e.g. p-values) after taking into account phenotypes, in addition to the pre-association QC.

Recent availability of the reference datasets such as the HapMap [1, 35] provides the linkage disequilibrium (LD) information between markers [3]. With this LD information, the post-association results can provide clues to identify spurious associations. Given a statistically significant association, p-values at nearby markers under LD are expected to be comparably significant and decay with LD (Figure 4.1A). However, if the association shows a singular peak of the statistical significance in the region under LD, it is highly probable that the association is caused by errors (Figure 4.1B). Since examining p-values at nearby markers is so intuitive, it is widely performed in many studies at least implicitly. For example, it is widely accepted for studies to report that k out of m associations were excluded from the report because the nearby markers under LD did not show comparably significant p-values [74].

However, somewhat surprisingly, no statistical framework exist for this procedure to the best of our knowledge. It should be noted that if an investigator manually examines the p-values at nearby markers, the investigator is applying a decision making procedure whether or not a formal statistical framework is used. Without a formal statistical framework, these decisions are based on an individual's intuitions and will be inconsistent from person to person and from study to study. The inconsistencies will be exacerbated if the LD pattern is complicated.

In this paper, we propose a formal statistical framework called post-association RESults QUality Enhancement (RESQUE) which quantitatively measures the evidence of spurious association using the LD information. The idea is to look at the association results at nearby markers and compute the likelihood under the two models, say, true and spurious association models. We use z-scores instead of p-values in order to take into account the positive/negative correlations between statistics

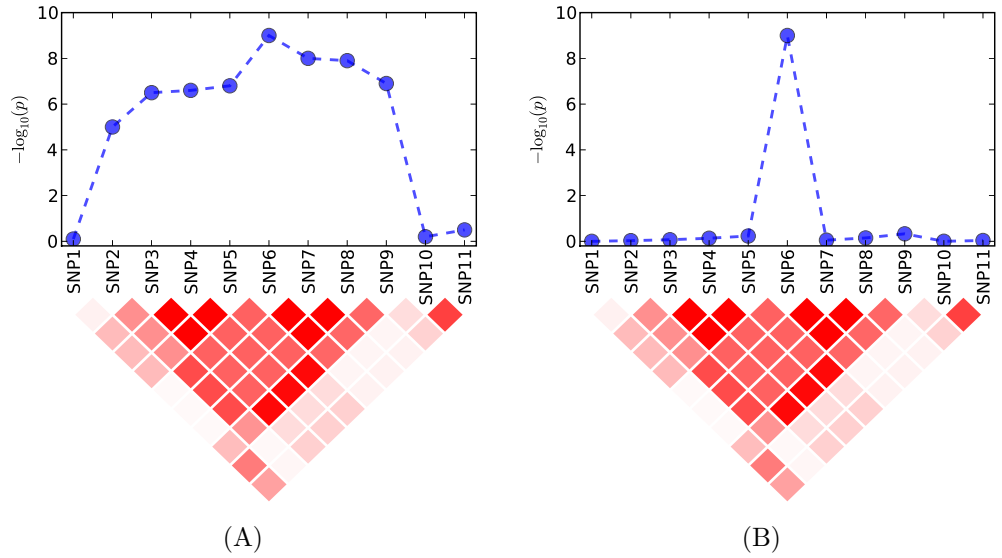


Figure 4.1: Underlying intuition of RESQUE. (A) If an association is a true genetic association, nearby markers in LD will show comparably significant p-values, forming a p-value plot which looks like a mountain. (B) If an association is a spurious association caused by marker-specific errors, nearby markers in LD will not have comparably significant p-values. In LD plot, red color denotes high r^2 between corresponding SNPs and white color denotes low r^2, similarly to the output of Haploview [3].

as well as the strength of the statistical significance. By using z-scores, we can use the well-established results that multiple z-scores asymptotically follow multivariate normal distribution (MVN) of which covariance can be estimated from the genotype correlation structure in the reference dataset [64, 67, 14, 30]. The only complication is that we do not know the effect size which is necessary to compute the likelihood. This is not of major concern at the post-association stage because the effect size can be estimated from the observed association we would like to test.

In addition to detecting spurious associations, RESQUE can also function as a method to rescue candidate associations among the markers which are excluded by QC. Since RESQUE is based on the likelihood ratio, it can detect the significant evidence of the "true" association as well as the spurious association. The underlying intuition is that, assuming that a systematic bias such as population structure does not exist, it is highly unlikely that nearby markers show comparably significant p-values decaying with LD unless they are caused by a genetic effect. This approach can be useful if the QC is too harsh and removes true genetic associations. For example, even if the rate of missing genotypes is high, it may have little effect on the statistical significance if the missing rate is independent of the alleles and phenotypes. Then, removing these markers may take away the chance to detect true genetic associations at these markers. We suggest a new workflow of treating the markers excluded by QC (Figure 4.2). We keep these markers in the dataset and test for statistical associations. We apply RESQUE to significant associations, and if a strong evidence of true association is found, the association is considered as a supplemental candidate association.

It should be noted that the associations rescued in this way are more likely to be false positives than the associations at "normal" markers. To minimize false positives, we use a conservative approach. First, we use a stringent likelihood ratio threshold of 1,000, which is in spirit similar to the LOD score of 3 in linkage studies. We analytically show that if there are no combined effects of genetic effect and errors, using this large threshold approximately controls the false positive rate of the study. Second, since combined effects can occur, we correct for the effect of the errors by using a similar procedure to genomic control [20]. If an association is rescued in spite

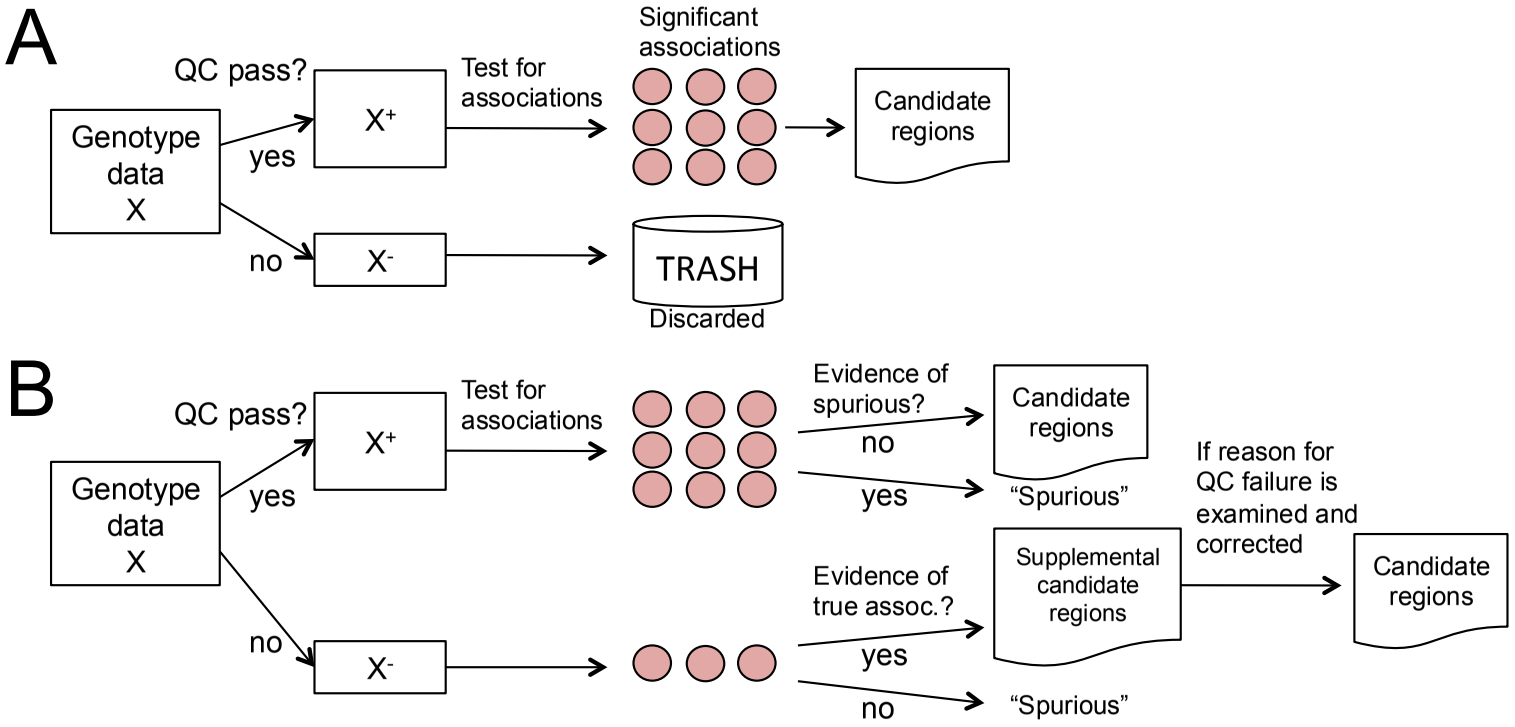


Figure 4.2: A schematic of the workflow of studies applying RESQUE. (A) In the current approach, the markers which fail for QC are completely discarded. (B) In our proposed approach, we rescue the associations at the markers which fail for QC as supplemental candidate regions, if there is a strong evidence of true associations. Also, the associations at the markers which pass QC are tested for the evidence of spurious associations.

of this conservative approach, it can be an indication of an interesting candidate region which may have genetic effect. We believe that similar to associations at "normal" markers rescued associations should also be followed up for verification such as through replication studies.

We apply RESQUE to the Wellcome Trust Case Control Consortium data (WTCCC) [74]. Among the reported associations by WTCCC, RESQUE finds strong evidence of spurious associations at 6 SNPs for type 1 diabetes. Since these are all in the MHC region where a large number of associated SNPs exist, whether they are spurious or not does not make a difference in the analysis. However, it will be useful if the statistical significance are subsequently used for fine-mapping causal variants in this large region. Among the unreported associations excluded by QC, RESQUE finds strong evidence of true genetic effects at 101 SNPs: 1 for coronary artery disease, 3 for CrohnâĂŹs disease, 31 for rheumatoid arthritis, and 65 for type 1 diabetes. One of the associations for type 1 diabetes is located at PGM1 gene (1p31.3) where only a moderate significance ($p < 10^{-4}$) was reported by WTCCC. Interestingly, this gene is one of the 18 regions recently identified and replicated by a meta-analysis performed 2 years after the WTCCC analysis [4]. This shows that RESQUE can help studies identify novel candidate regions and increase the study power. Other associations overlap with the reported regions by WTCCC, suggesting that the rescued associations are likely to indicate true genetic effects.

Our method is very different from the quality control methods based on the hidden Markov model [40, 66]. These methods can be thought of as an extension of imputation methods [51, 80, 38, 65]. The advantage of these methods is that the error probability of each individual genotype can be estimated and subsequently used for error corrections [66]. On the other hand, our method can be thought of as a formalization of what is typically performed at the post-association stage using intuition. Our method is different from the pre-association QC methods including the imputation-based methods, because our method uses the post-association results obtained after taking into account case/control status in order to capture spurious associations not captured at the pre-association stage. Moreover, our method can indicate the evidence of "true" associations, thereby providing a procedure treating the QC-excluded markers which may increase the study power (Figure 4.2). We

expect that both the pre-association QC methods including imputation-based methods and our post-association method can complement each other to provide a higher standard of the quality of association results in future studies.

4.2 Materials and Methods

4.2.1 Definitions

We call the associations with true genetic effect *true genetic associations*. There are two kinds of false positives. We call the false positives caused purely by the stochastic nature of the sampling procedure *stochastic false positives*. This type of false positives are usually unavoidable and controlled to a desired level in terms of false positive rate such as family-wise error rate (FWER) [30, 69]. We call the false positives caused by other unexpected bias and errors *spurious associations*. If spurious associations occur, the false positive rate is no longer controlled. Our goal is to detect spurious associations from the remainder, *candidate true associations*, which will be the union of true genetic associations and stochastic false positives. Since the discrimination between true genetic associations and stochastic false positives can be done by replication and functional studies but is not the focus of our paper, we will often refer to the candidate true associations as *true associations* in short.

4.2.2 RESQUE

RESQUE is a likelihood ratio test. Given a statistically significant association, the likelihood of the z-scores at nearby markers are computed for both the "true" and "spurious" association models. RESQUE utilizes the known results that z-scores at multiple markers asymptotically follows a multivariate normal distribution (MVN) [64, 67, 14, 30]. In this paper, we use the allelic χ^2 test under the multiplicative disease model, but extensions to other tests and models are possible in the MVN framework as shown in Han *et al.* [30].

Multivariate normal approximations

Consider we collect $N^+/2$ cases and $N^-/2$ controls. (N^+ case haplotypes and N^- control haplotypes). Let p_i be the population MAF at SNP i estimated from the reference dataset. Let p_i^+ and p_i^- be the case and control population MAFs. For the causal SNP c, given the disease prevalence F and relative risk γ,

$$p_c^+ = \frac{\gamma p_c}{(\gamma-1)p_c+1} \quad \text{and} \quad p_c^- = \frac{p_c - F p_c^+}{1-F} \ .$$

If we directly collect the causal SNP c, the test statistic at c

$$S_c = \sqrt{N}\frac{\hat{p}_c^+ - \hat{p}_c^-}{\sqrt{2\hat{p}_c^s(1-\hat{p}_c^s)}}$$

asymptotically follows $S_c \sim \mathcal{N}(\lambda_c\sqrt{N}, 1)$, where

$$\lambda_c\sqrt{N} = \frac{p_c^+ - p_c^-}{\sqrt{2p_c^s(1-p_c^s)}}\sqrt{N}$$

is the non-centrality parameter (NCP), $N = \frac{2N^+N^-}{N^++N^-}$ is the so-called effective sample size, $p_c^s = \frac{N^+p_c^+ + N^-p_c^-}{N^++N^-}$ is the expected sample MAF, where hats ($\hat{\cdot}$) denote observed values [29]. S_c^2 equals to the popular χ^2 test statistic.

Let r_{ij} be the correlation coefficient between SNPs i and j estimated from the reference dataset. It is known that if we collect a non-causal SNP i that has the correlation coefficient r_{ic} to the causal SNP c, the NCP of the statistic S_i is approximately $r_{ic}\lambda_c\sqrt{N}$ [58]. Also, given two SNPs i and j, if γ is small, the covariance between S_i and S_j is approximately r_{ij} [67, 14, 30]. Therefore, by the multivariate central limit theorem [73], the vector of statistics $(S_1, S_2, ..., S_n)$ at n nearby markers asymptotically follows approximately $\mathcal{N}_n(\Lambda\sqrt{N}, \Sigma)$ where $\mathcal{N}_n$ denotes a n-dimensional MVN,

$$\Lambda = \lambda_c(r_{1c}, r_{2c}, ..., r_{nc}) \ ,$$

and Σ is the $n \times n$ matrix of correlation coefficients between n markers. If there is no true genetic association (i.e. $\gamma = 1$), the MVN is centered at zero (i.e. $\Lambda = 0$).

It should be noted that the use of Σ is a valid approximation only if γ is small. The exact covariance matrix given any γ can also be computed using the procedure in Han *et al.* [30]. However, if the association shows genome-wide significance, $\hat{S}_s$

will be large and NCP (Λ) will be far from zero. Thus, NCP will be the major factor affecting the likelihood and a small difference in covariance will not make a difference in results.

Likelihood ratio

Using the MVN framework, we make a decision of true and spurious association based on the likelihood ratio. Suppose that we observe a significant association at SNP s. Let S_s be the statistic at s. Let $1, ..., n$ be the index of n nearby SNPs and $S_1, ..., S_n$ be the corresponding statistics. We assume that errors only affect a single SNP in a region. Therefore, even if the association at s is caused by errors, $S_1, ..., S_n$ are not affected. We compare likelihoods of observed statistics $(\hat{S}_1, ..., \hat{S}_n)$ under two models.

- **True association model (M_T)**: The association is a candidate true association. If the association is a true genetic association, there exists a causal SNP and $(S_1, ..., S_n)$ will follow $\mathcal{N}_n(\Lambda\sqrt{N}, \Sigma)$. We assume that s is causal and then estimate the effect size based on the observation at s. This is algebraically equivalent to using
$$\widehat{\Lambda\sqrt{N}} = \hat{S}_s(r_{1s}, r_{2s}, ..., r_{ns}) \ .$$
If the association is a stochastic false positive, $(S_1, ..., S_n)$ will follow a conditional distribution given $\hat{S}_s$. The conditional mean turns out to be exactly $\widehat{\Lambda\sqrt{N}}$. Therefore, for both cases, $(S_1, ..., S_n)$ approximately follows $\mathcal{N}_n(\widehat{\Lambda\sqrt{N}}, \Sigma)$.

- **Spurious association model (M_S)**: The association is a spurious association caused by errors. $(S_1, ..., S_n)$ follows $\mathcal{N}_n(0, \Sigma)$.

Under true association model, it is possible to use the conditional variance given the observation $\hat{S}_s$, which is $\Sigma - \mathbf{x}\mathbf{x}^T$ where $\mathbf{x} = (r_{1s}, r_{2s}, ..., r_{ns})$. We find that using the conditional variance is more powerful but often less stable because a small error in $\hat{S}_s$ considerably affects the results. We use the marginal variance Σ.

Let f_T and f_S be the probability densities for two models. Given the observation $(\hat{S}_1, ..., \hat{S}_n)$, the likelihood ratio (LR) statistic is
$$LR_{\text{RESQUE}} = \frac{f_T(\hat{S}_1, ..., \hat{S}_n)}{f_S(\hat{S}_1, ..., \hat{S}_n)} \ .$$

We do not include $\hat{S}_s$ in the likelihood computation to avoid the complication of determining a probability model for errors under spurious association model.

LR has the following interpretations. A large LR (e.g. 1,000) is the evidence that the association is a true association. A small LR (e.g. 1/1,000) is the evidence that the association is spurious. LR close to 1 shows that it is ambiguous which model is true.

An interesting property of our method is that if we add a uncorrelated SNP to the nearby SNP set, the likelihood ratio does not change. Given the nearby SNPs $1, ..., n$, suppose we add a SNP j that is uncorrelated to $1, ..., n$. Since they are uncorrelated, the new $n+1$-dimensional density function f_T^* or f_S^* can be decomposed into $f_T^*(\hat{S}_1, ..., \hat{S}_n, \hat{S}_j) = f_T(\hat{S}_1, ..., \hat{S}_n) f_j(\hat{S}_j)$ where f_j is the marginal distribution of S_j. Thus, the new likelihood ratio is

$$LR^*_{\text{RESQUE}} = \frac{f_T^*(\hat{S}_1, ..., \hat{S}_n, \hat{S}_j)}{f_S^*(\hat{S}_1, ..., \hat{S}_n, \hat{S}_j)} = \frac{f_T(\hat{S}_1, ..., \hat{S}_n) f_j(\hat{S}_j)}{f_S(\hat{S}_1, ..., \hat{S}_n) f_j(\hat{S}_j)} = LR_{\text{RESQUE}} \ .$$

This property allows us to generously choose nearby SNPs without worrying much about the adverse effect of including uninformative SNPs.

With similar derivation, we can show that if a marker is in LD with no other SNPs, all density functions cancel out and LR turns out to be exactly 1.

Decision making

Once the likelihood ratio is computed, we make a decision of true and spurious associations. The simple and intuitive approach we adapt is the decision making directly based on the likelihood ratio, which categorizes the results into three categories: true, spurious, and ambiguous. Given a threshold C (e.g. 1,000), if $LR \geq C$, we predict a true association. If $LR \leq 1/C$, we predict a spurious association. If the evidence is moderate ($1/C < LR < C$), we make an ambiguous prediction. This approach resembles Bayesian model selection based on a Bayes factor.

A less preferable approach is the decision making based on a p-value. If we want to detect true associations, we put M_S as the null hypothesis and M_T as the alternative. We measure the p-value toward M_T as

$$p_T = \int_{LR_{\text{RESQUE}}(x_1,...,x_n) \geq LR_{\text{RESQUE}}(\hat{S}_1,...,\hat{S}_n)} f_S \, dx_1 ... dx_n \ . \tag{4.1}$$

If the p-value is smaller than a threshold, we reject the null hypothesis of M_T. If we want to detect spurious associations, we put M_T as the null hypothesis and M_S as the alternative. The p-value toward M_S is

$$p_S = \int_{LR_{\mathrm{RESQUE}}(x_1,\ldots,x_n) \leq LR_{\mathrm{RESQUE}}(\hat{S}_1,\ldots,\hat{S}_n)} f_T \, dx_1 \ldots dx_n \ .$$

If the p-value is smaller than a threshold, we reject the null hypothesis of M_S. This approach is less preferable because the rejection of null hypothesis is not necessarily the evidence of alternative model. For example, if the means of null and alternative densities are far apart, which is often the case of our situation, the p-value can be extremely small even though the observation is closer to the null mean than to the alternative mean. Another reason is simplicity, because dealing with two types of p-values can be confusing.

In this paper, we use the decision making based on the likelihood ratio. However, the decision making based on a p-value has an advantage that the theoretical analysis of false positive rate is straightforward as shown below. Thus, for a more strict control of false positives, one can use a combined approach that predicts a true association only when $LR \geq C$ and p_T is small. Nevertheless, our simulations show that if we use a large C such as 1,000, p_T is almost always very small. Therefore, the combined approach will be nearly equivalent to the decision making based only on the likelihood ratio.

False positive rate control

We describe some analyses of false positive rate, how our method can help to control it and what conditions have to be met for the control. We assume the control of FWER, but similar analyses for different error measures [69] are possible. We first make a simplifying assumption that a statistically significant association is either solely caused by errors or unaffected by errors. Thus, we ignore the case that both the genetic effect and the errors result in a combined effect.

Let M be the set of m markers which pass QC and K be the set of k markers which do not pass QC. Applying only the QC is equivalent to assuming that associations in M and K are true and spurious respectively. RESQUE can be thought of as refining the results of QC if the strong evidence of the opposing decision is

seen. Thus, the refinement is a two-part procedure; (1) detects spurious associations among M and (2) detects true associations among K. In part (1), let q_{SPU} be the power to detect spurious associations and e_{SPU} be the probability to falsely predict "spurious" for actual true associations. In part (2), let q_{TRU} be the power to detect true associations and e_{TRU} be the probability to falsely predict "true" for actual spurious associations. Using a large LR threshold C in the decision making procedure will ensure small e_{SPU} and e_{TRU}, while q_{SPU} and q_{TRU} may not be maximized.

We first consider the effect of part (1) on FWER considering only the markers in M. Since part (1) only removes associations, FWER will always decrease. Given the desired FWER level α, assume that the Bonferroni correction is used to use the significance threshold of α/m. Although stochastic false positives will be controlled by using this threshold, there can be spurious associations. Let β be the probability that errors will induce spurious associations among M. The probability of one or more spurious associations is $1-(1-\beta)^m \approx m\beta$. Therefore, if we only apply QC, FWER will not be controlled at α but at $\alpha_{\mathrm{QC}} = 1-(1-\alpha)(1-m\beta) \approx \alpha + m\beta$. If we apply RESQUE to remove spurious associations, FWER is now controlled at $\alpha(1-e_{\mathrm{SPU}}) + m\beta(1-q_{\mathrm{SPU}})$, which is strictly smaller than α_{QC} and close to α if q_{SPU} is large. Let q_{STA} be the statistical power of the study. By applying our method, the power drops to $q_{\mathrm{STA}}(1-e_{\mathrm{SPU}})$, but the drop will be small if e_{SPU} is kept small by using a large LR threshold.

Now we consider the joint effect of part (1) and (2) on FWER considering all markers in M and K. Let β' be the probability of spurious associations among K. In many cases, $\beta' > \beta$. If one uses the Bonferroni correction to use the significance threshold of $\alpha/(m+k)$, FWER will be controlled at $\frac{m\alpha}{m+k}+m\beta$ among M and $\frac{k\alpha}{m+k}+k\beta'$ among K. Thus, the overall FWER will be approximately $\alpha + m\beta + k\beta'$, which can be very large if β' is large. If we apply RESQUE, FWER is now controlled at

$$\alpha_{\mathrm{RESQUE}} = \frac{m\alpha}{m+k}(1-e_{\mathrm{SPU}}) + \frac{k\alpha}{m+k}q_{\mathrm{TRU}} + m\beta(1-q_{\mathrm{SPU}}) + k\beta' e_{\mathrm{TRU}} \ .$$

Since $\alpha_{\mathrm{RESQUE}} \leq \alpha_{\mathrm{QC}} + k\beta' e_{\mathrm{TRU}}$, as long as we keep e_{TRU} very small, we can control FWER approximately at least at α_{QC}, the level that we would obtain with only applying QC, and possibly even lower because spurious associations will be removed.

Now we provide a practical example. Suppose $n = 480,000$ and $m = 20,000$.

We usually do not know β', but the number of significant associations among K, say l, can be an estimate of $k\beta'$. This will usually be a conservative over-estimate because they will contain true genetic associations as well. If we want to keep $k\beta' e_{\mathrm{TRU}}$ as small as $.05\alpha$, we need to control e_{TRU} at $.05\alpha/l$, which is 10^{-5} if $\alpha = .05$ and $l = 250$. This can be achieved by using the aforementioned decision making procedure based on both the LR and p-value, so that true associations can be predicted only when $p_T < 10^{-5}$.

However, stochastically estimating p_T for each association can be a time consuming procedure. Since p_T is often extremely small, the standard sampling procedure requires a very large number of samples. We can overcome this problem by using importance sampling [73]. Since we make decisions based on both LR and p-value, we need to estimate p_T only if $LR \geq C$. Since this large LR already shows that the observation is close to the mean of f_T, we have a suitable sampling distribution for importance sampling, f_T. Let $\mathbf{x}_1$, ..., $\mathbf{x}_B$ be the samples from f_T. Then, the p-value formula (4.1) can be estimated by

$$\sum_{i=1}^{B} \frac{I\left(LR_{\mathrm{RESQUE}}(\mathbf{x}_i) \geq LR_{\mathrm{RESQUE}}(\hat{S}_1, ..., \hat{S}_n)\right)\left(f_S(\mathbf{x}_i)/f_T(\mathbf{x}_i)\right)}{B},$$

where $I(\cdot)$ is an indicator function. This estimate will have a much smaller variance compared to directly sampling from f_S.

If there are t true genetic associations among M and t' among K, if we only apply the standard QC, the expected number of true positives is tq_{STA}. Now if we apply our method, it is $tq_{\mathrm{STA}}(1-\epsilon)(1-e_{\mathrm{SPU}}) + t'q_{\mathrm{STA}}q_{\mathrm{TRU}}$, where ϵ is the power loss due to the increased multiple testing burden, induced by using the significance threshold of $\alpha/(m+k)$ instead of α/m. If $t' > 0$ and e_{SPU} is kept small, this quantity can be larger than tq_{STA}. Thus, the study power can increase.

Now we consider the case that the aforementioned simplifying assumption does not hold. If the major cause of an association is genetic effect, RESQUE will rescue the association among QC-excluded markers, but there can also be the effect of errors. To alleviate this problem, we apply a similar idea to genomic control [20]. We estimate the standard deviation of the statistics among the QC-passed associations, σ, and the standard deviation of the statistics among the QC-excluded associations, σ'. Usually, $\sigma' \geq \sigma \approx 1$. Then we can adjust the statistics at the QC-

excluded associations by a factor of σ/σ' to compensate the increased variance. We call the resulting association p-values after this adjustment *penalized p-values*. If a large number of associations are predicted as true among QC-excluded associations, σ' can be estimated from the associations that are only predicted true instead. In WTCCC data, we find that $\sigma \approx 1.04$ and $\sigma' \approx 1.17$ resulting in $\sigma'/\sigma \approx 1.13$. If we only consider the associations that are predicted true, $\sigma' \approx 1.16$.

4.2.3 Genotype data

We downloaded the HapMap genotype data (release 23a, NCBI build 36) from the HapMap project web site [35, 1] and phased the data into haplotypes using HAP [81], which can handle the trio information. We downloaded the genotype data from the Wellcome Trust Case Control Consortium web site [74]. We downloaded the control data and the case data for bipolar disorder (BD), coronary artery disease (CAD), CrohnâĂŹs disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D).

4.3 Results

4.3.1 Simulations in HapMap ENCODE regions

We test the performance of RESQUE using the HapMap ENCODE regions of the CEU population consisting of 120 parental haplotypes. The ten ENCODE regions contain 10,710 SNPs. We assume that the 1,464 SNPs in the Affymetrix Human SNP Array 6.0 are genotyped, which we will call tag SNPs.

First we generate 1,000 studies with true associations. Given one ENCODE region, we randomly select a causal SNP among all SNPs and pick a relative risk from Uniform(1.1,1.4). We assume that the minor allele is causal and the disease prevalence is .01. Then we sample 2,000 cases and 2,000 controls (total 8,000 haplotypes) based on the relative risk. In practice, there exists a difference in MAF and LD between the collected sample and the reference dataset. Since our simulated study is directly generated from the reference dataset, in order to simulate more realistically, for every SNP of each haplotype, we mutate (change the value of the

allele) with probability .01. If the most significant p-value among all tag SNPs is between 10^{-6} and 10^{-10} (roughly the genome-wide significance level), we accept this study. Otherwise, the dataset is not used. We repeat this to construct 100 studies per each of 10 regions.

Second, we generate 1,000 studies with spurious associations. Given a region, we randomly select an error SNP among all tag SNPs and pick a case mutation rate μ_1 and a control mutation rate μ_0 each from Uniform(0,.5). We randomly sample 2,000 cases and 2,000 controls and change the allele at the error SNP with probability μ_1 in the cases and μ_0 in the controls, to induce a systematic bias. For all other SNPs, we apply the mutation probability of .01. We start over until the p-value at the error SNP is between 10^{-6} and 10^{-10}. We repeat this to construct 100 studies per each of 10 regions.

Then we apply RESQUE to all 2,000 studies to examine if RESQUE can correctly discriminate 1,000 studies with true associations and 1,000 studies with spurious associations. For each association-showing SNP, we use the nearby SNPs with $r^2 > .3$. If there is no nearby SNP with $r^2 > .3$, we use the most correlated SNP. We threshold the LR with various levels and the results are shown in Table 4.1.

Table 4.1 shows the predictions made by RESQUE depending on the threshold levels. Using a low threshold means that we make less ambiguous predictions with the cost of less accurate predictions. Using a high threshold means that we make more accurate predictions with the cost of more ambiguous predictions. For all threshold levels tested, RESQUE makes correct predictions on >95% of the true associations and >84% of the spurious associations. Surprisingly, at any threshold, RESQUE makes at most only 1 incorrect prediction on true and spurious associations respectively, showing a very high discriminative power.

The source of this high power turns out to be the genome-wide significance threshold. To show this, we similarly construct another set of 2,000 studies with true and spurious associations so that the most significant p-value in each study can be between 10^{-2} and 10^{-6}. Then we construct another set so that the p-value can be 10^{-10} and 10^{-14}. Thus, we have a study set of $10^{-2} < p < 10^{-6}$, a set

Table 4.1: RESQUE's predictions in simulations using the HapMap ENCODE regions. We construct 1,000 true and 1,000 spurious associations of genome-wide significance level ($10^{-10} < p < 10^{-6}$). Predictions are made based on the observed likelihood ratio (LR) at each association. LR is the likelihood under the true model divided by the likelihood under the spurious model.

	1,000 actual true associations			1,000 actual spurious associations		
LR threshold C	predicted true (LR$\geq C$)	predicted ambiguous ($1/C <$LR$< C$)	predicted spurious (LR$\leq 1/C$)	predicted true (LR$\geq C$)	predicted ambiguous ($1/C <$LR$< C$)	predicted spurious (LR$\leq 1/C$)
10	991	8	1	1	30	969
100	978	21	1	0	77	923
1000	958	42	0	0	156	844

of $10^{-6} < p < 10^{-10}$, and a set of $10^{-10} < p < 10^{-14}$. We plot the histogram of $\log_{10}$(LR) for each of three sets. Figure 4.3 shows that the LR distributions are more drastically separated between true and spurious associations for study sets with more significant p-values. The more the LR distributions are separated, our method obtains higher power. Therefore, the high power of our method can be thought of as the beneficial side-effect of the efforts to achieve genome-wide significance level.

In Methods, we have shown that our method has a theoretical property that the performance is not greatly affected by the choice of nearby SNPs. Instead of choosing nearby SNPs with $r^2 > .3$, we choose SNPs with $r^2 > .5$ and repeat the analysis. Table 4.2 shows that the results are similar to the results in Table 4.1, with power differing only by a maximum of 2% between two experiments.

One drawback of our approach is that we do not have any discriminative power for a marker that does not have any markers in LD. However, the proportion of these markers is small because even if a marker is covered with a low r^2 by nearby markers, our method has a reasonable amount of power. To show this, we take the study set of genome-wide significance level ($10^{-6} < p < 10^{-10}$). For each study, we look at the maximum r^2 that the signal-showing SNP is covered by nearby SNPs. We split the studies according to the range of maximum r^2, and plot the LR distributions for each r^2 range. Figure 4.4 shows that at the low r^2 range of $0.3 < r^2 \leq 0.5$, the LR distributions are already separated by a considerable amount. Thus, our method is likely to have a reasonable power if the marker is covered with r^2 as low as between 0.3 and 0.5.

When we estimate the likelihood under the true association model, we assume that the signal-showing SNP is causal and estimate the effect size based on the observation of the signal-showing SNP. However, both the assumption and estimation can be incorrect. Since we know the actual causal SNP and the effect size for these simulations, we examine whether the LR distribution changes if we use the actual causal SNP and the effect size. Results show that, if we use the true values, the LR

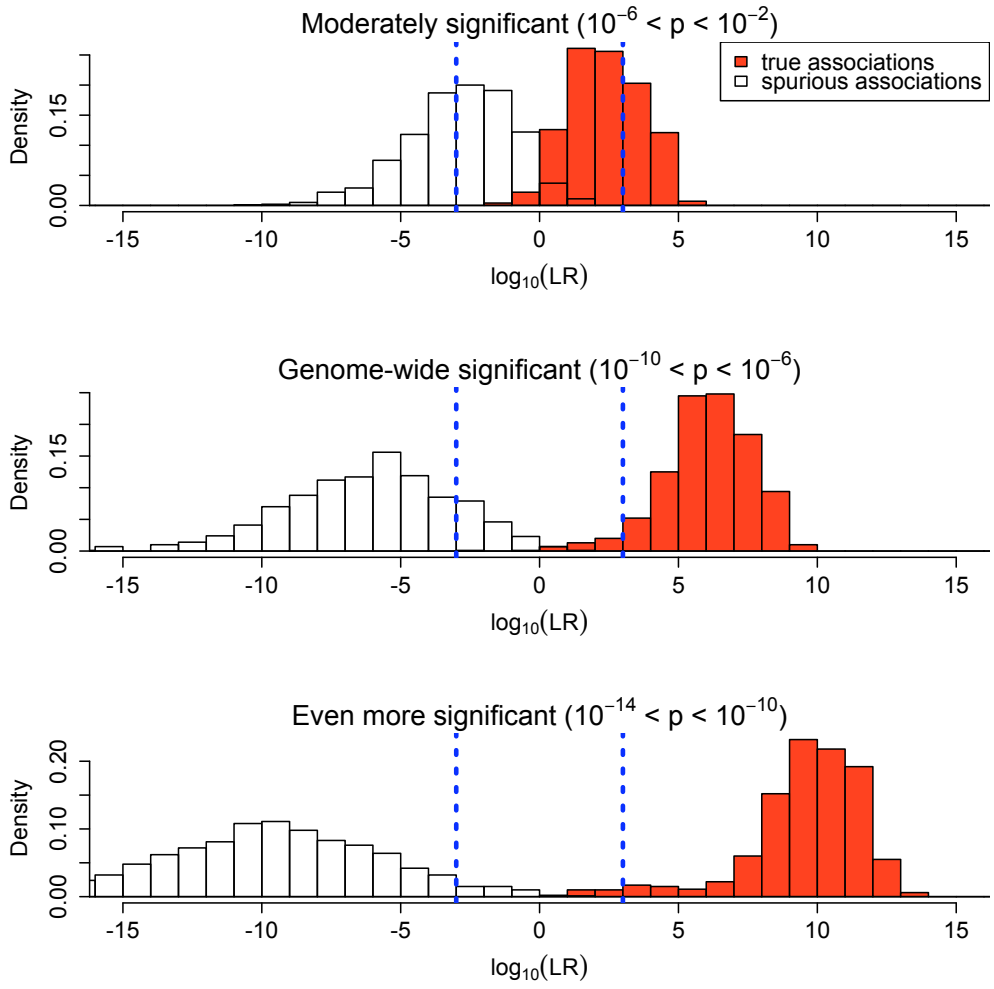


Figure 4.3: Relation between RESQUE's discriminative power and the p-value level. As the p-value becomes more significant, the LR distributions between true and spurious associations become highly separated, giving our method high discriminative power. LR is the likelihood under the true model divided by the likelihood under the spurious model. The dashed vertical lines denote the decision boundaries if the LR threshold $C = 1,000$ is used. RESQUE predicts true associations for observations in the right region, predicts spurious associations for observations in the left region, and makes ambiguous predictions for observations in the middle region.

Table 4.2: QRC's prediction in simulations using the HapMap ENCODE region, when we use the nearby markers with $r^2 > .5$. The results are very similar to the results in Table 1, showing that QRC is robust against the choice of nearby markers.

	1,000 actual true associations			1,000 actual spurious associations		
LR threshold C	predicted true LR$\geq C$	predicted ambiguous $1/C <$LR$< C$	predicted spurious LR$\leq 1/C$	predicted true LR$\geq C$	predicted ambiguous $1/C <$LR$< C$	predicted spurious LR$\leq 1/C$
10	991	8	1	1	31	968
100	976	23	1	0	86	914
1000	941	59	0	0	171	829

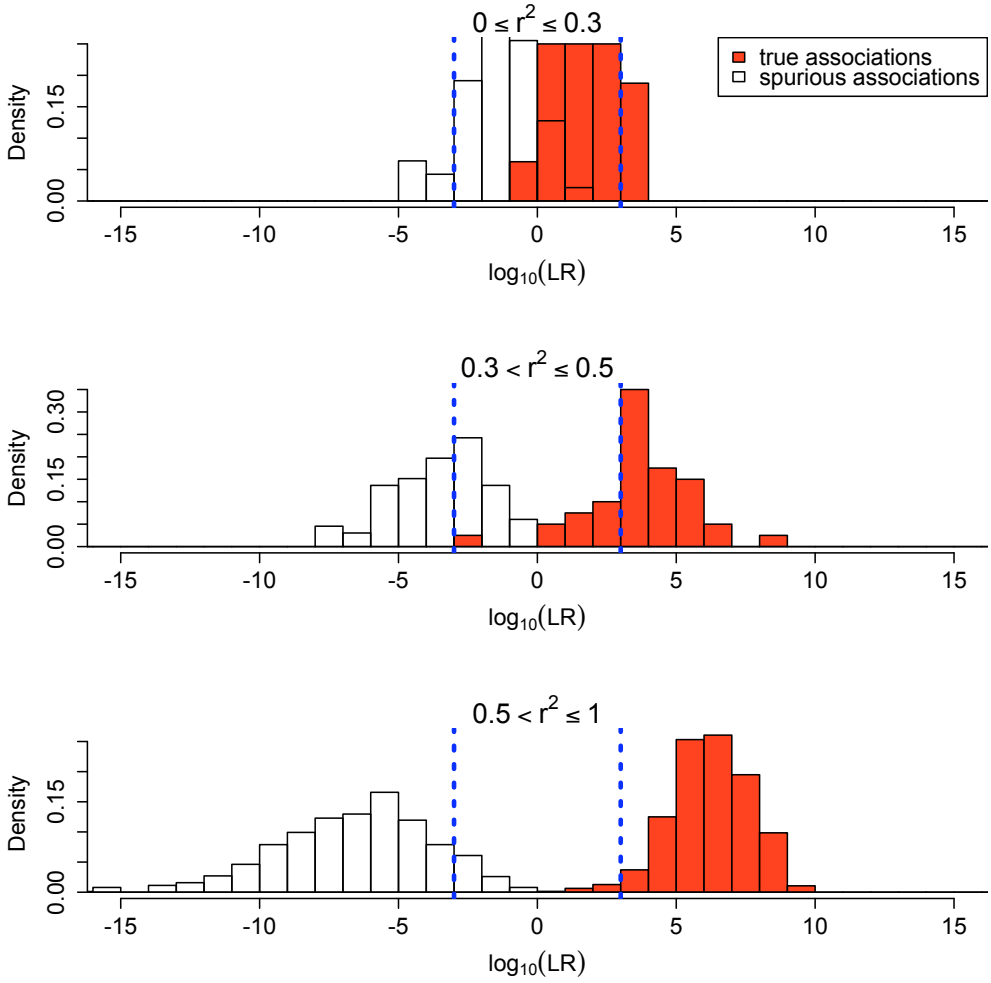


Figure 4.4: LR distributions of true and spurious associations, separated into r^2 bins. As we expect, the higher the maximum r^2 to nearby markers is, the two LR distributions are more highly separated giving QRC high power. At a r^2 level as low as 0.3-0.5, two distributions are already separated by a considerable amount.

distribution of the true associations moves slightly to the right (data not shown). That is, our method obtains slightly more power. However, the amount of power increase is very small, showing that our assumption and estimation are not a critical factor affecting the performance. Note that the LR distribution of the spurious associations is not affected by these assumptions, because there is no true causal SNP or effect size.

In our simulations, the discriminative power of our method is high enough that we can use a low LR threshold of $C = 10$ without making many incorrect predictions. However, in practice, there can be other causes that can degrade the accuracy of predictions. Therefore, we suggest a conservative threshold of 1,000 so that the predictions are made only if the evidence is strong.

4.3.2 Application to WTCCC data

We apply RESQUE to the WTCCC data using the HapMap CEU data as a reference dataset. Among 500,568 SNPs in the WTCCC data, we remove 63,550 SNPs that are non-polymorphic in the HapMap CEU data, 8,013 SNPs that did not have rs-ids, and 6,626 SNPs that were not in the version of the HapMap data we used. We use the remaining 422,379 SNPs in the following analysis. Although it will be possible to re-do the analysis with the current version of the HapMap, we assume that it will make little difference because our method is only applied to the significant associations of which the vast majority are in the remaining SNPs. We exclude all problematic individuals specified by WTCCC.

The QC filters that WTCCC applied are as follows. They first remove SNPs that have (1) a high rate of missing genotypes, (2) deviation from HWE, or (3) a significant difference between the two control groups. QC (1)-(3) exclude 20,794 SNPs among the SNPs we analyze. They apply the association tests to the remaining SNPs. Given significant associations, they additionally exclude SNPs that have (4) a bad genotype clustering plot or (5) a bad p-value plot. Examining the p-value plot is the same idea as RESQUE; nearby SNPs should show comparable p-values. QC (4) and (5) are manually performed in their analysis.

WTCCC reports tens of associated regions over 7 diseases based on the SNPs

that both pass QC and show genome-wide significance ($p < 5.7 \times 10^{-7}$) in either the trend test or the 2-degrees of freedom genotypic test. We will not consider their advanced analysis such as the combined cases or multi-locus analysis. A total of 740 associations are shown to be significant in these regions.

In order to examine if there are any spurious associations among the 740 reported associations, we apply RESQUE to these associations. we look at 50 neighboring SNPs in each direction and choose nearby SNPs with $r^2 > .3$. This is roughly equivalent to looking at 300kb to each direction. We apply a conservative LR threshold of $C = 1,000$ so that we make a prediction only when the evidence is strong. We make an ambiguous prediction if the evidence is not strong or fewer than 3 nearby SNPs are used in LR computation.

The results show that RESQUE finds strong evidence of true associations in 660 (89%) out of 740 reported associations (left column of Table 4.3). Thus, based on the LD information, RESQUE is confirming that the majority of reported associations are true associations. RESQUE makes ambiguous predictions for 134 associations and detects possibly spurious associations for 6 associations. Al these 6 associations are for type 1 diabetes and in MHC region where a huge number of associations exist 4.4. Thus, whether they are spurious does not make a difference in the analysis. However, this at least shows that RESQUE can strongly confirm the true associations and there can be spurious associations not captured by the standard QC.

Next, we examine if there are any associations RESQUE rescues among the markers excluded by QC. We put back the 20,794 SNPs excluded by QC (1)-(3) in the dataset and perform association test. We apply the χ^2 test with the same genome-wide threshold ($p < 5.7 \times 10^{-7}$) to obtain additional associations. We also put back the associations that passed QC (1)-(3) but did not pass QC (4)-(5). As a result, we obtain 1,822 new associations in the autosomes.

We apply RESQUE to these 1,822 new unreported associations using the same protocol as above. The results show that RESQUE finds strong evidence of spurious

Table 4.3: RESQUE's predictions for the WTCCC associations. Predictions are made based on the threshold of $C = 1,000$. LR is the likelihood under the true model divided by the likelihood under the spurious model.

Reported (QC passed) associations			Unreported (QC excluded) associations		
predicted true ($\text{LR} \geq C$)	predicted ambiguous ($1/C < \text{LR} < C$)	predicted spurious ($\text{LR} \leq 1/C$)	predicted true ($\text{LR} \geq C$)	predicted ambiguous ($1/C < \text{LR} < C$)	predicted spurious ($\text{LR} \leq 1/C$)
660	74	6	101	976	745

Table 4.4: We apply QRC to WTCCC associations and find these 6 associations as putative spurious associations. These passed the WTCCC QC, but the likelihood under the spurious model is orders of magnitude larger than the likelihood under the true model. These are all in MHC region where a huge number of significant associations exist, so even if these are really spurious, it does not make a difference in WTCCC analysis. Positions are in NCBI build-35 coordinates.

Disease	Chr.	Position	SNP	χ^2 p-value	$\log_{10}$(LR)	#nearby(max r^2)
T1D	6	30871270	rs3131060	8.1E-21	-5.3	16(1.0)
T1D	6	31133692	rs2517524	7.0E-10	-5.8	12(0.5)
T1D	6	31206936	rs13200022	2.6E-17	-7.0	3(0.7)
T1D	6	32766288	rs9469220	2.3E-101	-15.4	19(0.6)
T1D	6	32890127	rs17501267	4.1E-15	-7.4	4(0.6)
T1D	6	33208405	rs2395351	1.8E-12	-7.4	5(0.4)

associations in 745 (41%) out of 1,822 unreported associations (right column of Table 4.3). Thus, based on the LD information, RESQUE is confirming that about a half of new associations are spurious associations with high confidence. RESQUE makes ambiguous predictions for 976 associations. RESQUE finds that 101 associations show strong evidence of true associations based on the LD information and remains genome-wide significant after correcting for the increased variance with the factor of 1.12, which is the estimated standard deviation increase of the statistic.

Among these 101 possibly rescued true associations, 1 SNP is for CAD, 3 are for CD, 31 are for RA, and 66 are for T1D (Table 4.5). Except for 1 association at 1p31.3 for type 1 diabetes, all these new candidate associations overlap with the reported regions by WTCCC. Thus, effectively, RESQUE rescues only 1 association for all 7 diseases. This association is located at PGM1 gene where only a moderate ($p < 10^{-4}$) significance was reported by WTCCC 4.6. Interestingly, this gene is one of the 18 regions recently identified and replicated by a meta-analysis performed 2 years after the WTCCC analysis [4]. This shows that RESQUE may help studies to identify supplemental candidate regions.

We plot the LR distributions for all of 740 reported and 1,822 unreported associations in Figure 4.5. Two highly separated bimodal distributions show that there is plenty of information in the LD that we can take advantage of to discriminate true and spurious associations. Although two bimodal distributions do not look symmetric, it should be noted that it is not a systematic bias which moves the two distributions to the positive direction. The high peak around 0 (LR of 1) is expected because it contains markers which do not have nearby markers in LD. The reason we did not see this peak in simulations using the ENCODE regions can be that the marker set we use (1M chip) covers the majority of markers in the ENCODE regions in LD.

In this analysis, we made decisions based only on the LR. However, as described in Methods, it is possible to make decisions based on both the LR and the

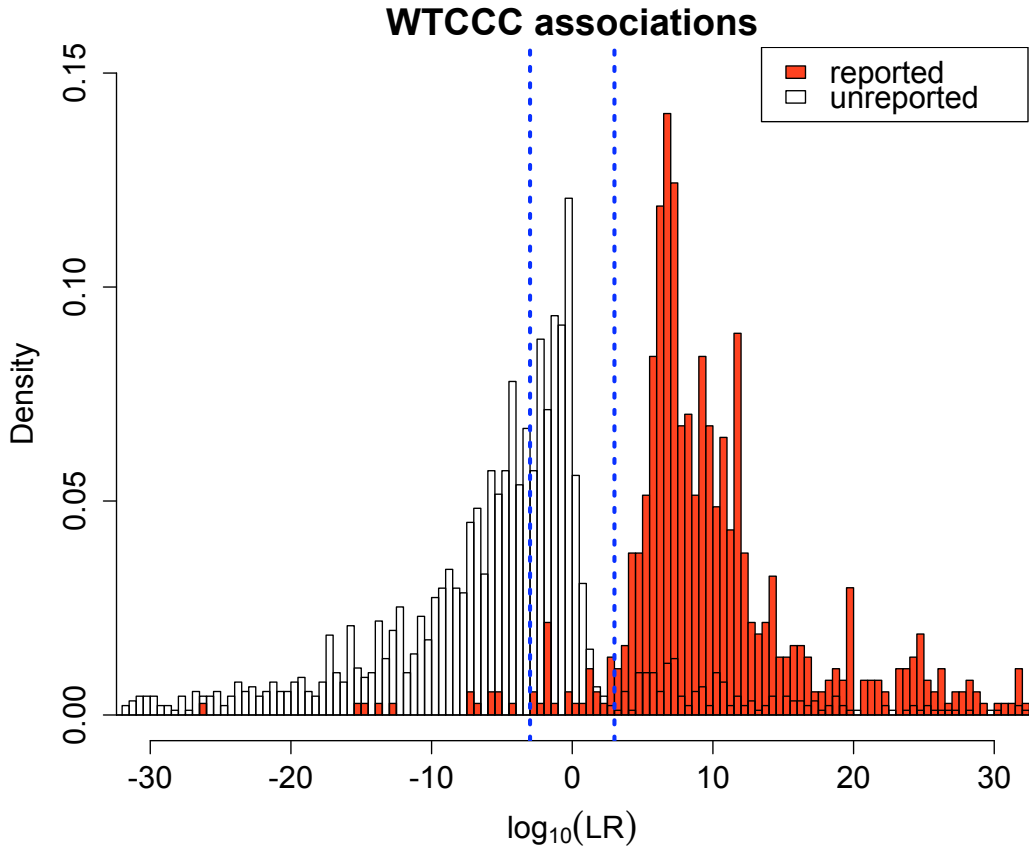


Figure 4.5: Likelihood ratio distribution of WTCCC associations. Reported associations denote the associations reported by WTCCC as genome-wide significant. Unreported associations denote the associations that achieve genome-wide significance but did not pass QC. The dashed vertical lines denote the decision boundaries of RESQUE given the LR threshold $C = 1,000$. LR is the likelihood under the true model divided by the likelihood under the spurious model. RESQUE predicts true associations for observations in the right region, predicts spurious associations for observations in the left region, and makes ambiguous predictions for observations in the middle region. The peak around 0 includes markers which do not have nearby markers in LD. Ambiguous predictions are made for these markers because they have no LD information.

p-value p_T, so that the false positive rate can be approximately controlled. If we want to control FWER to .05 per each disease, this can be approximately achieved by predicting true associations among QC-excluded markers only when $LR \geq C$ and $p_T < 10^{-5}$. After applying this decision procedure, the results are almost unchanged. The only difference from the results in Table 4.3 is that 1 out of 660 associations that were predicted as true among reported associations is now predicted as ambiguous. Thus, applying the decision procedure based on both the LR and p-value is nearly equivalent to the decision procedure based only on the LR in this analysis.

4.4 Discussion

We propose an approach which quantitatively measures the evidence of spurious associations at post-association stage using the LD information. Our approach can be thought of as a formalization of what is typically done in studies by intuition. Since our approach can indicate the evidence of "true" associations as well, the study power can increase by rescuing the associations at the QC-excluded markers.

To the best of our knowledge, our method is the first formal approach for quality refinement at the post-association stage based on the post-association results. We expect that more ideas can be developed for the post-association quality refinement area, so that they can interplay with existing pre-association stage QC to provide a strong protection against spurious associations. Our method is also the first attempt to find meaningful results from the markers which are typically discarded.

Since we use z-scores and the MVN framework in likelihood computation, obtaining the correct covariance matrix is important. It will be necessary to select appropriate reference population which closely resembles the LD structure of the target population. Also, if there are allele labeling inconsistencies or strand errors between the sample and reference dataset (e.g. A/T are erroneously considered as T/A), it will adversely affect the performance because maintaining the correct sign of correlation is important to compute accurate likelihood. These errors can also be problematic in imputation methods and meta-analysis [17]. There are known methods based on MAF or LD for detecting strand errors between two datasets [59], which can be used for our method to check strand errors between the sample and

reference dataset.

The application of our approach is limited to the case where systematic bias does not exist such as the population structure. Therefore, it is important to examine the evidence of such bias before applying our procedure. A similar method modeling the systematic bias or an extension to the next-generation sequencing will be an interesting area for further investigation.

4.5 Web Resources

The URL for methods presented herein is as follows:
`http://resque.cs.ucla.edu`

Chapter 4 is currently in submission for publication of the material. Buhm Han, Brian M. Hackel, and Eleazar Eskin, "Finding new associations and detecting spurious associations at post-association stage". The dissertation author is the primary investigator and author of this paper.

Table 4.5: 101 associations rescued among markers excluded by QC in WTCCC data. All 101 associations show strong evidence of true associations based on the LD information ($LR > 1,000$) and genome-wide significant ($p < 5.7 \times 10^{-7}$) after correcting for the increased variance. Unreported regions are the regions where no SNPs were reported as genome-wide significant by WTCCC.

	Overlap with reported regions	In unreported regions
BD	0	0
CAD	1	0
CD	3	0
HT	0	0
RA	31	0
T1D	64	1
T2D	0	0

Table 4.6: 1 putative true association found by RESQUE. This associations did not pass WTCCC QC and the region is reported as only moderately significant ($p < 10^{-4}$) by WTCCC. Penalized p-value is the p-value after applying our correction procedure with a standard deviation scaling factor of 1.12. LR is the likelihood under the true model divided by the likelihood under the spurious model. Positions are in NCBI build-35 coordinates.

Disease	Chromosome	Position	SNP	Allelic χ^2 p-value	Penalized p-value	$\log_{10}$(LR)	Risk allele	Minor allele	Control MAF	Case MAF	#nearby SNPs (Best r^2)	Note
T1D	1p31.3	63,823,887	rs3790857	3.4E-9	1.3E-7	3.6	A	A	0.190	0.241	11(1.00)	PGM1 [4]

Chapter 5

Conclusion and remarks

5.1 Summary and Conclusion

In this dissertation, I propose that major challenges in genetic association studies can be solved by effectively utilizing the LD information. For the multiple hypothesis testing correction problem, I take advantage of the characteristic of LD patterns that LD is localized by using an efficient sliding-window approach. For the power estimation problem, unlike the previous approaches which assumes the markers are independent, I take into account the LD patterns to estimate accurate power. For the tag SNP selection problem, instead of using only the correlations, I extract power information from LD by accounting for allele frequencies. By using the power itself as a SNP selection criterion, my method outperforms others. For the spurious association detection problem, I formalize the idea that the markers in LD should show similar p-values if the association is not caused by errors. I also extend this framework to "rescue" associations at the markers which are typically excluded by QC.

I expect that these method will help the future association studies to design and analyze better, so that more true genetic variations responsible for the diseases can be identified without incurring excessive false positives. The identified true genetic associations will improve our knowledge about the disease and hopefully and ultimately contribute to human health.

5.2 Future works

There are still many statistical and computational challenges for genetic association studies. An interesting challenge is the statistical framework which maximizes the statistical power for sequenced data. Since sequenced data include many rare variants, the current approach applying tests to each individual variant has poor statistical power. The underlying idea is that we can collapse the information within a meaningful genomic region such as a gene, so that the association between the region and the phenotypes can be detected. There are previously proposed frameworks such as the collapsing method [45] or weighted groupwise association test [50]. However, it is not clear whether there exists an optimal method, and for which situation and disease model each proposed method is suitable. For example, if the relative risk for each variant is known, the statistic based on the exact likelihood ratio will be different from the previous methods.

Genetic association studies based on model organisms such as the mouse are also a promising area because experiments can be done on these organisms which are not possible for humans, such as making genetic crosses or controlling the environment. Since the phenomenon of population structures is severe in model organisms, methods for controlling population structure are typically applied to assess the correct statistical significance of associations in the presence of population structure [39]. However, it is not clear if these methods are applied, how we should estimate the power and how much the multiple testing burden will be. It will be interesting to develop a power estimation procedure and a multiple hypothesis testing correction procedure for model organisms.

Finally, systems genetics is also an interesting area integrating multiple layers of data, such as the genotypes, intermediate phenotypes, and disease phenotypes. Since the gene expression data are now extensively gathered by the microarray technology, it will be interesting how these data integrate and provide useful information about the causality, such as which genes' expression levels affect the disease and are affected by the disease [11]. If we can infer the causal relationships between intermediate phenotypes and disease phenotypes, it will help us understand the underlying pathways how the disease phenotypes are developed.

Bibliography

[1] D Altshuler, LD Brooks, A Chakravarti, FS Collins, MJ Daly, P Donnelly, and International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–320, 2005.

[2] David J Balding. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, 7(10):781–791, 2006.

[3] J C Barrett, B Fry, J Maller, and M J Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005 Jan 15.

[4] JC Barrett, DG Clayton, P Concannon, B Akolkar, JD Cooper, HA Erlich, C Julier, G Morahan, J Nerup, C Nierras, V Plagnol, F Pociot, H Schuilenburg, DJ Smyth, H Stevens, JA Todd, NM Walker, and SS Rich. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet*, 2009 May 10.

[5] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300, 1995.

[6] Tushar R Bhangale, Mark J Rieder, and Deborah A Nickerson. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet*, 40(7):841–843, 2008 Jul.

[7] Brian L Browning. Presto: rapid calculation of order statistic distributions and multiple-testing adjusted p-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics*, 9:309, 2008.

[8] Sharon R Browning and Brian L Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, 81(5):1084–1097, 2007.

[9] M C Byng, J C Whittaker, A P Cuthbert, C G Mathew, and C M Lewis. SNP subset selection for genetic association studies. *Ann. Hum. Genet.*, 67(Pt 6):543–556, 2003.

[10] C. Carlson, M. Eberle, M. Rieder, Q. Yi, L. Kruglyak, and D. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, 74:106–20, 2004.

[11] Lin S Chen, Frank Emmert-Streib, and John D Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biol*, 8(10):R219, 2007.

[12] J. M. Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87:52–8, 2001.

[13] F. S. Collins, L. D. Brooks, and A. Chakravarti. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, 8:1229–31, 1998.

[14] K. N. Conneely and M. Boehnke. So many correlated tests, so little time! rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet*, 81, 2007.

[15] 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, 2003.

[16] Emmanuelle Cousin, Jean-Francois Deleuze, and Emmanuelle 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 Genet.*, 7:20, 2006.

[17] Paul I W de Bakker, Manuel A R Ferreira, Xiaoming Jia, Benjamin M Neale, Soumya Raychaudhuri, and Benjamin F Voight. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*, 17(R2):R122–8, 2008 Oct 15.

[18] PI de Bakker, R Yelensky, I Pe'er, SB Gabriel, MJ Daly, and D Altshuler. Efficiency and power in genetic association studies. *Nat Genet.*, 37(11):1217–23, 2005.

[19] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics.*, 29:311–22, 1995.

[20] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, Dec 1999.

[21] F Dudbridge and B Koeleman. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies. *American journal of human genetics*, 75:424–435, Sep 2004. 10.1086/423738.

[22] Frank Dudbridge and Arief Gusnanto. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*, 32(3):227–234, 2008.

[23] Bradley Efron. Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7(1):1–26, 1979.

[24] Eleazar Eskin. Increasing power in association studies by using linkage disequilibrium structure and molecular function as prior information. *Genome Res.*, 18(4):653–660, 2008.

[25] A. Genz. Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1:141–150, 1992.

[26] A. Genz. Comparison of methods for the computation of multivariate normal probabilities. *Computing Science and Statistics*, 25:400–405, 1993.

[27] V. Hajivassiliou, D. McFadden, and P. Rudd. Simulation of multivariate normal rectangle probabilities and their derivatives: theorerical and computational results. *Journal of Econometrics*, 72:85–134, 1996.

[28] Eran Halperin, Gad Kimmel, and Ron Shamir. Tag SNP selection in genotype data for maximizing snp prediction accuracy. *Bioinformatics*, 21 Suppl 1:i195–203, 2005.

[29] B. Han, H. M. Kang, M. S. Seo, N. Zaitlen, and E. Eskin. Efficient association study design via power-optimized tag SNP selection. *Ann Hum Genet*, 72(Pt 6):834–847, Nov 2008.

[30] Buhm Han, Hyun Min Kang, and Eleazar Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*, 5(4):e1000456–, 04 2009.

[31] HapMap. The International HapMap Project. *Nature*, 426(6968):789–796, 2003.

[32] David A Hinds, Laura L Stuve, Geoffrey B Nilsen, Eran Halperin, Eleazar Eskin, Dennis G Ballinger, Kelly A Frazer, and David R Cox. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307(5712):1072–1079, 2005.

[33] Louise Hosking, Sheena Lumsden, Karen Lewis, Astrid Yeo, Linda McCarthy, Aruna Bansal, John Riley, Ian Purvis, and Chun-Fang Xu. Detection of genotyping errors by hardy-weinberg equilibrium testing. *Eur J Hum Genet*, 12(5):395–399, 2004 May.

[34] Bryan N Howie, Christopher S Carlson, Mark J Rieder, and Deborah A Nickerson. Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Hum. Genet.*, 120(1):58–68, 2006.

[35] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–862, 2007.

[36] G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nat Genet.*, 29:233–7, 2001.

[37] Eric Jorgenson and John S Witte. Coverage and power in genomewide association studies. *Am. J. Hum. Genet.*, 78(5):884–888, 2006.

[38] Hyun Min Kang, Noah A Zaitlen, Buhm Han, and Eleazar Eskin. An adaptive and memory efficient algorithm for genotype imputation. *RECOMB*, 2009.

[39] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, Mar 2008.

[40] Justin Kennedy, Ion I. Mandoiu, and Bogdan Pasaniuc. Genotype error detection using hidden markov models of haplotype diversity. *Journal of Computational Biology*, 15(9):1155–1171, 2008.

[41] G. Kimmel and R. Shamir. A fast method for computing high-significance disease association in large population-based studies. *Am. J. Hum. Genet.*, 79:481–92, 2006.

[42] Robert J Klein. Power analysis for genome-wide association studies. *BMC Genet.*, 8:58, 2007.

[43] Leonid Kruglyak. Power tools for human genetics. *Nat. Genet.*, 37(12):1299–1300, 2005.

[44] Suzanne M Leal. Detection of genotyping errors and pseudo-snps via deviations from hardy-weinberg equilibrium. *Genet Epidemiol*, 29(3):204–214, 2005 Nov.

[45] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 83(3):311–321, Sep 2008.

[46] J. Li and L. Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95:221–7, 2005.

[47] D. Y. Lin. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21:781–7, 2005.

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

[49] Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233, 1982.

[50] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, Feb 2009.

[51] J. Marchini, B. Howie, S. Myers, G. McVean, and P Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39:906–913, 2007.

[52] V Moskvina and KM Schmidt. On multiple-testing correction in genome-wide association studies. *Genet Epidemiol*, 32(8):567–573, 2008.

[53] Dan L Nicolae. Testing untyped alleles (TUNA)-applications to genome-wide association studies. *Genet Epidemiol*, 30(8):718–727, Dec 2006.

[54] Dale R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am. J. Hum. Genet.*, 74:765–9, 2004.

[55] 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(Pt 6):733–746, 2005.

[56] Itsik Pe'er, Paul I W de Bakker, Julian Maller, Roman Yelensky, David Altshuler, and Mark J Daly. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.*, 38(6):663–667, 2006.

[57] Itsik Pe'er, Roman Yelensky, David Altshuler, and Mark J Daly. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol*, 32(4):381–385, 2008.

[58] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, 69:1–4, 2001.

[59] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, Mark J Daly, and Pak C Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575, 2007.

[60] Zhaohui S Qin, Shyam Gopalakrishnan, and Goncalo R Abecasis. An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. *Bioinformatics*, 22(2):220–225, 2006.

[61] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science.*, 273:1516–7, 1996.

[62] Scott F Saccone, John P Rice, and Nancy L Saccone. Power-based, phase-informed selection of single nucleotide polymorphisms for disease association screens. *Genet. Epidemiol.*, 30(6):459–470, 2006.

[63] P. D. Sasieni. From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–1261, Dec 1997.

[64] Daniel J. Schaid, Charles M. Rowland, David E. Tines, Robert M. Jacobson, and Gregory A. Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, 70:425–434, 2002.

[65] Paul Scheet and Matthew Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78(4):629–644, 2006 Apr.

[66] Paul Scheet and Matthew Stephens. Linkage disequilibrium-based quality control for large-scale genetic studies. *PLoS Genet*, 4(8):e1000147, 2008.

[67] S. R. Seaman and B. MÃŒller-Myhsok. Rapid simulation of P values for product methods and multiple-testing adjustment in association studies. *Am. J. Hum. Genet.*, 76:399–408, 2005.

[68] Andrew D Skol, Laura J Scott, Goncalo R Abecasis, and Michael Boehnke. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.*, 38(2):209–213, 2006.

[69] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, 100(16):9440–9445, 2003.

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

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

[72] T van Agthoven, AM Sieuwerts, J Veldscholte, ME Meijer-van Gelder, M Smid, A Brinkman, AT den Dekker, IM Leroy, WF van Ijcken, S Sleijfer, JA Foekens, and LC Dorssers. Cited2 and ncor2 in anti-oestrogen resistance and progression of breast cancer. *Br J Cancer*, Nov 2009.

[73] Larry A. Wasserman. *All of statistics: a concise course in statistical inference.* Springer, 2003.

[74] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[75] P. H. Westfall and S. S. Young. *Resampling-based multiple testing.* Wiley, New York, 1993.

[76] D. A. Williams. Tests for differences between several small proportions. *Applied Statistics*, 37(3):421–434, 1988.

[77] F. Yates. Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.

[78] Kai Yu, Nilanjan Chatterjee, William Wheeler, Qizhai Li, Sophia Wang, Nathaniel Rothman, and Sholom Wacholder. Flexible design for following up positive findings. *Am. J. Hum. Genet.*, 81(3):540–551, 2007.

[79] G. Udny Yule. On the application of the χ^2 method to association and contingency tables, with experimental illustrations. *Journal of the Royal Statistical Society*, 85(1):95–104, 1922.

[80] N. Zaitlen, H. Kang, E. Eskin, and E. Halperin. Leveraging the HapMap correlation structure in association studies. *Am. J. Hum. Genet.*, 80:683–91, 2007.

[81] Noah A Zaitlen, Hyun Min Kang, Michael L Feolo, Stephen T Sherry, Eran Halperin, and Eleazar Eskin. Inference and analysis of haplotypes from combined genotyping studies deposited in dbSNP. *Genome Res*, 15(11):1594–1600, 2005.