Statistical methods for genome-wide association studies of gene expression, with applications to the genetic study of psoriasis

by

Jun Ding

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2010

Doctoral Committee:

Professor Gonçalo R. Abecasis, Chair
Professor Michael L. Boehnke
Professor James T. Elder
Associate Professor Kerby A. Shedden
Assistant Professor Sebastian K. Zöllner

To my parents and my wife

Acknowledgements

It is a great privilege to receive my Ph.D. training in statistical genetics from Michigan. There are so many people I want to thank for their advice, help and support during my years at Michigan.

I want to thank my advisor Dr. Gonçalo Abecasis. He has led me to the research area of statistical genetics and has shown me how a combination of curiosity, creativity, and hard work can make an outstanding scientist. His advice on and support for my research are essential for me to finish my Ph.D. study. I also want to express my gratitude to my dissertation committee member Dr. Michael Boehnke. He has been giving me not only suggestions on my research but also advice and guidance on my career development. He has been a source of constant encouragement and support throughout my study. I also want to express my appreciation to my dissertation committee members Drs. James Elder, Kerby Shedden, and Sebastian Zöllner. They have been providing me constantly with scientific guidance and suggestions during my Ph.D. study.

I have received many helpful suggestions on my research from many colleagues in the Department of Biostatistics and at the Center for Statistical Genetics. I had very insightful discussion with Debashis Ghosh, Thomas Nichols, William Stewart, Paul Scheet on the project presented in Chapter 2. Liming Liang gave me very helpful input on the project in

Chapter 3. I had a lot of useful discussion with Jian Kang when I performed the study in Chapter 4. My collaborators in the Department of Dermatology, Johann Gudjonsson, Rajan Nair, Philip Stuart provided me with all the data and biological insights for my project presented in Chapter 5. I owe a special thank you to all of them.

I have had very pleasant experiences working with many colleagues and have got their help and support. They are: Tom Blackwell, Wei Chen, Shyam Gopalakrishnan, Hyun Min Kang, Yun Li, Steve Qin, Laura Scott, Tanya Teslovich, Rui Xiao, and Matt Zawistowski. I thank them for the collegiality they have shown me.

Last but not least, I would like to thank my parents and my wife. My parents have made every effort to let me receive the best education available at every step of my growth. I hope this dissertation can be my way to say thank you to them. I owe a big deal to my late mother, Zenghui Shen. She has been and will always be my role model as a person. She has taught me to be kind, to be fair, and to be generous. Even though she is no longer with me, I know she would be proud of me as long as I try my best to be a good person. My wife, Jincao, has been a supporter, a comforter, and a cheer leader for me during all these years. Without her unyielding love and unselfish support, I would not have been able to finish this dissertation. I deeply thank her for everything she has done for me.

Table of Contents

Chapter

List of Figures

Figure

List of Tables

Table

Abstract

Gene transcript levels can bridge genotypes and more complex phenotypes, including common human diseases and traits. Understanding the processes that regulate the expression of disease associated transcripts and, in parallel, understanding the impact of disease associated genetic variants on gene expression, could enhance our understanding of the biology of these complex traits. Advances in high-throughput gene expression profiling and genotyping technologies have made it possible to search for these connections on a genomic scale. My dissertation focuses on statistical methods for genome-wide studies that aim to identify genetic variants associated with gene expression levels. Such variants are called expression quantitative trait loci (eQTLs).

In Chapter 1, I use two case studies to discuss how genome-wide association studies of gene expression have the potential to address some of the new challenges raised by current genetic studies. In Chapter 2, I describe a practical method to identify genetic variants that are associated with the levels of many transcripts. In Chapter 3, I propose a novel method for estimating the eQTL overlap between two tissues. In Chapter 4, I extend the method proposed in the previous chapter by removing the constraint on the sample-splitting strategy and use simulation studies to assess the performance of the method. In Chapter 5, I perform eQTL mapping in skin tissues from psoriatic patients and normal controls, and build a large catalog of genetic variants influencing transcript levels in both normal and psoriatic skin. My work has the potential to lead to a better

understanding of the mechanisms of gene regulation and a better dissection of the effects of genetic variants on complex phenotypes, such as many common diseases.

Chapter 1

Introduction

1.1 The genetics of global gene expression

Gene transcript levels can serve as an intermediate phenotype that bridges genotypes and more complex organismal phenotypes, such as human diseases. Transcriptional regulation of gene expression is essential for almost every process in a cell and abnormal transcriptional regulation is likely to be involved in the etiology of many diseases. Advances in high-throughput gene expression profiling and genotyping technologies have recently enabled researchers to study the genetic variants that regulate gene expression at a genomic scale [Rockman and Kruglyak 2006]. Such genome-wide association studies of gene expression have identified thousands of genetic loci impacting the expression of specific transcripts. Each of these loci is called an expression quantitative trait locus (eQTL). The identification of eQTLs will enhance our understanding of global transcriptional regulation and regulatory variation. Furthermore, as genome-wide association (GWA) studies of diseases (see 1.2 for a detailed example) are likely to identify many susceptibility variants with no known functional effects [McCarthy, et al. 2008], eQTL studies provide potentially useful functional information for some of those variants, and hence facilitate efforts to understand the functional impact of newly identified susceptibility variants [Cookson, et al. 2009].

Genetic studies of global gene expression (i.e. eQTL analyses) were initially performed in model organisms (ranging from yeast [Brem, et al. 2002; Yvert, et al. 2003] to flies [Wayne and McIntyre 2002] to mice [Chesler, et al. 2005]), but more recent studies have directly examined human cells [Dixon, et al. 2007; Goring, et al. 2007; Morley, et al. 2004; Stranger, et al. 2007]. In these human studies, the vast majority of validated eQTLs map to within a few hundred kilobase pairs of the transcription unit they putatively regulate. Loosely, these loci are termed *cis*-eQTLs. In contrast to *cis*-eQTLs, loci located far from the transcripts they regulate have been much harder to identify in humans and are, loosely, termed *trans*-eQTLs. One example of the utility of eQTL analysis is the genome-wide association study for asthma reported by Moffatt and colleagues [Moffatt, et al. 2007]. The study showed that a set of non-coding genetic variants that is strongly associated with childhood asthma also regulates expression levels of *ORMDL3*, focusing attention on *ORMDL3* as a target for further functional studies. Most human studies have measured transcript abundance in blood cells (peripheral blood lymphocytes and immortalized lymphoblastoid cell lines, LCLs); only a small number of studies have examined it in other tissues (e.g. liver tissue [Schadt, et al. 2008]).

In the next section, I present two case studies from my collaborative research to show that studying the genetics of global gene expression (i.e. eQTL analysis) can be used to address new questions and challenges raised in the current genetic research. The first case study is a GWA study of psoriasis; the second case study is a differential gene expression study of psoriasis.

1.2 Two case studies from collaborative research

One of the major goals for genetic research is to identify genetic variants that predispose to human diseases, and hence to better understand the mechanism for diseases and to suggest new targets for therapeutic interventions. Recently, researchers have been performing genome-wide association (GWA) studies to study the genetic basis of many complex traits, including common human diseases. Many disease susceptibility loci have been successfully identified in this way for diseases including type 1 [Hakonarson, et al. 2007; Todd, et al. 2007] and type 2 diabetes [Saxena, et al. 2007; Scott, et al. 2007; Zeggini, et al. 2008; Zeggini, et al. 2007], inflammatory bowel disease [Duerr, et al. 2006; Parkes, et al. 2007; Rioux, et al. 2007], psoriasis [de Cid, et al. 2009; Nair, et al. 2009; Zhang, et al. 2009], and others. In our GWA study [Nair, et al. 2009] that aimed to identify genetic susceptibility factors for psoriasis, we genotyped 438,670 single nucleotide polymorphisms (SNPs) in 1,409 psoriasis cases and 1,436 controls of European ancestry, and then followed up 21 promising SNPs in an additional 5,048 psoriasis cases and 5,041 controls. Our results provide strong support for the association of at least seven genetic loci and psoriasis (each with combined p-value $< 5\times10^{-8}$): *HLA-C*, three genes involved in *IL-23* signaling, two genes that act downstream of *TNF-*α and regulate *NF-*κ*B* signaling, and one gene involved in the modulation of Th2 immune responses. Figure 1.1 presents genome-wide Manhattan plots displaying association scan results and a quantile-quantile (Q-Q) plot summarizing the distribution of the observed test statistics. In the Q-Q plot, even after excluding all SNPs in regions of replicated association (blue line), we still observe more loci with small p-values than we expect under the hypothesis of no association, indicating that if we follow up more SNPs, it is likely that we can identify more psoriasis susceptibility loci. This observation leads to an open question: how do we prioritize markers for follow-up studies after the initial GWA study? One way is to prioritize in a follow-up study SNPs that define an eQTL (eQTL SNPs), if we can show that eQTL SNPs are more likely to be associated with diseases than non-eQTL SNPs. The other new challenge raised by GWA studies is that most of the identified disease susceptibility loci have no known biological function. In this regard, eQTL studies can provide potentially useful functional information for some of those variants by indicating that they may function via regulating the expression level of a certain gene.

The second case study is a differential gene expression study of psoriasis, where we characterized differences in gene expression among three types of skin tissues: normal skin from healthy controls (normal skin), non-diseased (uninvolved skin) and diseased skin (lesional skin) from psoriatic patients. Our analysis reveals that many genes are differentially expressed between involved and uninvolved/normal skin, many of which involve in immune response, defense response, and keratinocyte differentiation [Gudjonsson, et al. 2010; Gudjonsson, et al. 2009]. Figure 1.2 presents the first two principal components from the principal component analysis (PCA), where expression profiles for lesional skin are markedly different from those of normal and uninvolved skin. We achieved near-perfect separation of lesional skin from normal and uninvolved skin, while the latter two skin types were intermixed. The new open question is: regarding the variants associated with gene expression, do different physiological conditions affect the existence of those eQTLs?

1.3 The scope of the dissertation

All new open questions raised in the last session can be addressed to some extent by genome-wide association studies of gene expression. My dissertation focuses on the statistical methods for genome-wide association studies of gene expression and their applications to the genetic study of psoriasis.

In Chapter 2, I develop a new statistical method to identify DNA variants that are associated with expression levels of multiple genes, so-called “master regulatory” single nucleotide polymorphisms (SNPs). Although many examples of DNA variants that regulate the expression of a single gene have now been identified using high throughput technologies, identifying master regulators of expression in humans has proved more challenging. While conventional methods assess significance of association for individual SNP-gene pairs by p-values and then highlight SNPs that are associated with large numbers of gene transcript levels, our method proposes a new statistic to summarize evidence for association between each SNP and all measured transcripts. This statistic summarizes not just the number of signals that exceed a particular threshold, but also the strength of these signals. In a genome-wide scan, we rank SNPs based on this summary statistic and determine significance by permutation. Simulation studies show that the new summary statistic is more powerful than conventional methods for detecting master regulatory SNPs. As an example, we applied our method to gene expression and genotype data on 200 lymphoblastoid cell lines. We identified potential master regulators of gene expression; they still need to be verified in later studies. Our approach has the potential to shed light on the global regulation of gene expression by genetic variants.

In Chapter 3, I propose a novel method for estimating the eQTL overlap between two tissues, which allows researchers to quantify the benefits of studying eQTLs in different tissues. A simple measure of the sharing of eQTLs can be obtained by examining the overlap of significant eQTL lists from two different tissues. Unfortunately, this naive approach likely underestimates the true proportion of overlapping signals, as current studies are underpowered, particularly for eQTLs of modest effect. We have developed a more accurate method. Our multi-step procedure starts by splitting the study/tissue with the larger sample size (Study 1) into two parts. One part identifies eQTLs in Study 1, and the second part provides unbiased estimators for the eQTL effect sizes. The power of the second study/tissue (Study 2) to detect overlapping eQTLs is then estimated, and is used to adjust the observed overlap percentage to derive the power-adjusted overlap percentage. When applied to compare *cis*-eQTLs detected in analyses of 57 skin biopsies and of 340 lymphoblastoid cell lines, our method shows that ~70% of eQTLs are shared between the two tissues, a much larger proportion than the naive estimate of 30-50%. Our results provide guidance to researchers contrasting eQTL results across tissues and a specific means to estimate accurately the proportion of overlapping eQTLs between tissues.

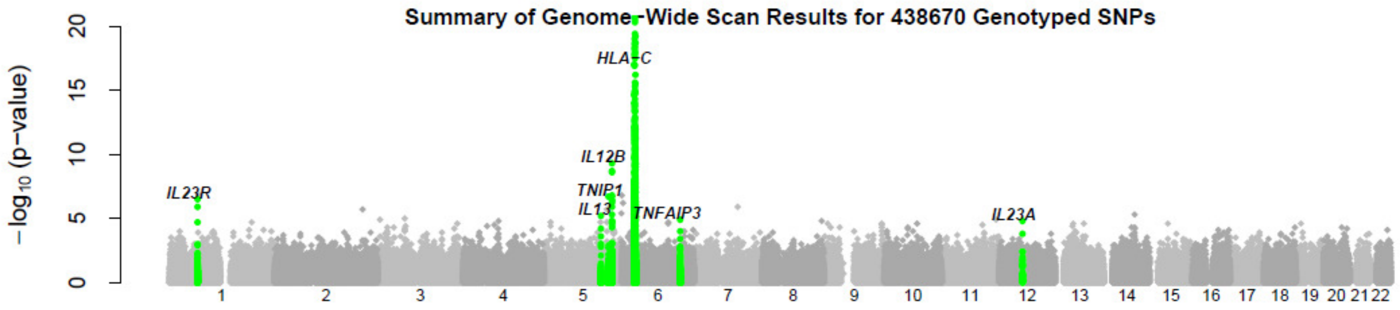
In Chapter 4, I use simulation studies to extend the method proposed in Chapter 3. More specifically, in Chapter 3, I apply a simplified version of our method, where I require that the second part of the split samples from Study 1 has the same sample size and same data structure as Study 2. Here, I extend this simplified version to the full version of the method by removing the constraint and use simulation to see if there exists an optimal sample splitting strategy. I also use simulation to test the performance of our method (in terms of the accuracy of the overlap estimation) when assumptions we make in Chapter 3 hold and when those assumptions are violated.

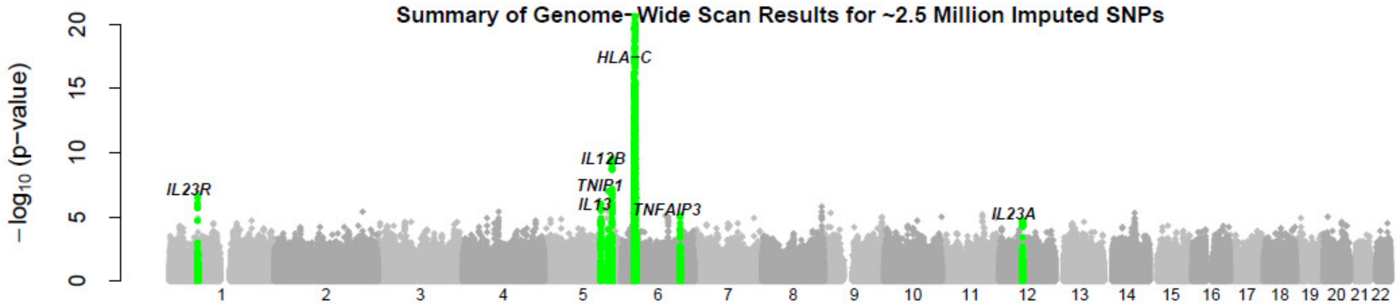
In Chapter 5, I perform the mapping of eQTLs in skin tissue from psoriatic patients and normal controls. Psoriasis, an immune-mediated, inflammatory disease of the skin and joints, provides an ideal system for eQTL mapping analysis, because it has a strong genetic basis and diseased tissue is readily accessible. To understand better the role of genetic variants regulating cutaneous gene expression in the pathogenesis of psoriasis, we identified 841 *cis*-acting eQTLs using RNA extracted from skin biopsies of 53 psoriatic individuals and 57 normal controls. We found substantial overlap between *cis*-eQTLs of normal control, uninvolved psoriatic, and lesional psoriatic skin. Consistent with recent studies and with the idea that control of gene expression can mediate relationships between genetic variants and downstream impact on disease risk, we found that eQTL SNPs are more likely to be associated with psoriasis than randomly selected SNPs. Our

results provide a catalog of *cis*-eQTLs in skin that can facilitate efforts to understand the functional impact of identified susceptibility variants of psoriasis and other skin traits.

1.4 Figures and Tables

Figure 1.1 Bird's eye view of psoriasis genome-wide association scan results.
The top panel summarizes the distribution of test statistics (i.e. p-values) at genotyped SNPs across the genome. We used a simple chi-squared test to compare SNP allele frequencies in cases and controls and plotted the resulting -log p-values across the genome. Several p-values < 10^{-20} in the MHC region were truncated. Loci where we obtained confirmatory evidence of association in follow-up samples are highlighted in green.
The middle panel summarizes the distribution of test statistics across the genome, after genotype imputation. We used a simple t-test to compare imputed allele counts in cases and controls and plotted the resulting –log p-values across the genome.
The bottom panel displays a Q-Q plot for our test statistics. Results are plotted including all SNPs (in red), after excluding SNPs in the MHC (in orange) and after excluding all SNPs in regions of replicated association (in blue). The shaded region represents a 90% confidence interval for the test statistics.


Summary of Genome-Wide Scan Results for 438670 Genotyped SNPs
−log10 (p−value)
0
5
10
15
20
HLA−C
IL12B
IL23R
TNIP1
IL13
TNFAIP3
IL23A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22


Summary of Genome−Wide Scan Results for ~2.5 Million Imputed SNPs
−log10 (p−value)
0
5
10
15
20
HLA−C
IL12B
IL23R
TNIP1
IL13
TNFAIP3
IL23A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

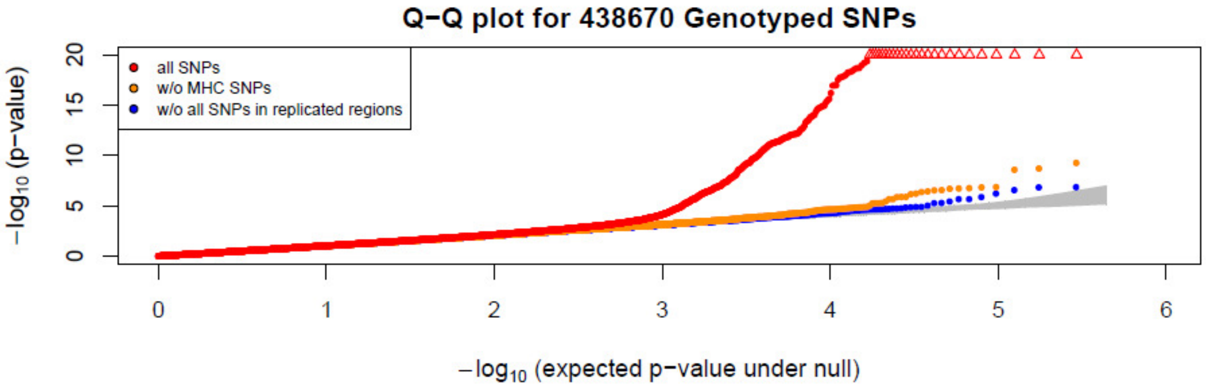
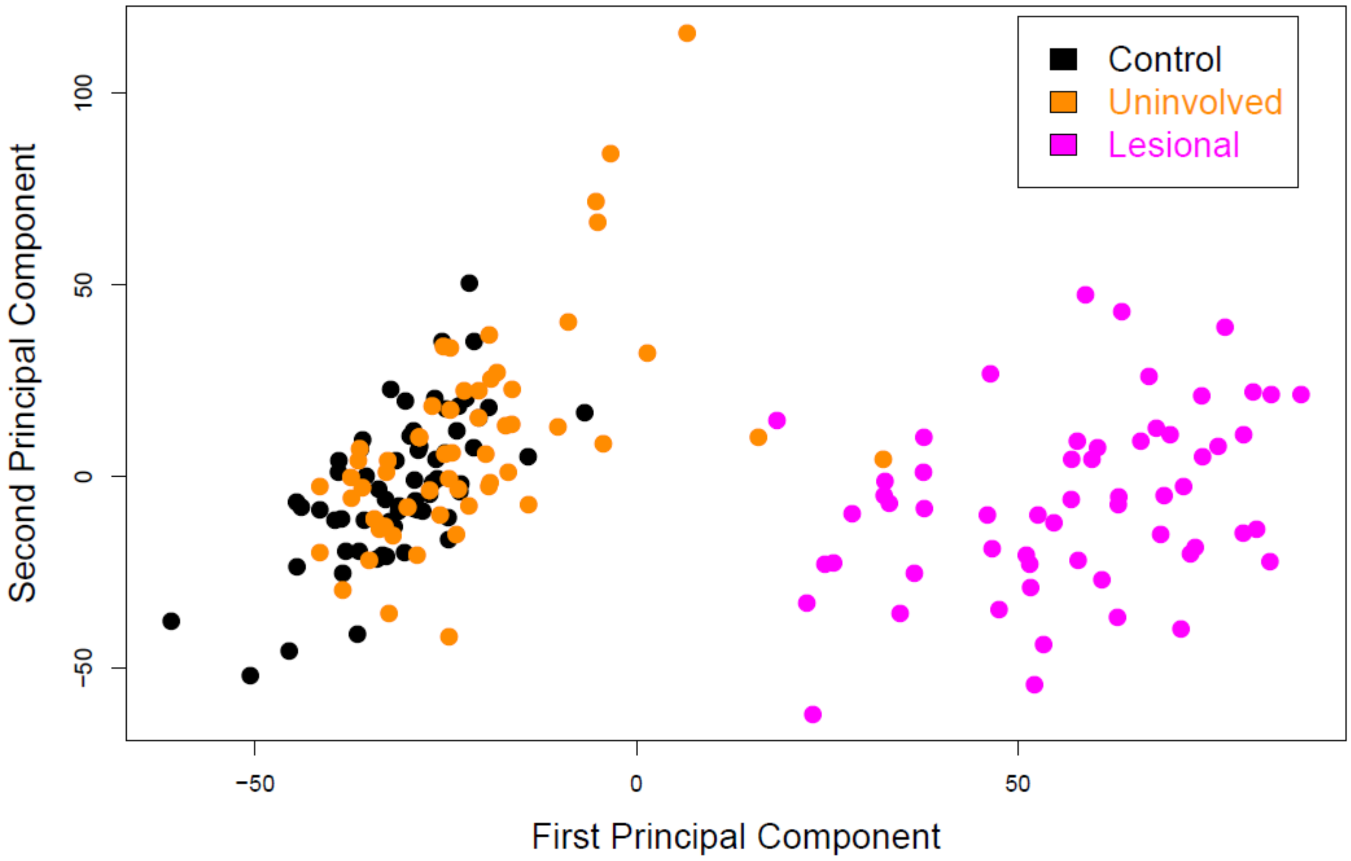

Q−Q plot for 438670 Genotyped SNPs
● all SNPs
● w/o MHC SNPs
● w/o all SNPs in replicated regions
−log10 (p−value)
0
5
10
15
20
0
1
2
3
4
5
6
−log10 (expected p−value under null)

Figure 1.2 Principal component analysis of all skin samples based on the gene expression data



Chapter 2

A statistical method for identifying genetic variants associated with the expression of many genes

2.1 Introduction

eQTL mapping has been conducted in several studies [Dixon, et al. 2007; Goring, et al. 2007; Morley, et al. 2004; Stranger, et al. 2007] using human cells (e.g. LCLs). Although many examples of *cis*-regulators of expression have now been mapped, identifying *trans*-regulators of expression has proved more challenging. This is probably because *cis*-regulators act very close to the gene they regulate and hence have large effect sizes while *trans*-regulators have smaller effect sizes. Within the framework of genome-wide association studies of gene expression, we focus on identifying genetic variants, or more specifically, single nucleotide polymorphisms (SNPs), that are associated with mRNA expression levels of multiple genes (master regulatory SNPs). Identification of master regulators of gene expression will pinpoint the key players (i.e. hub genes) in the transcriptional network, which in turn will lead to a better understanding of global regulation of gene expression. In addition, identification of master regulators of disease-related genes may pinpoint the causal variants of diseases, as those master regulators are likely to play a key role in the pathways leading to the diseases.

We envision that a master regulatory SNP could function via different mechanisms (Figure 2.1). It could act as a master regulator through a regulator chain (Figure 2.1A) or by affecting a single regulator, which in turn regulates the expression of multiple genes (Figure 2.1B). In both of these mechanisms, we expect that the expression levels are highly correlated for the group of genes that are regulated by the SNP. Meanwhile, a master regulatory SNP could hypothetically regulate two groups of genes via different mechanisms, as shown in Figure 2.1C. We expect that the two groups of genes are not correlated in their expressions. We have some evidence for the existence of master regulators of gene expression. For example, previous studies have shown that there are transcription factors that regulate the expression of multiple genes. In yeast, a study showed that the number of different promoter regions bound by a given transcription factor ranged from 0 to 181, with an average of 38 promoter regions per regulator [Lee, et al. 2002]. This observation indicates that there should be "master" genes (e.g. “master” transcription factors) that regulate mRNA expression levels of multiple genes.

Currently there is little research in the literature that aims to identify systematically the master regulatory SNPs of gene expression. A natural and relatively simple method to identify them is to first assess significance of association for individual SNP-gene pairs by p-values and then highlight SNPs that are significantly associated with large numbers of gene expression levels. However, this type of counting method ignores the significance of associations once the p-values pass the threshold at the first step. This will likely lead to the diminished statistical power to identify master regulatory SNPs. We propose a new summary statistic that simultaneously takes into account both significance levels and the number of association signals. In a genome-wide scan, we rank SNPs based on this summary statistic and determine significance by a permutation test. We expect that our approach may increase our ability to identify genetic variants that have a modest impact on the expression of many genes. Since we expect that *trans* eQTLs might influence the transcription levels of many genes, we hope our method will facilitate efforts to identify *trans* eQTLs in humans.

2.2 Methods

Preamble

Consider a dataset that includes genotype data at M genetic markers as well as expression levels for N transcripts. Let p_{ij} be a p-value that summarizes the evidence for association between a particular genetic marker $i = 1,\ldots,M$ and a particular transcript $j = 1,\ldots,N$. This p-value might be calculated by using regression based approaches to investigate the association between each marker and each transcript, but any valid p-value could be used.

A simple way to identify genetic variants that are associated with expression levels of particular transcripts would be to define an appropriate threshold T and declare that marker-transcript pairings for which $p_{ij} < T$ are significant. An appropriate choice of T could be defined using the Bonferroni rule, using False-Discovery Rate [Benjamini and Hochberg 1995], or using a permutation based approach. Although each of these choices would likely identify a set of markers that are strongly associated with the expression of one or more transcripts, they would not be particularly effective at identifying variants that are weakly associated with the expression of many transcripts. These variants might, for example, regulate the activity or expression of particular transcription factor that then influences the expression of many transcripts.

The Area Under the Log Curve (AULC) Statistic

To identify genetic variants that are modestly associated with the levels of many transcripts, potentially with none of the signals exceeding a stringent significance threshold, it is necessary to combine evidence for association across transcripts. One simple strategy would be to calculate a statistic such as:

$$AUC_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij}$$

We call it *AUC* (Area Under the Curve) because if we plot the ordered p-value vector $p_{i,(1)} \leq p_{i,(2)} \leq ... \leq p_{i,(N)}$ against *j/N, j*=1,…,*N*, AUC_i is actually proportional to the area under that curve (see Figure 2.2 for this geometric interpretation). However, because we expect *N* will be much larger (typically, 5,000 – 50,000) than the number of transcripts potentially associated with a single genetic marker, we expect that this statistic would be dominated by noise. A simple improvement might be to focus summation on a set of smallest p-values for each marker [Wille, et al. 2003], but the choice of the appropriate number of p-values to sum is not obvious. Instead, to increase the contribution of truly associated markers to the statistic while avoiding an arbitrary thresholding procedure, we propose to (i) sum log-pvalues, instead of raw p-values, so that transcripts showing tentative evidence for association make larger contributions to the overall statistic and, in addition, (ii) use a weighting scheme that focuses attention on the smallest observed p-values for each marker.

Our proposed statistic, which we term Area Under the Log-Curve (AULC) is defined thus:

$$AULC_i = \frac{1}{C}\sum_{j=1}^{N}\{w_{ij} \times [-\log(p_{ij})]\}$$

Here w_{ij} is the weight for $-\log(p_{ij})$. In our case we want w_{ij} to be a decreasing function of the rank of p_{ij} (i.e. the smallest observed p-value has rank 1, the second smallest p-value has rank 2, and so on) relative to all other p-values for marker i such that smaller observed p-values have larger weights. C is a normalizing constant defined as $C = \sum_{j=1}^{N} w_{ij}$. In practice, the choice of C does not matter as we use permutation to assess significance. The AULC statistic is thus a weighted sum of -log(p-values). To determine w_{ij}'s, we borrow strength from the following log-log plot of ordered p-value vector: for $p_{i,(1)} \leq p_{i,(2)} \leq ... \leq p_{i,(N)}$, we plot -log($p_{i,(j)}$) against -log($j/N$), j=1,…,N. Weights w_{ij}'s are determined such that AULC statistic is proportional to the area under this log curve. This geometric interpretation of the statistic is given in Figure 2.2 and the exact formula for the AULC statistic is provided in Appendix A. Markers with larger AULC statistics are, in principle, more likely to be associated with expression levels of one or more transcripts.

Assessing Significance Levels

Given the number of transcripts N examined for each marker, the distribution of the AULC statistic under the null hypothesis can be derived analytically if we assume the expression levels of N transcripts are uncorrelated. However, the exact distribution of the statistic is harder to derive because it depends on the potentially complex correlation structure describing the relationships between different transcript levels and between different genetic markers. We thus use permutations to assess the significance of the AULC statistic (and also to assess the significance of alternative statistics presented in this manuscript).

To maintain the correlation structure of the gene expression and genotype data, we keep the gene expression data for each individual unchanged and shuffle the observed genotype data between individuals. Genotypes are shuffled in blocks so that all genotypes originally observed in any individual are re-assigned to a new individual, selected at random. After each permutation we calculate a new set of p-values p_{ij} and AULC$_i$ statistics. We then record the maximum observed AULC$_i$ statistic for each permuted dataset; comparing the original observed statistics to this distribution of maximum statistics allows us to control for the family-wise error (FWE) rate and calculate accurate p-values for the AULC statistic [Nichols and Hayasaka 2003].

Simulations

We first used simulations to assess the statistical power of the AULC method and compare its performance to methods that rely on a threshold T to identify significantly associated SNP transcript pairs and that then count the number of p-values exceeding these thresholds for each marker to identify potential master regulators of gene expression. We considered the definition of T based on the Bonferroni rule, and called it the Bonferroni counting method.

To mimic the correlation structures observed in actual gene expression datasets, our simulation procedure starts with an actual gene expression dataset to which a synthetic signal is then added. We used the Dixon et al. [2007] lymphoblastoid expression data as a starting point. Measurements of transcript levels were first transformed to approximate normality by applying an inverse normal transformation to each transcript in turn. We then simulated genetic variants that impact the expression of many transcripts by: (i) simulating genotypes for a di-allelic marker for each individual, under the assumption of Hardy-Weinberg equilibrium and with minor allele frequencies ranging of either 0.1 or 0.5; (ii) selecting a set of transcripts (typically, including 5 – 500 transcripts) whose expression levels are influenced by the marker; and (iii) adding a small marker effect onto each of the observed transcript levels (as described in Appendix C). Finally, we use simulated expression data and genotype data to calculate both a set of AULC statistics and, using permutations, the distribution of the maximum AULC statistic under the null.

In our simulations, we control the *average* impact of each marker on the expression levels of associated genes. Specifically, we define a quantity R_i^2 corresponding to the average variance explained by marker *i* for each of the associated transcripts. We let, for each associated transcript, the impact of the marker be larger or smaller than this average. In particular, we let R_{ij}^2 follow a truncated (from 0 to 1) exponential distribution with mean $R_i^2 = \lambda$. This implies that each marker has a slightly larger impact on expression levels of a few transcripts, and a more modest impact on the levels of most other transcripts.

We simulated datasets including transcript levels for either 200 or 1000 individuals, with 200 being a sample size representative of contemporaneous studies and 1,000 being a possible sample size for studies in the near future. We let mean $R_i^2 = \lambda$ take values of .01, .02, .05 and 0.10. To compare the statistical power of different methods to detect true master regulators, we always control the type I error rate at 10^{-5} (i.e. we expect one false positive among 100,000 SNPs).

Lymphoblastoid Expression and Genotype Data Analysis

To evaluate the performance of our method in a real dataset, we re-analysed the data of Dixon et al. [2007]. We focus on 200 unrelated subjects from that study. The dataset includes Affymetrix U133 Plus 2.0 Array gene expression measurements for 54,675 probe sets, which we used to calculate mean transcript levels for each of 20,596 genes. In addition, the dataset includes genotype data of 408,289 SNP markers obtained using Illumina Beadchips. Prior to analysis, we adjusted for sex effects by regressing out the impact of sex on each transcript.

For each SNP-expression pair, we fitted a linear model to test the association between a SNP and a gene expression. We treated gene expression measurements as the outcome variable and the observed number of minor alleles as the predictor variable. In addition to analyzing the original data, we analyzed 100 permuted datasets to derive empirical significance levels. We obtained a nominal p-value for each marker by comparing the observed AULC statistic with the null distribution of all AULC statistics from all permutations and the FWE-corrected p-value by comparing the observed AULC statistic with the null distribution of maximum AULC statistics from each permutation.

2.3 Results

Simulations

As described in the methods, we used a set of actual gene expression data as a starting point for our simulations. We then simulated genotypes for each individual and generated a small genetic effect which was summed to the observed expression levels. This strategy ensures that our simulated data closely mimic the complex correlation patterns observed in actual gene expression datasets. Since we used permutations to derive empirical significance levels, all methods have adequate type I error rates and here we focus on comparing power. We contrast the ability of our proposed AULC statistic to strategies that use a Bonferroni correction approach to identify genetic markers significantly associated with the expression levels for one or more transcripts.

We first examined simulated samples of 200 individuals and simulated SNPs with a minor allele frequency of 0.50 (Table 2.1 and Figure 2.3). At this sample size, it was not possible to detect SNPs associated with the expression of 5 – 500 transcripts explaining (on average) either 1% or 2% of the variation in transcript levels for each of these. For SNPs that explained (on average) 5% of the variance in transcript levels in 200 or more transcripts, the AULC method had at least 50% power; this rises to >80% power for SNPs associated with 275 or more transcripts. In contrast, even with 500 associated transcripts the Bonferroni approach had <30% power. Again, for SNPs that explained (on average) 10% of the variance in transcript levels for 50 or more transcripts, the AULC method had >80% power. The Bonferroni approach required >250 associated transcripts to achieve similar power.

Although these results show the performance of our method is promising, it is not clear that the simulated effect sizes are realistic. Thus, we repeated our experiment with a larger simulated dataset, including 1,000 individuals. We observe a substantial power gain (Table 2.2 and Figure 2.4). Now, even when the marker accounts on average for only 1% of the variance in each associated transcript, the AULC method can achieve 80% power when n≥300. When average variance explained rises to 2%, greater than 80% power is achievable with >50 associated genes (versus 325 for the Bonferroni method).

Repeating our simulations with different minor allele frequencies produced similar results (see Supplementary Tables 2.1 and 2.2 for power results when minor allele frequency is 0.1). This result is what we expect to observe because the allele frequency has been taken into account in the effect size parameter R_i^2 (see Appendix C for details).

Analysis of Dixon et al. [2007] Dataset

We next applied our method to the Dixon et al. [2007] lymphoblastoid gene expression genome-wide association study. In total, the dataset analyzed includes gene expression levels for 20,596 genes; genotype data for 408,289 SNPs; and 200 unrelated subjects. We first calculated the AULC statistic for each SNP and used permutations to assess significance for the resulting statistics.

The top ranked SNP (rs6787837) has a nominal p-value of 4.5×10^{-7}; after adjustment for multiple testing the family-wise error rate (FWE) for this SNP is 0.16; not reaching genome-wide significance. The same SNP was also the top ranked SNP using the Bonferroni counting method with a nominal p-value at 4.6×10^{-7} and an FWE-corrected p-value at 0.15. Thus, with the current sample of size of 200 individuals, neither method identified a signal exceeding genome-wide significance thresholds.

Even though the top ranked SNP does not quite reach genome-wide significance levels, we note that it maps near *TMF1* (TATA element modulatory factor 1). The TMF1 protein inhibits transcriptional activation by the TATA-binding protein (TBP). Since TBP is essential for the transcription of most genes, it seems plausible that variants in the *TMF1* locus could impact the expression of many genes. A gene ontology enrichment analysis of transcripts showing strong association (p-value $< 2.4\times10^{-6}$, which is the Bonferroni-corrected p-value threshold after adjusting for testing 20,596 gene transcripts for the marker) with rs6787837 suggested several categories of genes were over-represented in this set: genes involved in regulation of transcription (49 genes, p-value at 2.0×10^{-6}), regulation of metabolism (54 genes, p-value at 5.7×10^{-7}), and zinc ion binding (60 genes, p-value at 4.0×10^{-10}).

2.4 Discussion

We describe a method for identifying genetic markers associated with transcript levels of multiple genes. The method provides a strategy for systematically identifying master regulators of gene expression and may facilitate the identification of *trans* eQTLs. The method uses a summary statistic (the AULC) that summarizes both the number of association signals and the strength of each individual signal. The statistic has a geometric interpretation, corresponding to the area under the curve in a familiar log-quantile plot of log-pvalues. These plots are often used for data quality control in genome-wide association scans [McCarthy, et al. 2008]. In our simulations, the AULC method outperformed the Bonferroni counting method, which focuses on markers that are strongly associated with the expression of multiple genes using a threshold procedure.

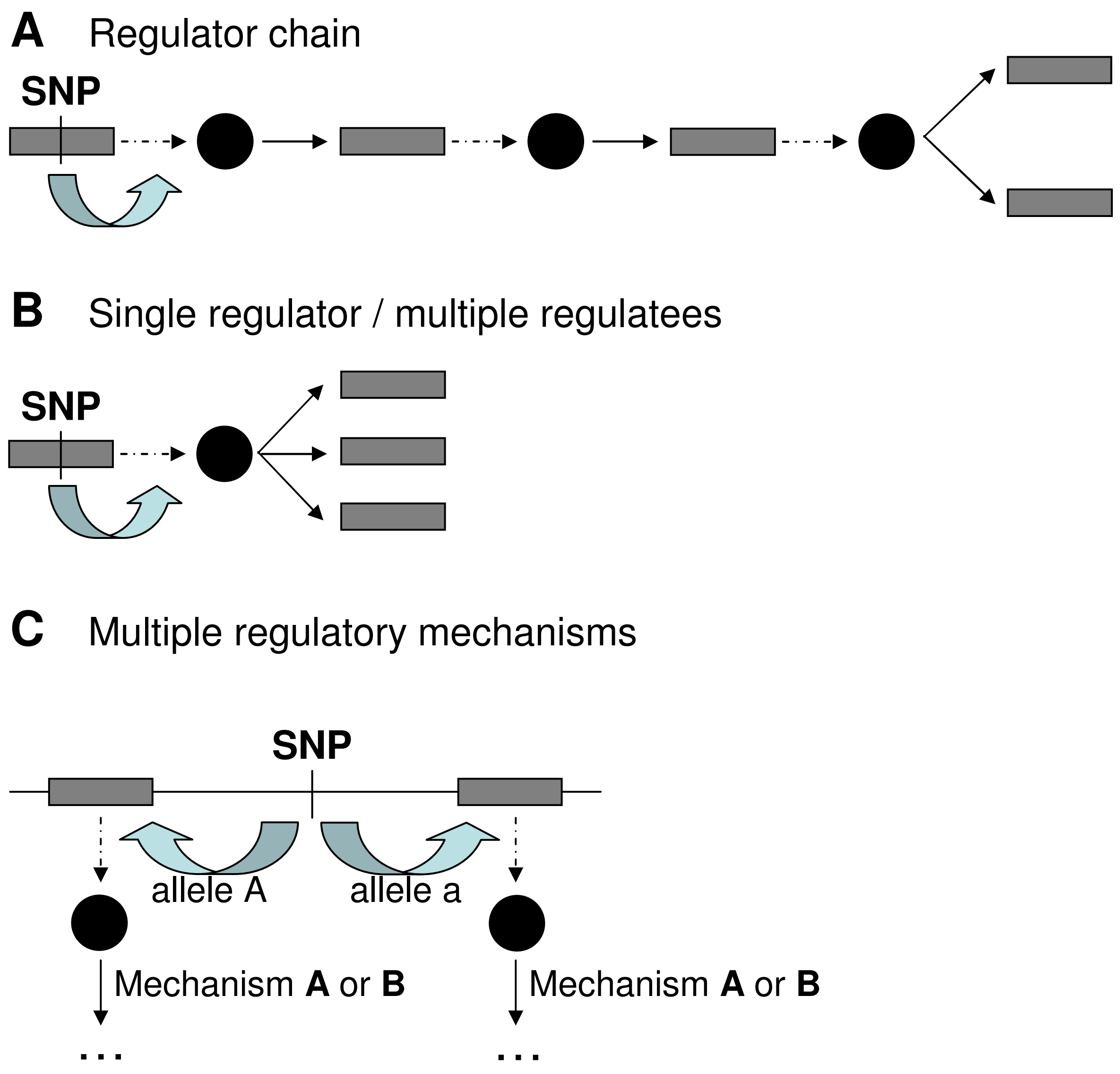
While the AULC method is more powerful than conventional methods that group signals into likely true positives and likely false positives using a Bonferroni or FDR threshold, our simulations also show that our method, as well as conventional methods, can benefit greatly from increased sample sizes. With a sample size of 200 individuals, even when the average variation in expression explained by the associated marker equals 5%, a study can achieve 80% power only when more than 275 genes are associated with the same marker. With a sample size of 1000 individuals, it should be possible to identify markers that on average account for 2% of the variance in 50 transcripts or more. Overall, we expect that large gene expression studies now planned or ongoing will lead to the identification of master regulators of gene expression.

Our re-analysis of the data of Dixon et al. [2007] identified a potentially interesting signal near the *TMF1* (TATA element modulatory factor 1) locus. This signal did not reach genome-wide significance (after adjusting for testing many markers) and did not replicate in the genetic study of gene expression on HapMap samples [Stranger, et al. 2007], where our method was applied to the data for 60 unrelated subjects from 30 Caucasian trios of northern and western European origin (CEU). The lack of replication could simply reflect a lack of power or it could suggest our original signal is a false-positive. In either case, future analysis of larger samples should clarify the situation.

We hope the AULC statistic will prove useful in a variety of settings where high-dimensional data can first be summarized in a series of p-values. A natural refinement of the analysis described here, where we sought markers associated with transcript levels of many genes, is to search for markers that are associated with transcripts that show evidence of differential expression between diseased and normal individuals. In this case, one could simply imagine stratifying the original analysis to focus on differentially expressed transcripts, or changing the weighting scheme so these transcripts have greater weights. But we expect that even more interesting extensions may be possible. For example, the method may aid in evaluating the impact of gene markers on brain activity by enabling searches for genetic variants that are associated with brain activity levels measured at many points in a functional MRI (functional magnetic resonance imaging) experiment. Brain activity at each point is typically summarized in an image voxel (whose intensity summarizes the degree of activation of that area of the brain) and could naturally serve as an outcome in a regression analysis. Because a typical brain image easily includes 100,000s of voxels, the ability of the AULC statistic to summarize evidence for association seems extremely attractive.

2.5 Figures and tables

Figure 2.1 Different potential mechanisms by which a master regulatory SNP regulates expression of multiple genes. (A) A master regulatory SNP acts through a regulator chain. (B) A master regulatory SNP affects a single regulator, which in turn regulates the expression of multiple genes. In (A) and (B), the SNP could be in the regulatory region of the corresponding gene, resulting in altered expression of the gene, or it could be in the coding region of the gene, resulting in altered function of the protein product of the gene. (C) A master regulatory SNP is in the vicinity region of gene 1 and 2. The allele *A* of the SNP affects gene 1, which regulates a group of genes' expression via mechanisms (A) or (B). Meanwhile, the allele *a* of the SNP affects gene 2, which regulates another group of genes' expression via mechanisms (A) or (B). In the diagrams, squares represent genes and circles represent proteins. Broken lines indicate protein production from a gene while solid arrows represent regulatory interactions. (A) and (B) are modified from Wittkopp (2007) [Wittkopp 2007].



A Regulator chain

B Single regulator / multiple regulatees

C Multiple regulatory mechanisms

Figure 2.2 Geometric interpretations for the AUC and AULC statistics. The area under the curve in the upper panel is proportional to the AUC statistic. The area under the curve in the lower panel is proportional to the AULC statistic. A vector of 200 p-values are plotted in this example (i.e. $N = 200$).

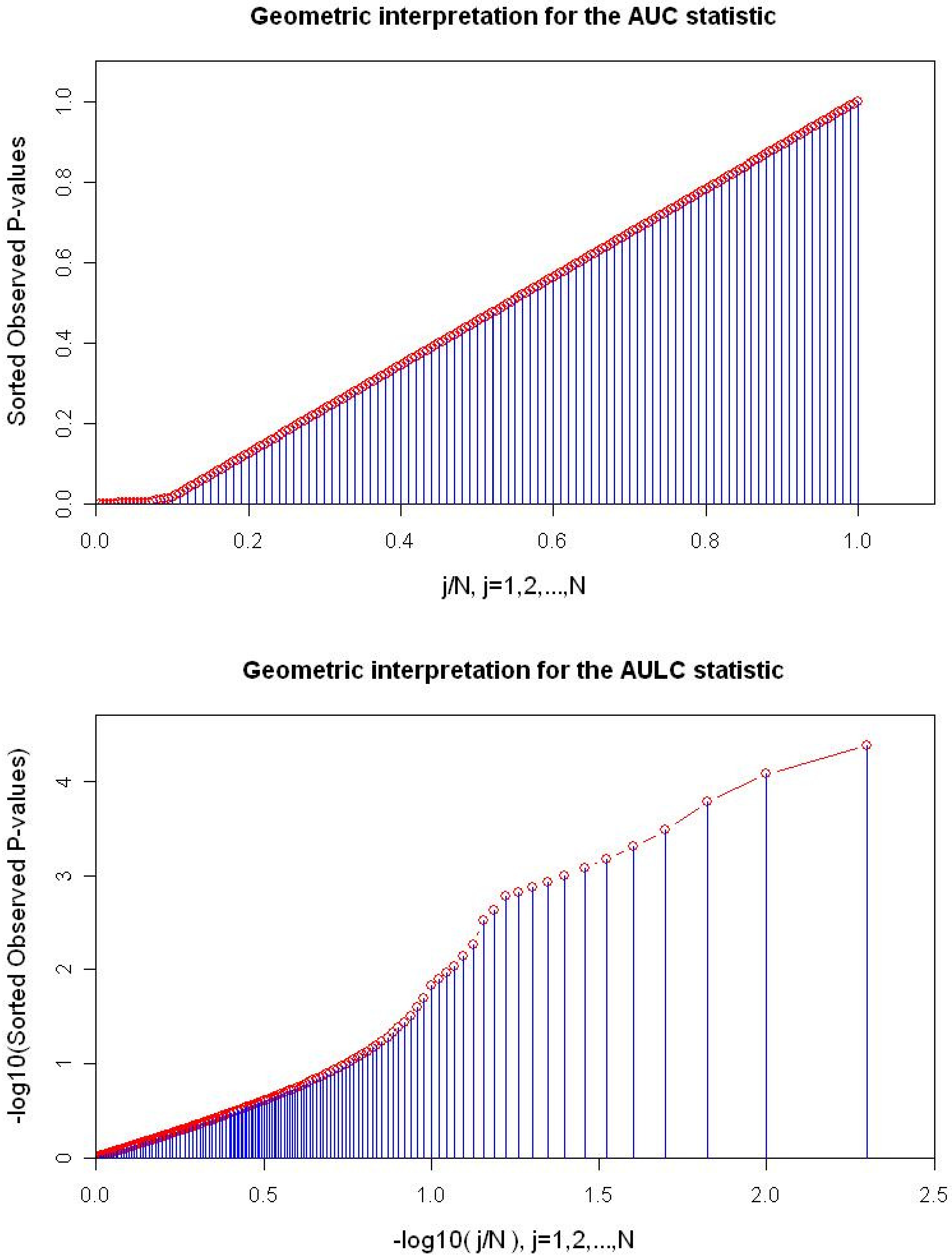


Figure 2.3 The statistical power from simulation studies for the AULC method and the Bonferroni counting method with the sample size at 200 and the minor allele frequency of the master regulatory SNP at 0.5. The results are shown when the effect size (mean R^2) equals to 0.05 or 0.1.

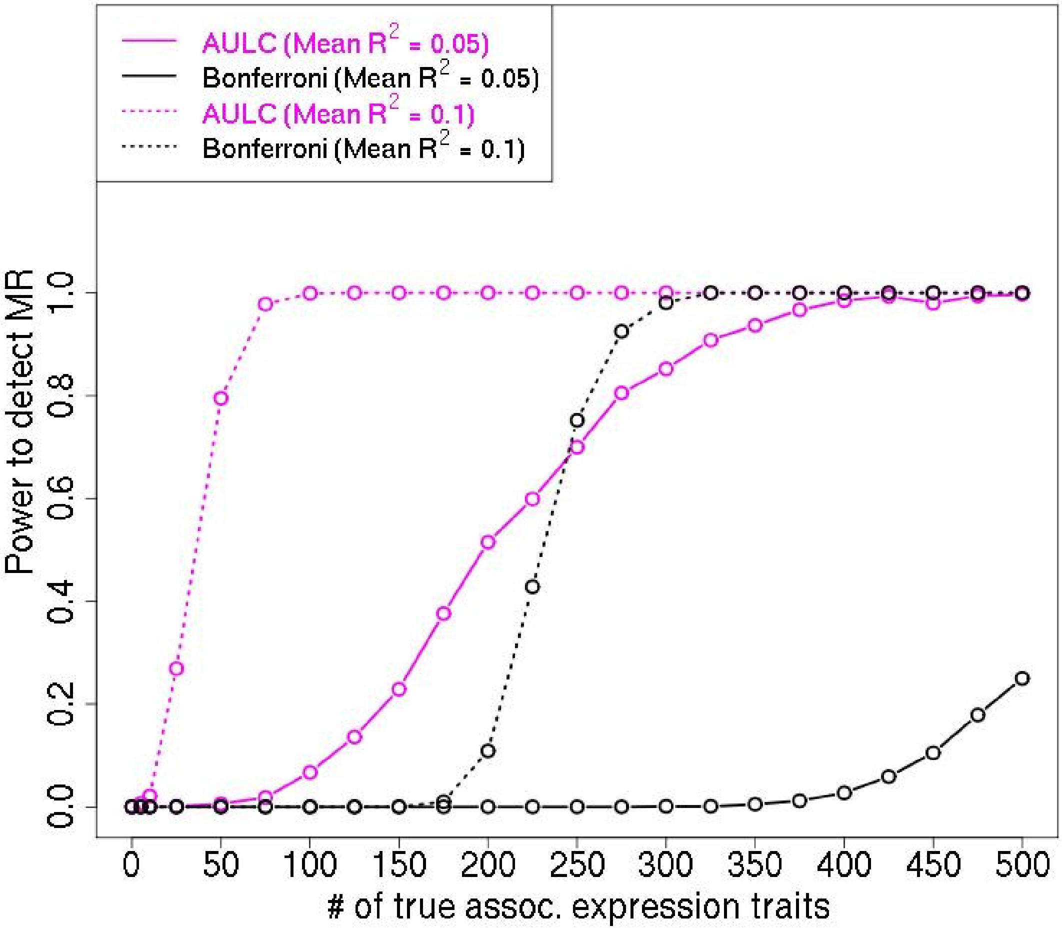


Figure 2.4 The statistical power from simulation studies for the AULC method and the Bonferroni counting method with the sample size at 1000 and the minor allele frequency of the master regulatory SNP at 0.5. a) The effect size (mean R^2) equals to 0.01 or 0.02; b) The effect size (mean R^2) equals to 0.05 or 0.1.

Figure 2.4a.

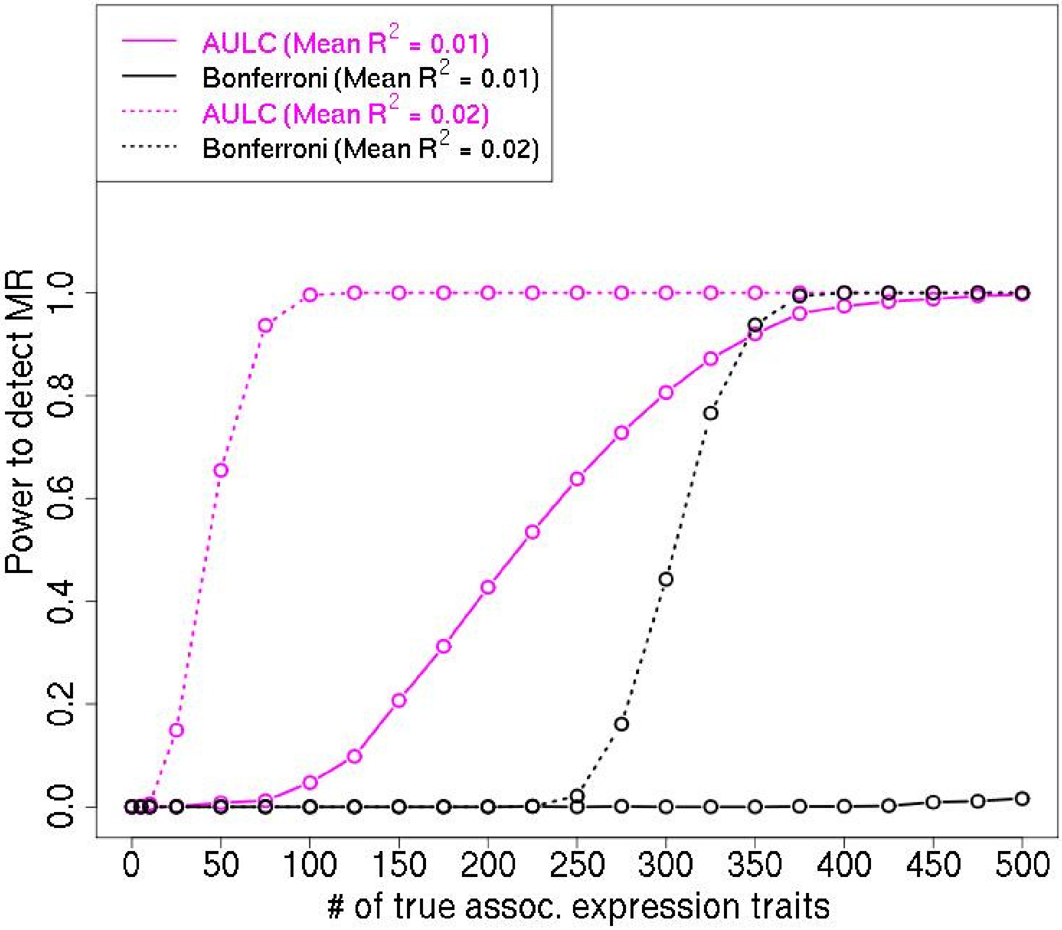


Figure 2.4b

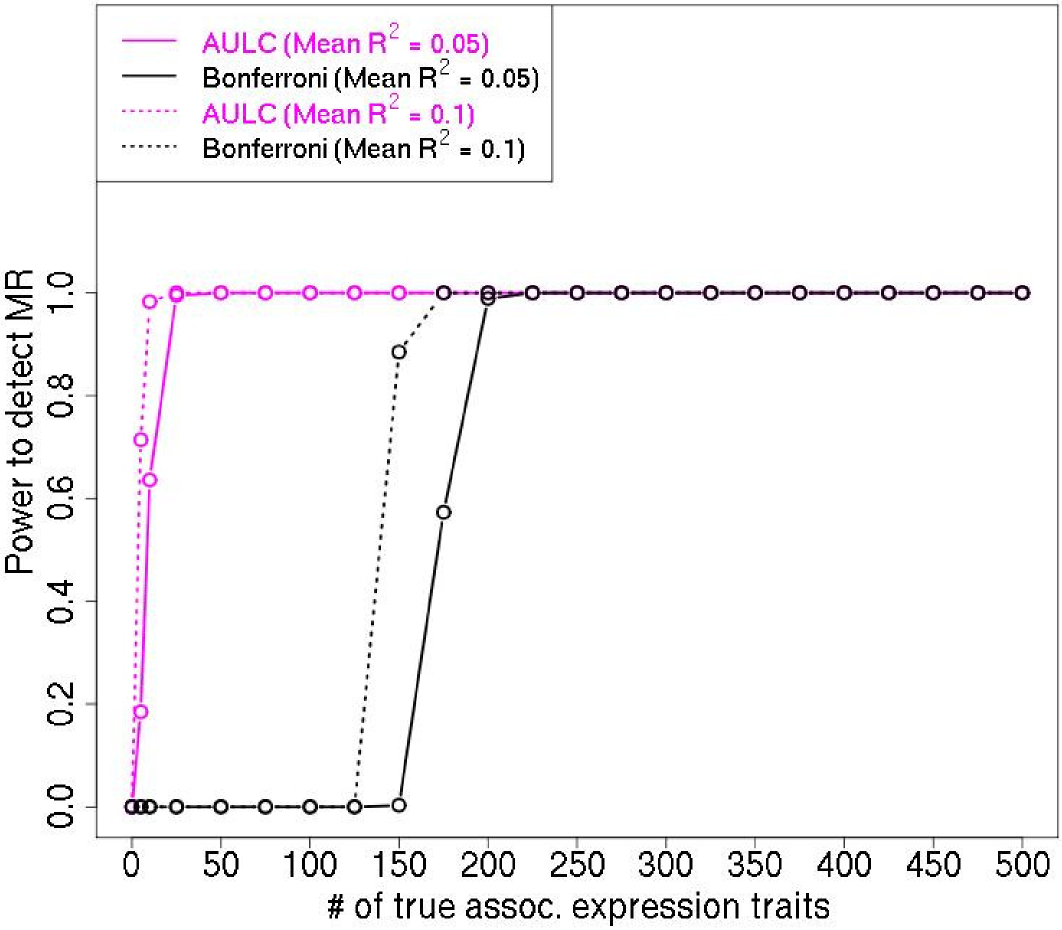

AULC (Mean R^2 = 0.05)
Bonferroni (Mean R^2 = 0.05)
AULC (Mean R^2 = 0.1)
Bonferroni (Mean R^2 = 0.1)
Power to detect MR
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
400
450
500
# of true assoc. expression traits

Table 2.1 The statistical power from simulation studies for the AULC method and the Bonferroni counting method with the sample size at 200 and the minor allele frequency of the master regulatory SNP at 0.5

	Statistical Power							
	mean R^2 = 0.01		**mean R^2 = 0.02**		**mean R^2 = 0.05**		**mean R^2 = 0.1**	
#associations	AULC	Bonferroni	AULC	Bonferroni	AULC	Bonferroni	AULC	Bonferroni
0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0.007	0
10	0	0	0	0	0	0	0.021	0
25	0	0	0	0	0.001	0	0.269	0
50	0	0	0	0	0.006	0	0.795	0
75	0	0	0	0	0.018	0	0.978	0
100	0	0	0	0	0.067	0	0.999	0
125	0	0	0	0	0.136	0	1	0
150	0	0	0	0	0.229	0	1	0
175	0	0	0	0	0.376	0	1	0.010
200	0	0	0	0	0.515	0	1	0.109
225	0	0	0	0	0.599	0	1	0.428
250	0	0	0	0	0.700	0	1	0.752
275	0	0	0.001	0	0.805	0	1	0.925
300	0	0	0.002	0	0.852	0.001	1	0.981
325	0	0	0.002	0	0.908	0.001	1	1
350	0	0	0	0	0.937	0.005	1	1
375	0	0	0.003	0	0.967	0.012	1	1
400	0	0	0	0	0.985	0.027	1	1
425	0	0	0.002	0	0.993	0.059	1	1
450	0	0	0.002	0	0.980	0.105	1	1
475	0	0	0.004	0	0.994	0.179	1	1
500	0	0	0.001	0	0.997	0.250	1	1

Table 2.2 The statistical power from simulation studies for the AULC method and the Bonferroni counting method with the sample size at 1000 and the minor allele frequency of the master regulatory SNP at 0.5

	Statistical Power							
	mean R^2 = 0.01		**mean R^2 = 0.02**		**mean R^2 = 0.05**		**mean R^2 = 0.1**	
#associations	AULC	Bonferroni	AULC	Bonferroni	AULC	Bonferroni	AULC	Bonferroni
0	0	0	0	0	0	0	0	0
5	0	0	0.001	0	0.185	0	0.714	0
10	0	0	0.006	0	0.636	0	0.983	0
25	0.001	0	0.149	0	0.995	0	1	0
50	0.008	0	0.655	0	1	0	1	0
75	0.012	0	0.937	0	1	0	1	0
100	0.047	0	0.996	0	1	0	1	0
125	0.098	0	1	0	1	0	1	0
150	0.207	0	1	0	1	0.003	1	0.885
175	0.312	0	1	0	1	0.573	1	1
200	0.427	0	1	0	1	0.989	1	1
225	0.535	0.001	1	0.001	1	1	1	1
250	0.638	0	1	0.021	1	1	1	1
275	0.728	0.001	1	0.161	1	1	1	1
300	0.806	0	1	0.443	1	1	1	1
325	0.872	0	1	0.766	1	1	1	1
350	0.92	0	1	0.938	1	1	1	1
375	0.96	0.001	1	0.994	1	1	1	1
400	0.974	0.001	1	1	1	1	1	1
425	0.983	0.002	1	1	1	1	1	1
450	0.988	0.009	1	1	1	1	1	1
475	0.994	0.011	1	1	1	1	1	1
500	0.997	0.016	1	1	1	1	1	1

Supplementary Table 2.1 The statistical power from simulation studies for the AULC method and the Bonferroni counting method with the sample size at 200 and the minor allele frequency of the master regulatory SNP at 0.1

	Statistical Power							
	mean R^2 = 0.01		**mean R^2 = 0.02**		**mean R^2 = 0.05**		**mean R^2 = 0.1**	
#associations	AULC	Bonferroni	AULC	Bonferroni	AULC	Bonferroni	AULC	Bonferroni
0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0.003	0
10	0	0	0	0	0	0	0.023	0
25	0	0	0	0	0	0	0.249	0
50	0	0	0	0	0.008	0	0.737	0
75	0	0	0	0	0.023	0	0.954	0
100	0	0	0	0	0.084	0	0.987	0
125	0	0	0	0	0.163	0	0.999	0
150	0	0	0	0	0.263	0	1	0
175	0	0	0	0	0.353	0	1	0.023
200	0	0	0	0	0.472	0	1	0.132
225	0	0	0	0	0.551	0	1	0.398
250	0	0	0	0	0.625	0	1	0.631
275	0	0	0	0	0.701	0	1	0.814
300	0	0	0.004	0	0.755	0.006	1	0.915
325	0	0	0.001	0	0.801	0.013	1	0.964
350	0	0	0	0	0.835	0.017	1	0.983
375	0	0	0	0	0.882	0.033	1	0.996
400	0	0	0.001	0	0.904	0.069	1	0.999
425	0	0	0.001	0	0.939	0.125	1	0.998
450	0	0	0.001	0	0.931	0.155	1	1
475	0	0	0	0	0.951	0.217	1	1
500	0	0	0.003	0	0.957	0.309	1	1

Supplementary Table 2.2 The statistical power from simulation studies for the AULC method and the Bonferroni counting method with the sample size at 1000 and the minor allele frequency of the master regulatory SNP at 0.1

	Statistical Power							
	mean R^2 = 0.01		**mean R^2 = 0.02**		**mean R^2 = 0.05**		**mean R^2 = 0.1**	
#associations	AULC	Bonferroni	AULC	Bonferroni	AULC	Bonferroni	AULC	Bonferroni
0	0	0	0	0	0	0	0	0
5	0	0	0	0	0.184	0	0.716	0
10	0	0	0.009	0	0.624	0	0.982	0
25	0	0	0.151	0	0.997	0	1	0
50	0.007	0	0.660	0	1	0	1	0
75	0.017	0	0.932	0	1	0	1	0
100	0.043	0	0.993	0	1	0	1	0
125	0.121	0	0.998	0	1	0	1	0
150	0.204	0	1	0	1	0	1	0.622
175	0.317	0	1	0	1	0.324	1	1
200	0.426	0	1	0	1	0.960	1	1
225	0.514	0	1	0	1	1	1	1
250	0.635	0	1	0.016	1	1	1	1
275	0.731	0	1	0.117	1	1	1	1
300	0.800	0	1	0.322	1	1	1	1
325	0.842	0	1	0.606	1	1	1	1
350	0.911	0	1	0.853	1	1	1	1
375	0.939	0	1	0.951	1	1	1	1
400	0.944	0	1	0.989	1	1	1	1
425	0.974	0	1	0.997	1	1	1	1
450	0.978	0.005	1	1	1	1	1	1
475	0.990	0.017	1	1	1	1	1	1
500	0.987	0.017	1	1	1	1	1	1

2.6 Appendices

Appendix A. Calculation of the AULC statistic

As described in the method session, for SNP i, we order the p-value vector p_{ij}'s such that $p_{i,(1)} \leq p_{i,(2)} \leq ... \leq p_{i,(N)}$, and then plot -log$_{10}$($p_{i,(j)}$) against -log$_{10}$($j/N$), j = 1,2,...,N. We connect the points in the plot to get a curve and the area under the curve is proportional to the AULC statistic, which will be calculated by summing up the areas of all trapezoids in the plot (Figure 2.2). Here we present the formula for calculating the area under the curve; for simplicity, log$_{10}$ will be represented by log.

$$\begin{aligned}
AULC = \frac{1}{2}\{ & (-\log p_{i,(1)} - \log p_{i,(2)}) \times (-\log \frac{1}{N} + \log \frac{2}{N}) \\
& + (-\log p_{i,(2)} - \log p_{i,(3)}) \times (-\log \frac{2}{N} + \log \frac{3}{N}) \\
& + ... \\
& + (-\log p_{i,(N-1)} - \log p_{i,(N)}) \times (-\log \frac{N-1}{N} + \log \frac{N}{N}) \} \\
= -\frac{1}{2} & \{\log p_{i,(1)} \times \log 2 + \sum_{j=2}^{N-1} (\log p_{i,(j)} \times \log \frac{j+1}{j-1}) + \log p_{i,(N)} \times \log \frac{N}{N-1} \}
\end{aligned}$$

Appendix B. Properties of the AULC statistic

As shown in Appendix A., AULC is a weighted sum of log$_{10}$(p-values). AULC is within the framework of combining functions of p-values. If we assume the proposed statistic has the following form:

$$S = \psi(p_1, p_2, p_3, ..., p_N)$$

where ψ is the combining function. We can prove that the combining function ψ for the AULC statistic satisfies following properties:

(P.1) ψ is non-increasing in each argument: $\psi(...,p_i,...) \geq \psi(...,p_i',...)$ if $p_i < p_i', i \in \{1,...,N\}$ (i.e. the statistic will increase if one of the p-values gets smaller).

(P.2) ψ attains its supremum value $\bar{\psi}$, possibly not finite, even when only one argument attains zero: $\psi(...,p_i,...) \to \bar{\psi}$ if $p_i \to 0, i \in \{1,...,N\}$ (i.e. if one of the p-values reaches 0, the statistic attains its supremum value).

(P.3) $\forall \alpha > 0$, the critical value of every ψ is assumed to be finite and strictly smaller than the supremum value: $S_\alpha < \bar{\psi}$.

(P.4) ψ is a symmetric function, which is invariant with respect to permutation of arguments: for instance, if $(u_1,...,u_N)$ is any permutation of $(1,...,N)$, then $\psi(p_{u_1},...,p_{u_N}) = \psi(p_1,...,p_N)$.

(P.5) When *N*=2, the AULC statistic is equivalent to the Fisher combining function, which is based on the statistic:

$$S_F = -2 \times \sum_j \log(p_j)$$

(P.6) The AULC statistics are equivalent to each other if we use different bases with respect to which logarithms are computed (currently base 10). We can prove that the weights in the summation do not change relatively to each other even though the absolute values of the weights will change.

Property (P.1) is related to the unbiasedness of combined tests; (P.2) and (P.3) are related to consistency. All combining functions ψ are generally required to satisfy at least these three properties [Pesarin 2001].

Appendix C. The procedure for simulating a master regulatory SNP

Prior to the simulation, we do an inverse normal transformation on the expression data such that each gene's expression follows a standard Gaussian distribution. We then simulate a master regulatory SNP by adding the SNP effect onto the expression of those genes that are picked to be associated with the SNP. Specifically, we do the following:

1) The SNP genotype data are simulated by using different allele frequencies p (p=0.5 or p=0.1). The SNP genotypes are coded as 0, 1, and 2 (i.e. the dosage of the minor allele).

2) We randomly pick n (n takes values from 0 to 500) genes whose expression will be associated with the master regulatory SNP.

3) For each of those n genes, we decide the association effect size, which is the proportion of gene expression variation explained by the SNP (R^2). R^2 follows a truncated (from 0 to 1) exponential distribution with mean λ (λ takes values of 0.01, 0.02, 0.05, and 0.1).

4) We add the SNP effect onto the original expression to get the new expression for each of those n genes (assume a linear additive model with the genotype of the master regulator coded as 0, 1 and 2):

$$new\_expression = \beta \times genotype + original\_expression$$

where the SNP effect β is determined by the association effect size R^2:

$$\beta = \sqrt{\frac{R^2}{(1-R^2)(2p(1-p))}}$$

This is because:

$$\begin{aligned} Var(new\_expression) &= Var(\beta \times genotype) + Var(original\_expression) \\ &= \beta^2 \times (2p(1-p)) + 1 \end{aligned}$$

Therefore,

$$
\begin{aligned}
R^2 &= \frac{variance\,explained\,by\,genotype}{total\,variance} \\
&= \frac{\beta^2 \times (2p(1-p))}{\beta^2 \times (2p(1-p))+1}
\end{aligned}
$$

This leads to the formula above to calculate β based on R^2. Note $Var(original\_expression)$ is 1 because we have done an inverse normal transformation on the data such that each gene's expression follows a standard Gaussian distribution. Also, the formula above for calculating R^2 based on β shows that the allele frequency has been taken into account when calculating R^2.

Chapter 3

A method for estimating the overlap of eQTLs between two tissues

3.1 Introduction

Most human eQTL studies have measured transcript abundance in blood cells (peripheral blood lymphocytes and immortalized lymphoblastoid cell lines, LCLs) [Dixon, et al. 2007; Goring, et al. 2007; Morley, et al. 2004; Stranger, et al. 2007]; only a small number of studies have examined it in other tissues (e.g. liver tissue [Schadt, et al. 2008]). Because of this, researchers interested in whether a particular variant is associated with gene expression levels have typically examined this association in LCLs. There is a controversy on whether associations observed in LCLs will translate to other tissues, and recent results reported in a study by Dimas and colleagues [Dimas, et al. 2009] suggest that overlap between eQTL signals among tissues will typically be relatively small. They compared three lists of significant eQTLs identified in three cell types (LCLs, fibroblasts, and T-cells) of 75 individuals and claimed that 69 to 80% of *cis*-eQTLs operate in a cell type-specific manner. However, our analyses provide evidence that current studies with fewer than 100 subjects are typically underpowered and that as a result directly comparing lists of significant eQTLs likely leads to underestimation of the overlap percentage. Indeed, our results show that overlap in eQTLs across tissues can be substantial even when the raw overlap percentage is low.

An accurate estimate of overlap in eQTLs across tissues can allow researchers to quantify the benefits of studying eQTLs in different tissues. Here, we describe a new procedure that takes statistical power into account to provide a more accurate estimate of the percentage of overlapping eQTLs between two tissues. We use this method to compare our eQTL catalog in skin with a previously described catalog for LCLs [Dixon, et al. 2007]. This method has the potential to be applied to a wide range of biological studies.

3.2 Methods

Statistical methods

The simplest approach to compare eQTL lists between two experimental settings is to evaluate the overlap of lists of significant eQTLs compiled separately for each setting. Unfortunately, this method will underestimate the overlap percentage whenever either of the two studies is underpowered (in that case, many true eQTLs might be detected in one study but missing from the list of eQTLs detected in the second study). Here, we propose a new method that takes into account the statistical power of the studies. Very briefly, our procedure starts by splitting the study/tissue with the larger sample size into two parts. One part identifies eQTLs in the tissue, and the second part provides unbiased estimators for the power to replicate eQTL signals. This estimated power is then used to adjust the observed raw overlap percentage and hence to obtain a power-adjusted estimate of the overlap in eQTL signals.

Estimating the overlap percentage. In our method, we assume that eQTL analyses are performed in two studies: in Study 1 (here, the study using lymphoblastoid cell lines), we use a nominal p-value cut-off of α_1 to generate a list of significant eQTLs, which corresponds to a false discovery rate (FDR) of FDR_1, while in Study 2 (here, the study using skin tissue), we use a nominal p-value cut-off of α_2, corresponding to an FDR of FDR_2. Let π be the proportion of true eQTLs in Study 1 that are also true eQTLs in Study 2; let π_{raw} be the observed percentage of significant eQTLs in Study 1 that are also significant in Study 2. Since both eQTL lists are necessarily incomplete, π_{raw} will result in an underestimate of π. Our aim is, thus, to arrive at a better estimator of the true overlap percentage π. To do this, we attempt to estimate a power-adjusted expected overlap in significant eQTLs, $\pi_{adjusted}$.

To arrive at this power-adjusted expected overlap, we start with the list of significant eQTLs in Study 1 and consider (see Figure 3.1 for a detailed decision diagram) a series of possibilities that might lead these eQTLs to replicate in Study 2 (i.e. to be overlapping eQTLs): (a) a fraction (π) of true positive eQTLs in Study 1 are also true eQTLs in Study 2 and are expected to replicate in Study 2 with a particular power; (b) a fraction ($1- \pi$) of true positive eQTLs in Study1 will not be true eQTLs in Study 2 but might "replicate" by chance, with the probability determined by the significance threshold α_2, which is simply the false positive rate; (c) false positive eQTLs in Study 1 might also replicate by chance with probability also determined by significance threshold α_2. We note that it is possible a small fraction of false positive eQTLs in Study 1 will represent true eQTLs in Study 2, but for simplicity, we assume that, since the vast majority of expression level - SNP pairings are not expected to be associated, this event will be extremely rare (see Supplementary Methods for the complete decision diagram with all possibilities and the full description of the method that leads to the simplified version presented here).

Therefore:

$$\pi_{raw} = (1 - FDR_1) \times \pi \times power_2 + (1 - FDR_1) \times (1 - \pi) \times \alpha_2 + FDR_1 \times \alpha_2 \tag{1}$$

where $power_2$ is the statistical power of Study 2 to detect eQTLs that are both true positives in Study 1 and Study 2 (overlapping eQTLs). Algebraic manipulation of equation (1) above gives:

$$\pi = \frac{\pi_{raw} - \alpha_2}{(1 - FDR_1)(power_2 - \alpha_2)}$$

Thus, we can estimate π as:

$$\hat{\pi}_{adjusted} = \frac{\hat{\pi}_{raw} - \alpha_2}{(1 - FDR_1)(power_2 - \alpha_2)} \tag{2}$$

In (2), $\hat{\pi}_{raw}$ is an observed quantity. α_2 is the p-value threshold in Study 2 that is set by researchers. Given α_1, FDR_1 can be estimated using the Benjamini and Hochberg [1995] procedure. Therefore, to estimate π, the major work is to estimate $power_2$.

In theory, $power_2$ is determined by the effect sizes of overlapping eQTLs in Study 2, the sample size of Study 2, and the type I error rate of Study 2 (α_2). Because we are not able to observe all overlapping eQTLs, in our method we make two key assumptions (see below) and use the unbiased estimates of effect sizes of all identified true Study1 eQTLs to approximate the effect sizes of overlapping eQTLs in Study 2. Equivalently, we estimate *power*$_2$ by *power*$_{2raw}$, which we define as the statistical power of a study on Tissue 1 with the same sample size as Study 2 to detect observed significant Study 1 eQTLs when type I error rate is controlled at the same α_2 level. Because observed significant Study 1 eQTLs include both identified true Study 1 eQTLs and false positives, we use the following formula to adjust for false positives when estimating *power*$_2$ (the formula is obtained using a decision tree idea similar to the one in Figure 3.1, see Supplementary Methods for the derivation):

$$po\hat{w}er_2 = \frac{po\hat{w}er_{2raw} - FDR_1 \times \alpha_2}{1 - FDR_1} \qquad (3)$$

A simple estimate of *power*$_{2raw}$ based on the observed effect sizes of each eQTL in Study 1 would be biased because of the winner's curse. To avoid the bias, we estimate *power*$_{2raw}$ using a sample splitting strategy: we split Study1 into mutually exclusive and independent sets Study1A and Study1B. Study 1A is used to identify significant eQTLs in Study 1, while Study 1B is used to provide unbiased estimates for effect sizes of these eQTLs. Given the sample size of Study 2, we can then estimate *power*$_{2raw}$ based on the effect size estimates. The sample splitting strategy can be further simplified if splitting is done such that Study 1B has the same number of subjects and the same data structure (e.g. the same case-control design) as Study 2. In this setting, the percentage of signals identified in Study 1A that are also significant in Study 1B equals to *power*$_{2raw}$. The results presented in this chapter use this simplified sample splitting strategy.

Our approach requires two key assumptions. First, it assumes that the distribution of overlapping eQTL effect sizes in Study 1 is well approximated by the distribution of effect sizes for all identified true eQTLs in Study 1. Second, it assumes that the distribution of overlapping eQTL effect sizes is similar between the two studies. Violation of these assumptions could lead to an underestimate of $power_2$ (for example, if overlapping eQTLs typically have larger effect sizes than the non-overlapping ones), or to an overestimate of $power_2$ (for example, if overlapping eQTLs typically have smaller effect sizes in Study 2 than in Study 1, where they were originally detected). Given these assumptions, we can use this estimated $power_2$ and plug it into equation (2) to obtain an adjusted estimate of the overlap in eQTLs between two tissues.

As an example, consider a setting where 1000 eQTLs are detected with 1% FDR in Study 1. If 80% of the true eQTLs in Study 1 are also true eQTLs in Study 2, and if we set $\alpha_2 = 0.05$ and assume the power of Study 2 is 50%, then we expect to see 406 significant eQTLs in Study 2 (using formula (1)). So $\hat{\pi}_{raw} = 40.6\%$, which is approximately half of the true overlap percentage 80%. However, if we apply equation (2) with $power_2 = 0.5$, $\alpha_2 = 0.05$ and FDR = 0.01, we get $\hat{\pi}_{adjusted} = 80.0\%$.

Estimating the variance of the overlap percentage. We use the jackknife resampling technique to estimate the variance of $\hat{\pi}_{adjusted}$. We randomly remove one subject from Study 1B and one subject from Study 2 to obtain new estimators for $\hat{\pi}_{raw}$ and $power_2$, and hence a new estimator for $\hat{\pi}_{adjusted}$. We repeat this procedure and obtain multiple estimators for $\hat{\pi}_{adjusted}$ and then estimate the variance of $\hat{\pi}_{adjusted}$ as:

$$\widehat{\mathrm{var}}(\hat{\pi}_{adjusted}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\pi}_{adjusted,i} - \bar{\hat{\pi}}_{adjusted}\right)^2$$

Data for skin. Genotype and expression data for skin will be described in detail in Chapter 5. Very briefly, we collected gene expression data using Affymetrix U133 Plus 2.0 arrays from three types of skin tissues: normal skin from 57 healthy controls (normal skin), non-diseased (uninvolved skin) and diseased skin (lesional skin) from 53 psoriatic patients. All 110 subjects were genotyped by Perlegen Sciences using four proprietary, high-density oligonucleotide arrays in partnership with the Genetic Association Information Network (GAIN). We then used MACH [Li, et al. 2010; Li, et al. 2009] to impute all HapMap SNPs, with phased HapMap CEU sample haplotypes as templates. We analyzed only SNPs that were genotyped or could be imputed with relatively high confidence (estimated r^2 between imputed SNP and true genotypes >0.3, so that patterns of haplotype sharing between sampled individuals and HapMap samples consistently indicated a specific allele; we use this r^2 based threshold, rather than one based on the posterior probability of each imputed genotype, because it naturally accommodates SNPs with different allele frequencies and because it is the same threshold used in many recent genomewide association studies, including for psoriasis [Nair, et al. 2009]).

Data for lymphoblastoid cell lines (LCLs). Genotype and expression data for LCLs were originally published in Dixon et al. [2007] for a set of 183 families (340 subjects total).

Affymetrix U133 Plus 2.0 arrays were used for gene expression profiling, and Sentrix HumanHap300 Genotyping BeadChips (Illumina, San Diego, CA) were used for SNP genotyping. As above, we then used MACH to impute all HapMap SNPs, and analyzed only SNPs that were genotyped or could be imputed with relatively high confidence. We split the 183 families randomly into two sets, where Set 1A contained 126 families and Set 1B contained one randomly selected individuals from each of the remaining 57 families (so that the sample size equals to that of the study on normal skin).

Applying the method to estimating the overlap of cis-eQTL signals in LCLs and in skin. We first performed eQTL analysis in Study 1A. We tested only associations between each transcript and those SNPs that are in the *cis*-candidate region (within 1Mb) of the transcript. We used a range of nominal p-value thresholds, which corresponded to an FDR (FDR_1) of 0.001, 0.0005 and 0.0001 for testing *cis*-eQTLs. To avoid multiple counting of the same *cis*-eQTL signal, for each transcript that had at least one significant *cis*-eQTL association we picked the most significant SNP to form the SNP-transcript pair which would be evaluated in Study 1B and Study 2. For the study in skin (Study 2), we focused our analysis on the data from the 57 normal controls (the data from 53 patients were also analyzed). We also used a range of α_2 values (p-value thresholds) in Study1B and Study2: 0.05, 0.001, and 0.0005.

3.3 Results

Using our new method, we estimated the percentage of true eQTLs in LCLs that are also true eQTLs in normal skin. As our method requires an approximation in the formula, we controlled the false discovery rate in Study 1 relatively tightly (i.e. controlling FDR_1 at 0.001, 0.0005 and 0.0001). We allowed α_2 to take a range of different values (0.05, 0.001 and 0.0005), and then estimated the overlap percentage for all combinations of FDR_1 and α_2. As summarized in Table 3.1, the different FDR_1 and α_2 thresholds give relatively consistent estimates for the percentage of overlapping eQTLs between tissues: around 70% of the true cis-eQTLs in LCLs are estimated to be present in normal skin. The naïve estimator $\hat{\pi}_{raw}$ suggests overlap percentages ranging from 30% to 50% depending on the statistical thresholds used in the analysis. As an example, if we set $FDR_1 = 0.0005$ and $\alpha_2 = 0.001$, the observed overlap percentage ($\hat{\pi}_{raw}$) was 0.316 and power was estimated at $power_2 = 0.462$. Using formula (2), we estimated the true overlap percentage to be 68.3% (95% confidence interval from the jackknife resampling 66.4% to 70.2%). We also estimated the overlap of *cis*-eQTLs between LCLs and uninvolved skin, as well as between LCLs and lesional skin. These additional comparisons produced similar estimates that ~70% *cis*-eQTLs in LCLs are also present in uninvolved and lesional skin (Table 3.1). These results suggest that a majority of *cis*-eQTLs are shared between skin and LCLs.

We also compared the LCL *cis*-eQTLs in our analysis with *cis*-eQTLs identified in fibroblasts and T cells generated by Dimas et al. [2009]. Even though the raw overlap percentages were rather low, after we adjusted for the power of the study (because sample sizes of LCLs, fibroblasts, and T-cells are the same in Dimas et al. [2009], the power of the study can be estimated by using the results from LCLs in that same study), we estimated that 65-70% of significant LCL *cis*-eQTLs were also present in fibroblasts and T-cells (Table 3.2 and Supplementary Methods). This finding is very consistent with the results from comparing LCLs and skin.

3.4 Discussion

We have developed a new method for estimating eQTL overlap between two tissues. Our method can provide a more accurate estimator for the overlap percentage whenever either of the two studies is underpowered and hence is likely to fail to detect all the significant signals, leading to undetected overlapping signals. Our procedure starts by splitting the study/tissue with the larger sample size into two parts. One part identifies eQTLs in the tissue, and the second part provides unbiased estimators for the eQTL effect sizes. The power of the second study/tissue to detect overlapping eQTLs is then estimated, and is used to adjust the observed overlap percentage to derive the power-adjusted overlap percentage. Our method can be useful in a variety of settings where estimation of the overlap of two signal lists is needed. For example, in theory the method can be applied to estimate the overlap of areas of the brain activated in two settings (e.g. two different stimuli) in an fMRI (functional magnetic resonance imaging) experiment.

Using our method, we have estimated that around 70% of the significant *cis*-eQTLs in LCLs are also observed in skin, a value that greatly exceeds the raw overlap percentage of 30-50% obtained using a naïve estimator. It may be argued, however, that because the overlapping *cis*-eQTLs may have different effect sizes in the two tissues, they may still be regarded as tissue specific.

3.5 Figures and tables

Figure 3.1 Simplified diagram for categorization of significant eQTLs from Study1 into groups for the estimation of the overlap percentage (a full version of decision diagram can be seen in the Supplementary Methods)

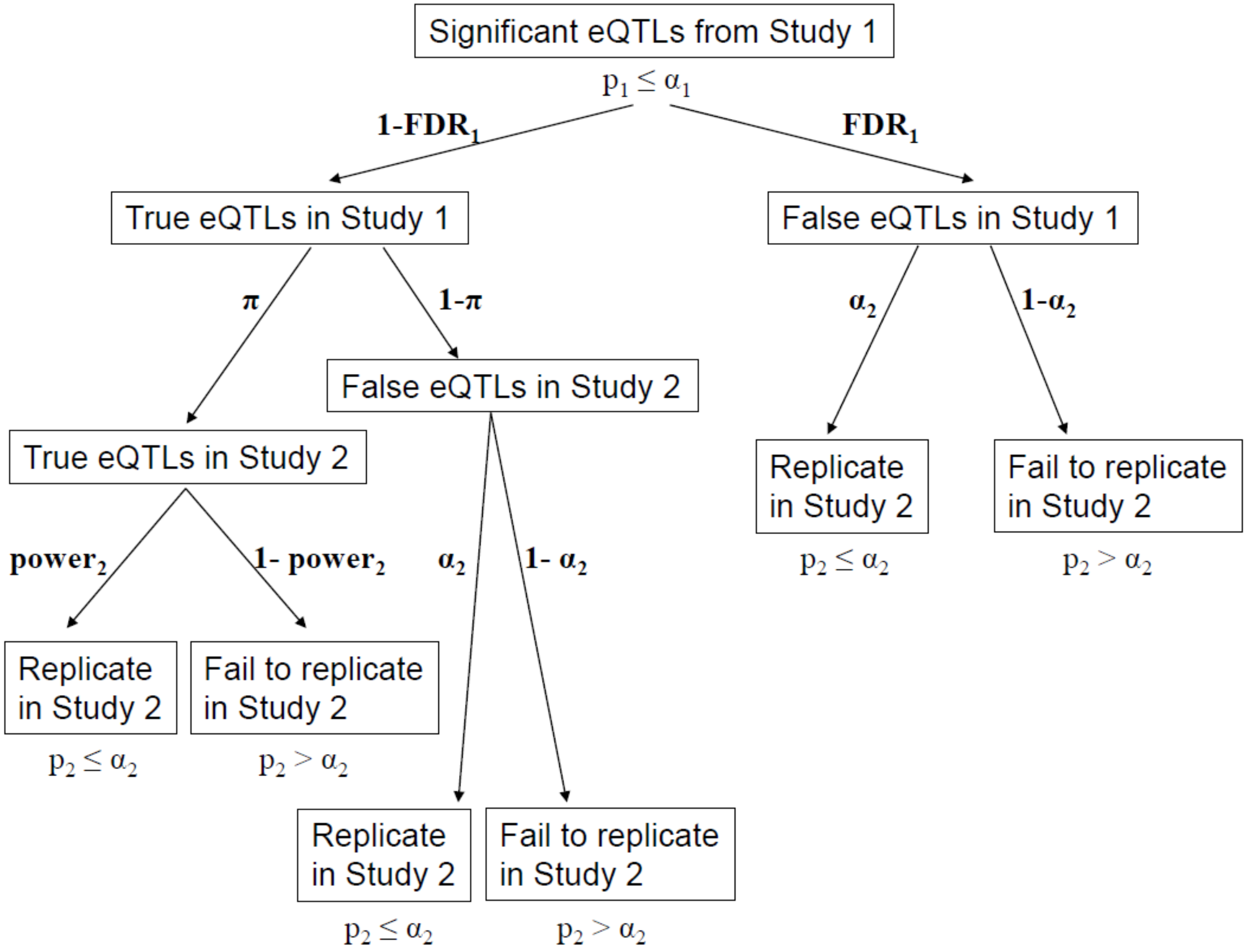


Table 3.1 Estimating the overlap of cis-eQTLs between LCLs and the three types of skin using different significance thresholds.

Different thresholds		57 normal skin			53 uninvolved skin			53 lesional skin		
FDR_1	α_2	$\hat{\pi}_{raw}$	$power_2$	$\hat{\pi}_{adjusted}$ (95% CI)	$\hat{\pi}_{raw}$	$power_2$	$\hat{\pi}_{adjusted}$ (95% CI)	$\hat{\pi}_{raw}$	$power_2$	$\hat{\pi}_{adjusted}$ (95% CI)
0.001	0.05	0.520	0.771	0.652 (0.642, 0.662)	0.495	0.764	0.623 (0.612, 0.634)	0.499	0.764	0.629 (0.618, 0.640)
0.001	0.001	0.296	0.429	0.689 (0.669, 0.709)	0.278	0.402	0.691 (0.671, 0.711)	0.266	0.402	0.661 (0.641, 0.681)
0.001	0.0005	0.264	0.367	0.719 (0.692, 0.746)	0.252	0.345	0.730 (0.703, 0.757)	0.239	0.345	0.692 (0.667, 0.717)
0.0005	0.05	0.538	0.805	0.646 (0.636, 0.656)	0.514	0.798	0.620 (0.609, 0.631)	0.519	0.798	0.627 (0.616, 0.638)
0.0005	0.001	0.316	0.462	0.683 (0.664, 0.702)	0.297	0.434	0.684 (0.665, 0.703)	0.285	0.434	0.656 (0.636, 0.676)
0.0005	0.0005	0.284	0.398	0.713 (0.686, 0.740)	0.271	0.374	0.724 (0.698, 0.750)	0.256	0.374	0.684 (0.660, 0.708)
0.0001	0.05	0.587	0.856	0.666 (0.656, 0.676)	0.564	0.849	0.643 (0.632, 0.654)	0.552	0.849	0.628 (0.617, 0.639)
0.0001	0.001	0.364	0.534	0.681 (0.663, 0.699)	0.342	0.506	0.675 (0.657, 0.693)	0.332	0.506	0.655 (0.637, 0.673)
0.0001	0.0005	0.330	0.471	0.700 (0.675, 0.725)	0.315	0.442	0.712 (0.688, 0.736)	0.301	0.442	0.681 (0.659, 0.703)

Table 3.2 eQTL Overlap between LCLs from Dixon et al. and LCLs, fibroblasts, and T-cells from Dimas et al.

	Overlap with LCLs (Dixon et al.)					
Dimas et al.	**α_2 = 0.01**		**α_2 = 0.005**		**α_2 = 0.001**	
	$\hat{\pi}_{raw}$	$\hat{\pi}_{adjusted}$	$\hat{\pi}_{raw}$	$\hat{\pi}_{adjusted}$	$\hat{\pi}_{raw}$	$\hat{\pi}_{adjusted}$
FDR₁ **= 0.001**						
LCLs	0.368	1.000*	0.312	1.000*	0.241	1.000*
Fibroblasts	0.237	0.633	0.202	0.643	0.151	0.624
T-cells	0.257	0.690	0.230	0.732	0.167	0.691
FDR₁ **= 0.005**						
LCLs	0.326	1.000*	0.272	1.000*	0.208	1.000*
Fibroblasts	0.210	0.632	0.177	0.644	0.131	0.630
T-cells	0.235	0.712	0.204	0.744	0.143	0.687

*Assumed

3.6 Supplementary materials

Supplementary Methods

A. The detailed method for estimating the overlap percentage (i.e. deriving formula (2) in the main text)

In our method, we assume that eQTL analyses are performed in two studies: in Study 1 (here, the study using lymphoblastoid cell lines), we use a nominal p-value cut-off of α_1 to generate a list of significant eQTLs, which corresponds to a false discovery rate (FDR) of FDR_1, while in Study 2 (here, the study using skin tissues), we use a nominal p-value cut-off of α_2, corresponding to an FDR of FDR_2. Let π be the percentage of eQTLs in Study 1 that are also eQTLsin Study 2; let π_{raw} be the observed percentage of significant eQTLs in Study 1 that are also significant in Study 2. Since both eQTL lists are necessarily incomplete, π_{raw} will result in an underestimate of π. Our aim is, thus, to arrive at a better estimator of the true overlap percentage π. To do this, we attempt to estimate a power-adjusted expected overlap in significant eQTLs, $\pi_{adjusted}$.

Our method starts with a list of significant eQTLs in Study 1 and dissects those significant eQTLs in three steps into 8 mutually exclusive groups, depending on whether they are true/false positives in Study 1, true/false positives in Study 2, and designated significant/non-significant eQTLs in Study 2 (The detailed dissection diagram is shown in Supplementary Fig 1A). Besides the abovementioned parameters, ρ is the percentage of false positive eQTLs in Study 1 that are true positives in Study 2; $power_2$ is the statistical power of Study 2 to detect eQTLs that are both true positives in Study 1 and Study 2 (overlapped eQTLs); *power₂'* is the statistical power of Study 2 to detect eQTLs that are true positives in Study 2 but false positives in Study 1. Among the 8 groups of eQTLs that are identified as significant in Study1, 4 groups will be identified as significant in Study 2 (observed overlaps). Therefore:

$$\pi_{raw} = (1 - FDR_1) \times \left[\pi \times power_2 + (1 - \pi) \times \alpha_2 \right] + FDR_1 \times \left[\rho \times power_2' + (1 - \rho) \times \alpha_2 \right] \quad (0)$$

If we can control the false discovery rate of Study 1 (*FDR₁*) well, we can expect *FDR₁* to be much less than 1- *FDR₁*. Therefore, the contribution of the second term in (0) is much less than that of the first term and hence the simplification of the second term below should have little impact on formula (0). It is reasonable to assume that *ρ* « 1- *ρ*, and then we can assume (1- *ρ*) ×*α₂* + *ρ* × *power₂'* ≈ (1-0) × *α₂* + 0 × *power₂'* = *α₂*. Hence, formula (0) can be simplified as formula (1) in the main text:

$$\pi_{raw} = (1 - FDR_1) \times \pi \times power_2 + (1 - FDR_1) \times (1 - \pi) \times \alpha_2 + FDR_1 \times \alpha_2 \quad (1)$$

This simplification essentially assumes that the probability that Study 2 will identify the false positives in Study 1 as significant signals is *α₂* (the probability of being identified by chance; see Fig 1 for the corresponding simplified diagram).

Based on (1), we can estimate *π* as:

$$\hat{\pi}_{adjusted} = \frac{\hat{\pi}_{raw} - \alpha_2}{(1 - FDR_1)(power_2 - \alpha_2)} \quad (2)$$

B. The detailed method for estimating *power₂* using *power₂raw* (i.e. deriving formula (3) in the main text)

With the two assumptions made in the main text, *power*$_2$ equals to *power*$_{2appro}$, which is defined as the statistical power of a study on Tissue 1 with the same sample size as Study 2 to detect all identified true Study 1 eQTLs when controlling type I error rate at α_2. Meanwhile, *power*$_{2raw}$ is defined as the statistical power of a study on Tissue 1 with the same sample size as Study 2 to detect all identified Study 1 eQTLs. The difference between the two lies in the fact that the list of "all identified Study 1 eQTLs" includes false positives while the list of "all identified true Study 1 eQTLs" does not. In the main text, Study 1A can be regarded as the original Study1, and Study 1B can be regarded as the study with the same sample size as Study2. Using a similar decision tree idea, we can categorize the list of significant eQTLs identified in Study 1A into 4 mutually exclusive groups (Supplementary Fig 1B), among which 2 groups will be identified as replicated in Study 1B (the proportion replicated in Study 1B equals to *power*$_{2raw}$). Therefore:

$$power_{2raw} = (1 - FDR_1) \times power_{2appro} + FDR_1 \times \alpha_2$$

Algebraic manipulation of the above equation gives:

$$po\hat{w}er_{2appro} = \frac{po\hat{w}er_{2raw} - FDR_1 \times \alpha_2}{1 - FDR_1}$$

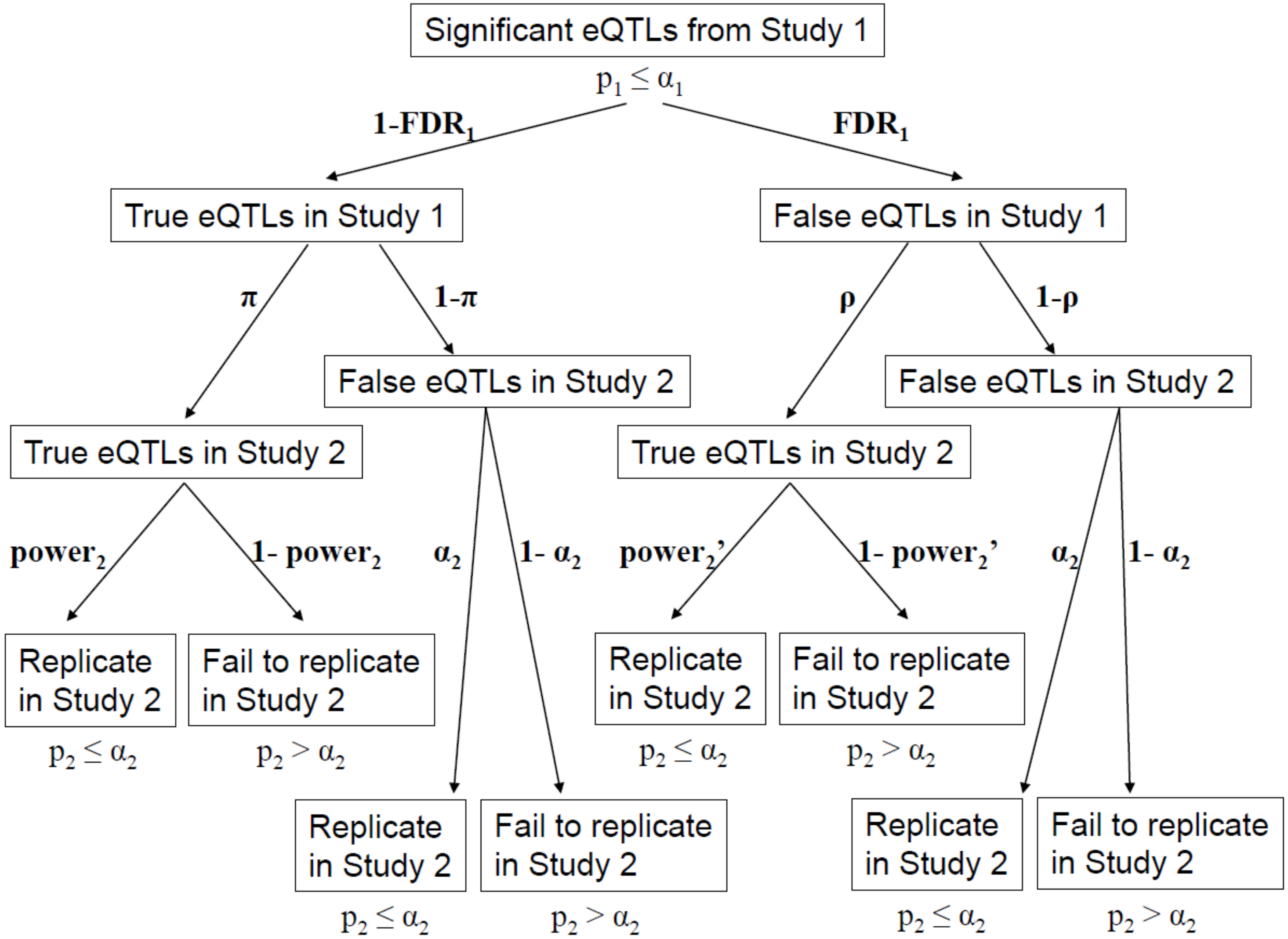
Therefore:

$$po\hat{w}er_2 = \frac{po\hat{w}er_{2raw} - FDR_1 \times \alpha_2}{1 - FDR_1} \qquad (3)$$

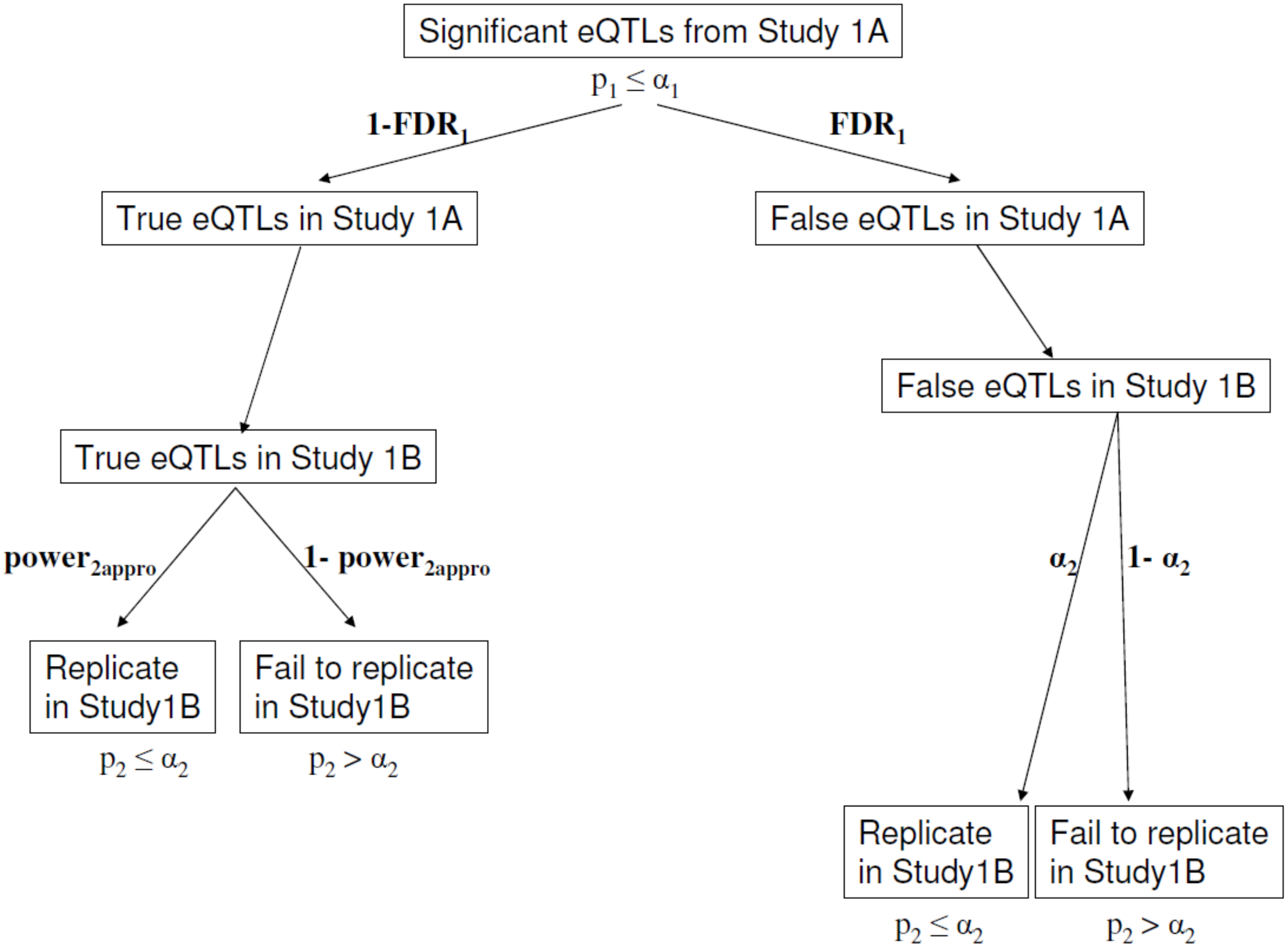
C. Estimating *cis*-eQTL overlap between LCLs from Dixon et al. [2007] and fibroblasts and T-cells from Dimas et al. [2009]

Antigone Dimas and Emmanouil Dermitzakis kindly provided us with lists of all significant SNP-gene expression pairs in LCLs, fibroblasts, and T-cells that have nominal p-values <= 0.01 from Dimas et al. [2009]. We use these lists of significant signals, together with the list of eQTLs identified in LCLs from Dixon et al. [2007], to estimate the eQTL overlap between LCLs from Dixon et al. [2007] and fibroblasts and T-cells from Dimas et al. [2009], respectively. For example, to estimate the eQTL overlap between LCLs from Dixon et al. [2007] and fibroblasts from Dimas et al. [2009], we treat the study on LCLs from Dixon et al. [2007] as Study 1A and the study on fibroblasts from Dimas et al. [2009] as Study 2. We first compare the two lists and get the raw observed overlap percentage ($\hat{\pi}_{raw}$). Because the sample sizes for LCLs, fibroblasts, and T-cells are the same in Dimas et al. [2009], the study on LCLs in Dimas et al. [2009] can be regarded as the Study1B in our method, and therefore, the raw overlap percentage between LCLs from Dixon et al. [2007] (Study 1A) and LCLs from Dimas et al. [2009] (Study 1B) is exactly $po\hat{w}er_{2raw}$, as defined. Then we use the formulas above to estimate the power-adjusted overlap percentage. More specifically, because the study of Dimas et al. [2009] used Illumina 550K SNP array, we also consider only SNPs on the Illumina 550K SNP array in the study of LCLs from Dixon et al. [2007] to make the lists of SNP considered comparable. Meanwhile, because the two studies used different platforms of gene expression arrays, we map the gene expression probe IDs to Entrez gene IDs to make expression traits comparable. We then compile a list of significant SNP-gene expression pairs (for expression traits associated with >1 *cis*-SNP, the most significant *cis*-SNP-expression pair will be counted as one independent signal) for LCLs from Dixon et al. [2007] This list of *cis*-eQTL signals are compared with eQTL lists from Dimas et al. [2009] to estimate the raw observed overlap percentages.

Supplementary Fig 1A Categorization of significant eQTLs from Study 1 into groups for the estimation of overlap percentage



Supplementary Fig 1B Categorization of significant eQTLs from Study 1A into groups for the estimation of $power_{2appro}$ using $power_{2raw}$



Chapter 4

Simulation studies on the method for estimating the eQTL overlap

4.1 Introduction

To quantify the benefits of studying eQTLs in different tissues, we propose a novel method for accurately estimating the overlap of eQTLs between two tissues in Chapter 3. Briefly, our procedure starts by splitting the study/tissue with the larger sample size (Study 1) into two parts (Studies 1A and 1B). One part (Study 1A) identifies eQTLs in the tissue, and the second part (Study 1B) provides unbiased estimators for the eQTL effect sizes. The power of the second study/tissue (Study 2) to detect overlapping eQTLs is then estimated, and is used to adjust the observed overlap percentage to derive the power-adjusted overlap percentage.

In theory, there are multiple ways to split samples in Study 1 into Studies 1A and 1B. The simplified version of the sample splitting strategy employed in Chapter 3 requires that splitting is done such that Study 1B has the same number of subjects and the same data structure (i.e. the same pattern of related and unrelated individuals) as Study 2. In this scenario, the power of the second study (more specifically, $power_{2raw}$) can be simply estimated as the proportion of signals identified in Study 1A that are also significant in

Study 1B. However, one of the limitations for this simplified version of the method is the lack of flexibility. In some situations, the splitting strategy will not even work. For example, when the sample sizes of Study 1 and Study 2 are the same, no splitting can be done that satisfies the requirement of the simplified method.

The full version of the splitting strategy can overcome the limitation of fixed sample size for Study 1B. Study 1B with an arbitrary sample size is used to provide unbiased estimates for effect sizes of the eQTLs; the power of Study 2 is then estimated based on the effect size estimates and the sample size of Study 2. Out of all plausible ways of splitting samples in Study 1, it is not clear which way is the best with regard to the estimation accuracy. Apparently, there are two competing interests here. With a fixed number of samples in Study 1, if we put more samples in Study 1A, Study 1A will have larger statistical power to detect significant eQTLs in Study1, which means a better estimation for the observed overlap percentage. On the other hand, if we put more samples in Study 1B, we can get better estimates for the eQTL effect sizes, which means a better estimation for the power to detect overlapping eQTLs. A good sample splitting strategy should be able to balance these two competing interests. In a broad sense, this is similar to the problem encountered by many researchers in genetic studies where reliably detecting signals and accurately estimating effect sizes of those signals are both research goals [Sun and Bull 2005]. People have proposed the sample splitting strategy [Goring, et al. 2001] to achieve the two goals at the same time. But typically, it is not clear how to identify the best way to split the samples such that both goals can be approached in an acceptable manner. We plan to address this question here under the framework of estimating eQTL overlap using simulation studies. We propose to use the parameter estimation accuracy as the criterion to compare the performance of different sample splitting strategies.

In this chapter, we plan to achieve two goals using the same simulation framework. First, for the simplified version of the method, we want to assess its performance in the perfect scenario where both assumptions for effect sizes hold or in scenarios where at least one of the two assumptions is violated. The two assumptions for effect sizes are: Assumption 1) the distribution of overlapping eQTL effect sizes in Study 1 is well approximated by the distribution of effect sizes for all identified true eQTLs in Study 1; Assumption 2) the distribution of overlapping eQTL effect sizes is the same between two studies. We want to assess how good the overlap percentage estimation is when both assumptions hold and how much the overlap percentage estimation is affected when the two assumptions are violated to different extents.

Second, for the full version of the method, we want to use the simulation framework to determine the ideal way to split samples in Study 1 that can achieve highest estimation accuracy.

4.2 Methods

Simulation framework

In our simulation, we test *cis*-associations between gene transcripts and their *cis*-SNPs in the genome. We assume each gene transcript has either zero or one associated *cis*-SNP.

We use a simple linear model to describe the true association between the *cis*-SNP genotype (coded as 0, 1, 2) and the corresponding gene's expression level. The effect size of each *cis*-association is determined by a quantity R^2 (the coefficient of determination), which corresponds to the proportion of transcript expression variation explained by the genotype. For all the true *cis*-associations, we let R^2 follow a truncated (from 0 to 1) exponential distribution with mean λ. As the effect size parameter, λ corresponds to a specific statistical power of a study with a certain sample size to detect *cis*-eQTLs when controlling a certain Type I error rate. We set λ to different values in our simulation that correspond to a range of study power.

We assume that 20,000 gene transcripts are tested on the whole genome. In both Study 1 and Study 2, 3,000 gene transcripts have true *cis*-associations with eQTL SNPs. Among the two lists of 3,000 *cis*-eQTLs, a proportion of them are shared in the two studies, which is determined by the overlap percentage parameter π. We let π take three values in our simulation. One value is $\frac{2}{3}$ (i.e. 2,000 eQTLs are shared in the two studies), which reflects the overlap percentage estimated from empirical data (see Chapter 3). We also let π take the values 0.05 (i.e. 150 shared eQTLs) and 0.95 (i.e. 2,850 shared eQTLs) so that we can test the performance of the method when π is at the limits of the parameter space.

Assess the performance of the simplified method under different scenarios

We consider three scenarios in the simulation: 1) "perfect scenario": two effect size assumptions hold and the overlapping eQTLs have same effect sizes in the two studies/tissues; 2) "violation in Assumption 1": Assumption 1 is violated (at different levels) such that overlapping eQTLs have larger mean effect size than non-overlapping eQTLs in Study 1; 3) "violation in Assumption 2": Assumption 2 is violated (at different levels) such that all overlapping eQTLs have larger effect sizes in Study 1 than in Study 2.

We perform the simulation using the following different parameter settings:

a) sample size *n*: *2n* unrelated subjects in Study 1 and *n* unrelated subjects in Study 2 such that Study 1A, 1B, and 2 all have *n* samples. *n* takes the values 60 (i.e. 120 for Study 1 and 60 for Study 2, a sample size similar to currently published studies), or 120.

b) *cis*-eQTL mean effect size (λ) takes the values 0.10, 0.16, or 0.32, which correspond to the average statistical power to detect signals at ~10%, ~25%, and ~40%, respectively, for a study with a sample size of $n = 60$ at the false discovery rate (FDR) of 0.001, or at ~30%, ~40%, and ~65%, respectively, for a study with a sample size of $n = 120$ at FDR of 0.001.

c) true overlap percentage (π) takes the values $\frac{2}{3}$, 0.05, or 0.95.

d) "violation in Assumption 1": in Study 1, overlapping eQTLs' mean effect size (λ) is 10%, 20%, or 30% larger than the non-overlapping eQTLs' mean effect size.

e) "violation in Assumption 2": each overlapping eQTL has its effect size (λ) 10%, 20%, or 30% larger in Study 1 than in Study 2.

We simulate 100 data sets for each simulation scenario. When applying our method to the simulated data set, we use 16 different threshold combinations: the false discovery rate controlled in Study 1A (*FDR₁*) is set at 0.01, 0.008, 0.005, or 0.001; the Type I error rate controlled in Study 2 (α_2) is set at 0.05, 0.01, 0.001, or 0.0005. We then assess the performance of the method under different scenarios by looking at the bias, the standard error (*SE*), and the square root of the mean squared error ($\sqrt{MSE}$) of $\hat{\pi}_{adjusted}$. For each simulation scenario, the three summary statistics are estimated as the mean among the 16 different threshold combinations.

Identify the ideal sample-splitting strategy for the full-version method

We split samples in Study 1 into Study 1A and Study 1B with different splitting proportions. We use the mean squared error (*MSE*) for the overlap estimator as the criterion to assess the performance of different ways of splitting samples.

Similarly as above, we perform the simulation using the following different parameter settings:

a) sample size (*n*) takes the values 60, 120, or 240.

b) *cis*-eQTL mean effect size (λ) takes the values 0.10, 0.16, or 0.32.

c) true overlap percentage (π) takes the value $\frac{2}{3}$.

d) Study 1 sample splitting proportion (the proportion of Study 1 samples used in Study 1A) takes values from 5% to 95% in increments of 5%.

We use samples in Study 1B to estimate the effect size of each significant eQTL (R_i^2, *i* = 1 , …, *M*, *M* is the number of significant eQTLs) identified in Study 1A. For each eQTL, under the alternative hypothesis (i.e. the eQTL is a true eQTL), the test statistic (F statistic) follows a non-central F distribution with 1 and *n*-2 (*n* is the sample size in Study 2) degrees of freedom and a noncentrality parameter $\frac{R_i^2 \times n}{(1-R_i^2)}$. When controlling Type I error rate at α_2, we can then estimate the power of Study 2 to detect each individual eQTL based on the above F distribution. The average power of Study 2 to detect significant eQTLs identified in Study 1A is the mean of all individual power estimates.

We then estimate the mean *MSE* among *MSE*s from 16 different threshold combinations (4 FDR_1 thresholds and 4 α_2 thresholds) for each simulation scenario.

4.3 Results

Assess the performance of the simplified method under different scenarios

The first simulation scenario that we consider is a "perfect scenario", where two effect size assumptions both hold. When the sample size *n* = 60 and the true overlap percentage $\pi = \frac{2}{3}$, the mean $\hat{\pi}_{adjusted}$ is 0.6658, 0.6661, and 0.6663 when the mean effect size parameter λ takes the values 0.10, 0.16, and 0.32, respectively (Table 4.1). Therefore, the bias is -0.0009, -0.0005, and -0.0003, respectively. Furthermore, the mean *SE* is 0.032, 0.020, and 0.010; the mean $\sqrt{MSE}$ is 0.032, 0.020, and 0.010. Given the same sample size, when the mean effect size (λ) increases (i.e. the power of the study increases) the bias and $\sqrt{MSE}$ both decrease. We observe this same trend when sample size *n* increases to 120 (Table 4.1). We also obtain similar estimation accuracy when the true overlap percentage π takes the more extreme values 0.05 and 0.95 (Tables 4.2 and 4.3). For example, when the sample size n = 60 and the true overlap percentage π = 0.05, the mean $\hat{\pi}_{adjusted}$ is 0.0520, 0.0513, and 0.0506 when the mean effect size parameter λ takes the values 0.10, 0.16, and 0.32, respectively (Table 4.2). All these simulations show that under "perfect scenario" our method performs very well in estimating the overlap percentage.

We then check the performance of our method when the two effect size assumptions are violated. In the scenario of "violation in Assumption 1", overlapping eQTLs' mean effect size (λ) is larger than the non-overlapping eQTLs' mean effect size in Study 1. As expected, this leads to the over-estimation of the overlap percentage (as a result of the under-estimation of the power). Table 4.4 shows the estimation results for n = 60 and λ = 0.32. For $\pi = \frac{2}{3}$, the bias is 0.015 (2.2%), 0.030 (4.5%), and 0.044 (6.6%) when the overlapping eQTLs' mean effect size is 10%, 20%, and 30% larger than that of the non-overlapping eQTLs in Study 1, respectively. Meanwhile, the mean *SE* is 0.011, 0.011, and 0.011; the mean $\sqrt{MSE}$ is 0.018, 0.032, and 0.045. The mean *SE*s are very similar with each other at the three different violation levels, and they are also similar with the mean *SE* estimated when the assumption is not violated at all (see Table 4.1). This indicates that the increase in the mean $\sqrt{MSE}$ is majorly due to the increase of estimation bias when the Assumption 1 is violated. We have similar observations when π = 0.05 or 0.95 (Table 4.4): even when the overlapping eQTLs' mean effect size is 30%

larger than that of the non-overlapping eQTLs in Study 1, the difference between the estimator and the true value is considerably smaller.

We then do simulations under the scenario of “violation in Assumption 2”: all overlapping eQTLs have larger effect sizes in Study 1 than in Study 2, which will lead to the under-estimation of the overlap percentage (as a result of the over-estimation of the power). Table 4.5 shows the estimation results for n = 60 and λ = 0.32. For $\pi = \frac{2}{3}$, the bias is -0.015 (-2.3%), -0.031 (-4.6%), and -0.047 (-7.1%) when each overlapping eQTL has larger effect size in Study 1 than in Study 2 by 10%, 20%, and 30%, respectively. Meanwhile, the mean SE is 0.010, 0.010, and 0.011; the mean $\sqrt{MSE}$ is 0.019, 0.033, and 0.049. Again, the increase in mean $\sqrt{MSE}$ is majorly due to the increase of the estimation bias when the Assumption 2 is violated. When π takes the extreme values 0.05 or 0.95, we have similar observation that the bias of the estimator is considerably smaller comparing with the level of violation in the assumption.

All above simulations show that our method performs reasonably well in estimating the overlap percentage when the two assumptions both hold or when assumptions are moderately violated.

Identify the ideal sample-splitting strategy for the full-version method

In our simulation, we split samples in Study 1 into Study 1A and Study 1B with different splitting proportions: Study 1A takes a proportion of Study 1 samples ranging from 5% to 95% while Study 1B takes the remaining samples. We then use the full version of our method to estimate the overlap percentage. *MSE* is used to assess the performance of the method and the sample splitting proportion that provides minimum *MSE* is regarded as the ideal sample splitting strategy.

Figure 4.1 shows the *MSE* statistic across different sample splitting proportions in three mean effect size settings for sample size n = 60. Specifically, when $E(R^2) = 0.10$, 0.16, and 0.32, our method achieve minimum *MSE* when Study 1A takes 55%, 45%, and 35% of the samples in Study 1, respectively. As expected, when the mean effect size increases (i.e. the power of a study to identify eQTLs increases), the ideal sample proportion for Study 1A decreases, as less samples are needed to identify signals in Study 1A. Figure 4.2 and Figure 4.3 show the same plots for sample size n = 120 and n = 240, respectively; Table 4.6 summarizes the ideal splitting proportion for each of the nine settings considered here. Even though the exact ideal splitting proportion for achieving minimum *MSE* is different in different simulation settings, the curves always have a plateau at the bottom, indicating that there is a window for picking the splitting proportion where the *MSE* can be kept relatively low (e.g. less than 1.5 times the minimum *MSE*). For example, in Figure 4.1 middle panel, the estimation will have *MSE*s that are less than 1.5 times the minimum *MSE* as long as the splitting proportion is from 25% to 65%. In almost all the current simulation settings, the estimation can achieve relatively low *MSE* if the splitting proportion ranges from 25% to 40% (Table 4.6). We discuss the probable ideal splitting proportion under other settings in the “Discussion” section.

4.4 Discussion

Our first goal is to use simulation to assess the performance of our method under different scenarios. If the two assumptions of our method hold, our method provides accurate overlap percentage estimation. It is reassuring to observe that even when the two assumptions are moderately violated, the estimation is still reasonably good. For example, in one simulation setting where the true overlap percentage is $\frac{2}{3}$, when the mean effect size of the overlapping-eQTLs in Study 1 is 30% larger than that of the non-overlapping eQTLs, the mean of the estimator from our method is 0.711 (a difference of 0.044 from the true value), and the $\sqrt{MSE}$ is 0.045. When the true overlap percentage is 0.05 and 0.95, the bias is even smaller. Likewise, when the effect size of each overlapping eQTLs is 30% larger in Study 1 than in Study 2, the mean of the estimator from our method is 0.620 (a difference of -0.047) when the true overlap percentage is $\frac{2}{3}$. Clearly, moderate violations of the two assumptions do not lead to severe over-estimation or under-estimation of the overlap percentage.

Our second goal is to identify the ideal sample-splitting strategy for the full-version method. Apparently, the ideal sample-splitting proportion (with smallest mean squared error) depends on many different parameters, including the sample size of each study, the effect sizes of the eQTLs, and the true overlap percentage. Different parameter settings can lead to completely different ideal splitting proportion. Consider two extreme cases: in Case 1, the true overlap percentage is very small (i.e. close to 0) and the effect sizes of the eQTLs are also very small such that Study 1A has to have a large number of samples to be able to detect a reasonable number of overlapping eQTLs, or equivalently, to estimate the observed overlap percentage reasonably well. In this case, the proportion of samples for Study 1A should be large. Case 2 is just the opposite, where the true overlap percentage is very large (i.e. close to 1) and the effect sizes of the eQTLs are also very large such that Study 1A with only a small number of samples will be able to estimate the observed overlap percentage reasonably well. Therefore, a large proportion of samples from Study 1 can be used in Study 1B to obtain a good estimation for the power of the study. In this case, the proportion of samples for Study 1A should be small. Our simulation results support this reasoning: when the study sample size increases or when the eQTL effect sizes increase, the ideal sample proportion for Study 1A will generally decrease. Meanwhile, we observe from simulations that there is usually a set of continuous values for the splitting proportion that gives similarly good estimation of the overlap percentage with regard to the mean squared error (i.e. $MSE \leq 1.5 \times$ minimum MSE). In fact, when the true overlap percentage is chosen to be close to the empirical value observed from at least two studies (i.e. $\pi = \frac{2}{3}$), our simulation shows that for a range of study sample sizes and a range of eQTL effect sizes, the estimation can always be reasonably good if the splitting proportion of samples for Study 1A is 40%. In eight of the nine settings, the estimation can be reasonably good if the splitting proportion is from 25% to 40%. This can be used as guidance when samples in Study 1 are split for identifying eQTLs and estimating eQTL effect sizes.

4.5 Figures and Tables

Figure 4.1 *MSE*s for the overlap percentage estimator across different sample splitting proportions for sample size 60

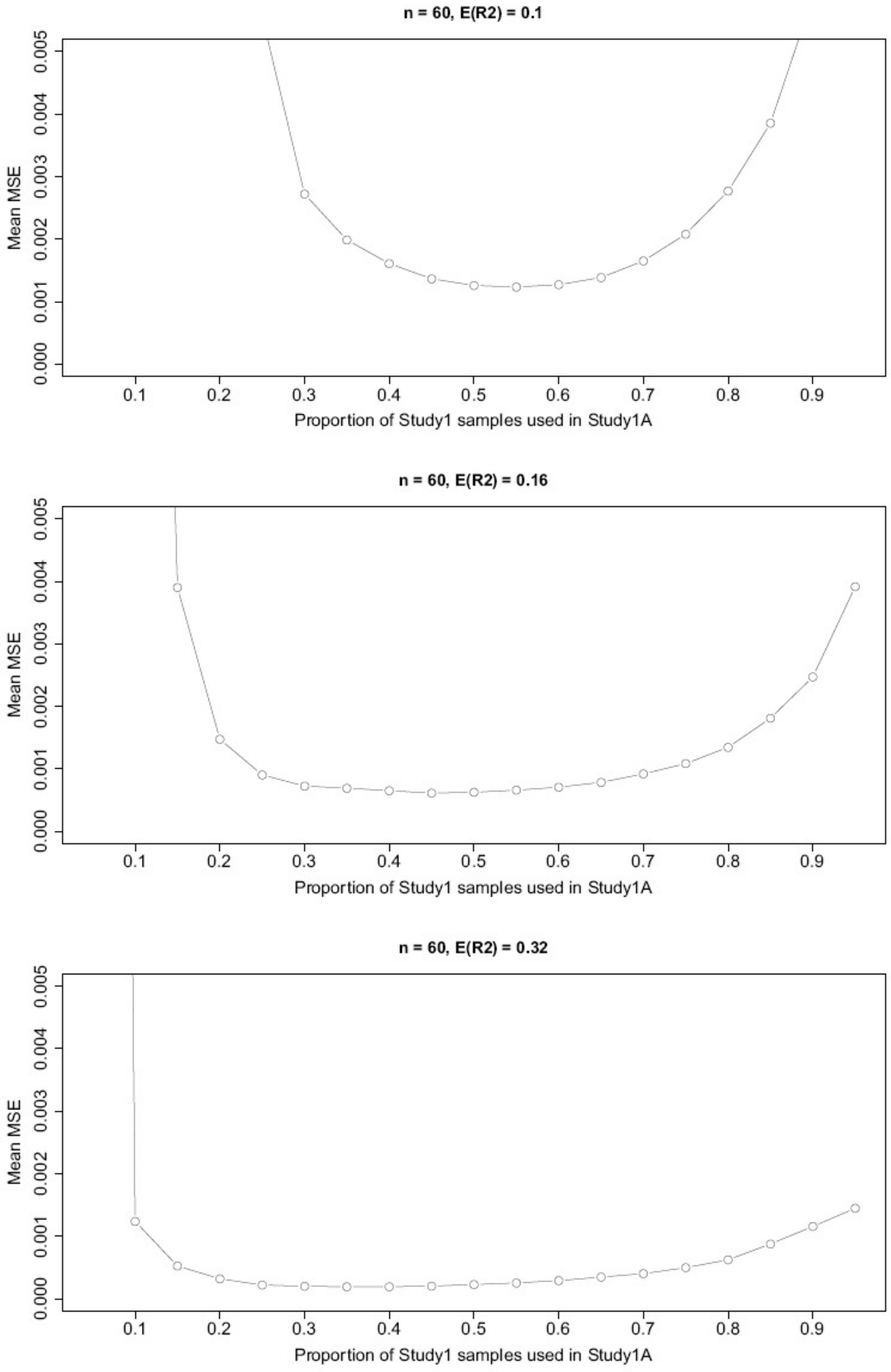


Figure 4.2 *MSE*s for the overlap percentage estimator across different sample splitting proportions for sample size 120

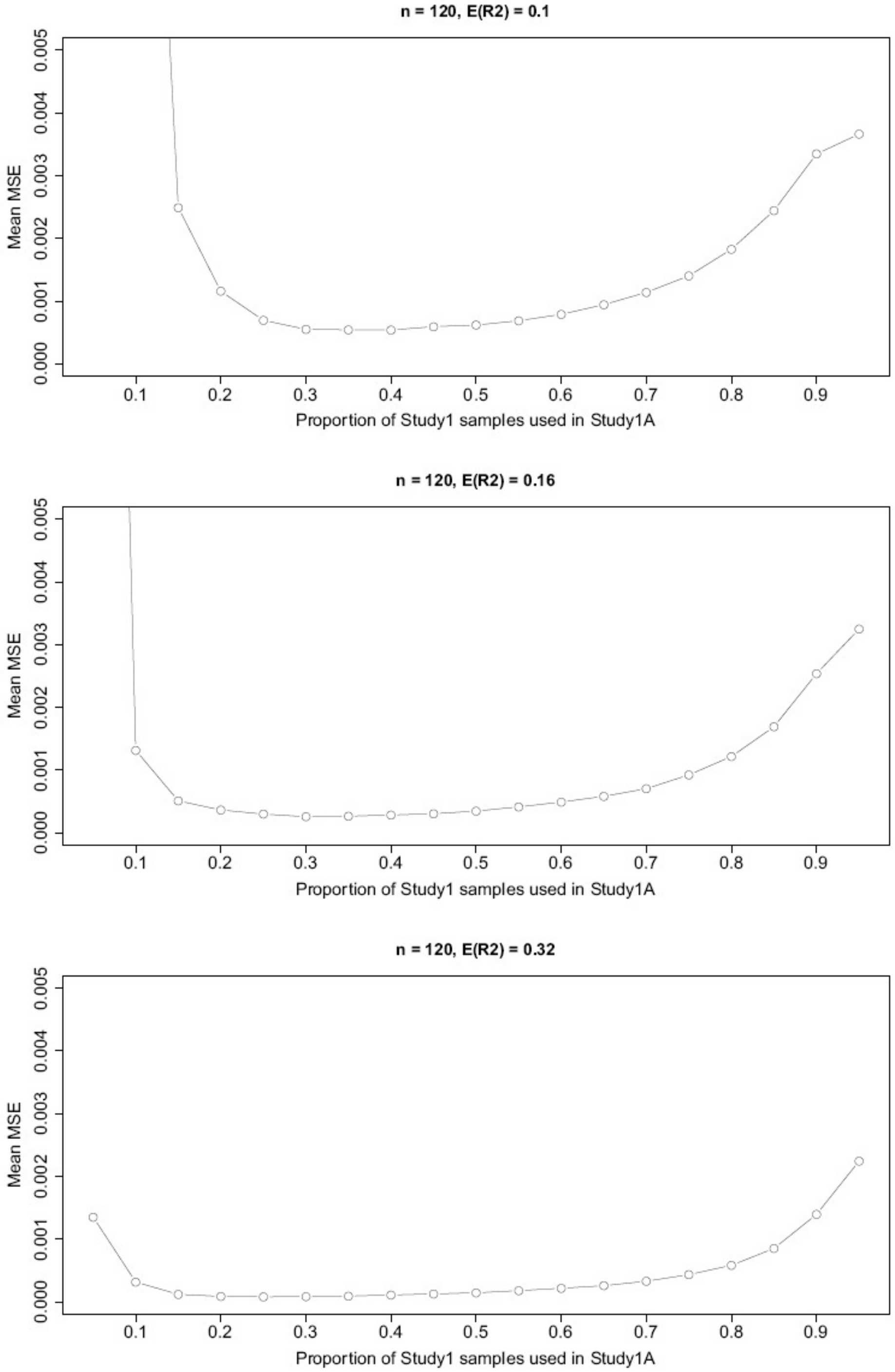


Figure 4.3 *MSE*s for the overlap percentage estimator across different sample splitting proportions for sample size 240

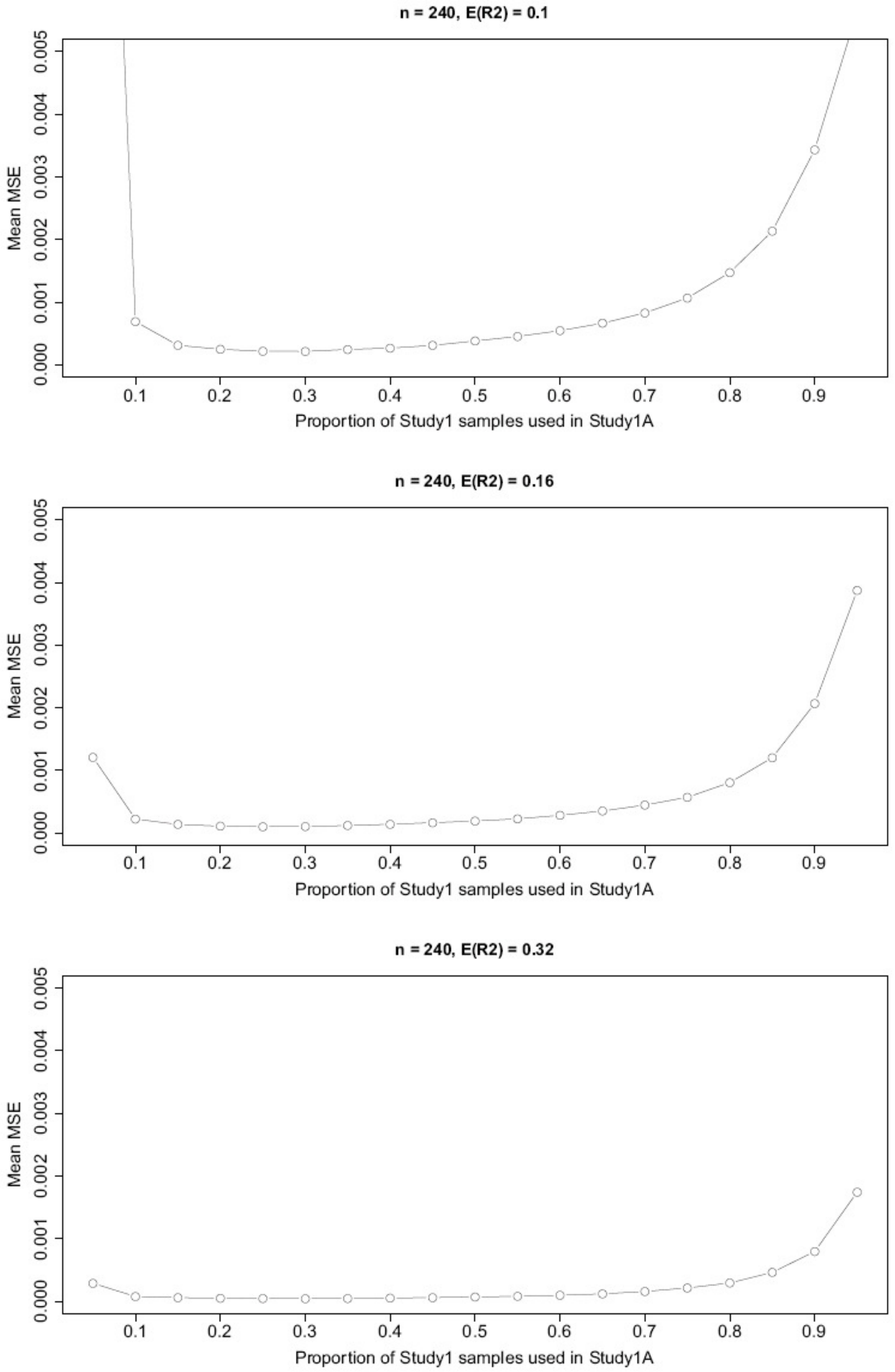


Table 4.1 Estimation results from simulations where the two effect size assumptions both hold and $\pi = \frac{2}{3}$

Parameter Settings				Estimation			
n	E(R^2)	Power*	True π	$\hat{\pi}_{adjusted}$	$\hat{\pi}_{adjusted} - \pi$	SE × 10^2	$\sqrt{MSE}$ × 10^2
60	0.10	~10%	0.667	0.6658	-0.00088	3.146	3.151
60	0.16	~25%	0.667	0.6661	-0.00053	1.981	1.983
60	0.32	~40%	0.667	0.6663	-0.00032	1.022	1.023
120	0.10	~30%	0.667	0.6647	-0.00194	1.589	1.602
120	0.16	~40%	0.667	0.6657	-0.00095	1.095	1.099
120	0.32	~65%	0.667	0.6663	-0.00035	0.678	0.679

*Average statistical power of a study with the sample size of Study 2 to detect signals at FDR of 0.001

Table 4.2 Estimation results from simulations where the two effect size assumptions both hold and $\pi = 0.05$

Parameter Settings				Estimation			
n	E(R^2)	Power*	True π	$\hat{\pi}_{adjusted}$	$\hat{\pi}_{adjusted} - \pi$	SE × 10^2	$\sqrt{MSE}$ × 10^2
60	0.10	~10%	0.05	0.0520	0.00197	1.297	1.313
60	0.16	~25%	0.05	0.0513	0.00132	0.796	0.809
60	0.32	~40%	0.05	0.0506	0.00063	0.542	0.547
120	0.10	~30%	0.05	0.0506	0.00064	0.688	0.692
120	0.16	~40%	0.05	0.0506	0.00059	0.495	0.500
120	0.32	~65%	0.05	0.0505	0.00046	0.391	0.397

*Average statistical power of a study with the sample size of Study 2 to detect signals at FDR of 0.001

Table 4.3 Estimation results from simulations where the two effect size assumptions both hold and $\pi = 0.95$

Parameter Settings				Estimation			
n	E(R^2)	Power*	True π	$\hat{\pi}_{adjusted}$	$\hat{\pi}_{adjusted} - \pi$	SE × 10^2	$\sqrt{MSE}$ × 10^2
60	0.10	~10%	0.95	0.9494	-0.00057	2.682	2.687
60	0.16	~25%	0.95	0.9491	-0.00090	1.585	1.593
60	0.32	~40%	0.95	0.9506	0.00056	0.850	0.855
120	0.10	~30%	0.95	0.9492	-0.00078	1.299	1.304
120	0.16	~40%	0.95	0.9502	0.00019	0.895	0.897
120	0.32	~65%	0.95	0.9501	0.00013	0.623	0.626

*Average statistical power of a study with the sample size of Study 2 to detect signals at FDR of 0.001

Table 4.4 Estimation results from simulations where overlapping eQTLs have larger mean effect size than non-overlapping eQTLs in Study 1

Parameter Settings				Estimation			
n	E(R^2)	True π	Violation Level*	$\hat{\pi}_{adjusted}$	$\hat{\pi}_{adjusted} - \pi$	SE × 10^2	$\sqrt{MSE}$ × 10^2
60	0.32	0.667	10%	0.682	0.0149	1.05	1.83
60	0.32	0.667	20%	0.696	0.0297	1.07	3.16
60	0.32	0.667	30%	0.711	0.0441	1.10	4.54
60	0.32	0.05	10%	0.054	0.0040	0.58	0.71
60	0.32	0.05	20%	0.058	0.0076	0.61	0.98
60	0.32	0.05	30%	0.062	0.0115	0.64	1.32
60	0.32	0.95	10%	0.954	0.0038	0.87	0.95
60	0.32	0.95	20%	0.957	0.0068	0.88	1.11
60	0.32	0.95	30%	0.960	0.0096	0.88	1.31

* The amount by which the mean effect size of overlapping eQTLs is larger than that of non-overlapping eQTLs

Table 4.5 Estimation results from simulations where overlapping eQTLs have larger effect sizes in Study 1 than in Study 2

Parameter Settings				Estimation			
n	E(R^2)	True π	Violation Level*	$\hat{\pi}_{adjusted}$	$\hat{\pi}_{adjusted} - \pi$	SE × 10^2	$\sqrt{MSE}$ × 10^2
60	0.32	0.667	10%	0.652	-0.0150	1.03	1.86
60	0.32	0.667	20%	0.636	-0.0307	1.03	3.27
60	0.32	0.667	30%	0.620	-0.0472	1.06	4.86
60	0.32	0.05	10%	0.050	-0.0005	0.55	0.56
60	0.32	0.05	20%	0.048	-0.0017	0.55	0.59
60	0.32	0.05	30%	0.047	-0.0030	0.54	0.66
60	0.32	0.95	10%	0.930	-0.0205	0.88	2.26
60	0.32	0.95	20%	0.907	-0.0430	0.90	4.41
60	0.32	0.95	30%	0.884	-0.0662	0.95	6.70

* The amount by which overlapping eQTLs' effect sizes in Study 1 are larger than those in Study 2

Table 4.6 Ideal splitting proportions in different parameter settings

Parameter Settings			Estimation		
n	E(R^2)	True π	minimum MSE×10^4	Ideal Splitting Proportion*	Splitting Proportion Window**
60	0.10	0.667	12.3	0.55	0.40 - 0.70
60	0.16	0.667	6.1	0.45	0.25 - 0.65
60	0.32	0.667	1.9	0.35	0.25 - 0.55
120	0.10	0.667	5.4	0.40	0.25 - 0.60
120	0.16	0.667	2.6	0.30	0.20 - 0.50
120	0.32	0.667	0.8	0.25	0.20 - 0.40
240	0.10	0.667	2.2	0.30	0.15 - 0.45
240	0.16	0.667	1.0	0.25	0.15 - 0.40
240	0.32	0.667	0.4	0.30	0.15 - 0.45

* The splitting proportion that gives the minimum MSE

** A set of values for the splitting proportion that gives MSE ≤ 1.5×minimum(MSE)

Chapter 5

eQTL mapping in skin

5.1 Introduction

In this chapter we report the mapping of eQTLs in skin tissue from psoriatic patients and normal controls. Psoriasis is an immune-mediated, inflammatory and hyperproliferative disease of the skin and joints. It provides an ideal system for eQTL mapping analysis, because psoriasis has a strong genetic basis and diseased tissue is readily accessible.

We describe a large catalog of genetic variants influencing transcript levels in both normal and psoriatic skin. This catalog is based on gene expression and genotype data that we have collected from normal skin from healthy controls (normal skin), normal-appearing skin from psoriatic patients (uninvolved skin), and diseased skin from psoriatic patients (lesional skin). This catalog represents a genetic map of gene regulation in skin and provides a useful tool for examining the functional impact of genetic variants associated with psoriasis and other skin diseases.

As mentioned in Chapter 1, we have previously performed differential gene expression study of psoriasis using the same gene expression data. We find that gene expression profiles for lesional skin are markedly different from those of normal and uninvolved skin. Here, we focus on identifying eQTLs in the different skin types and investigating whether different physiological conditions of skin types will affect the existence of these eQTLs.

5.2 Methods

Mapping eQTLs in skin

Subjects. We enrolled 58 psoriatic patients and 64 normal healthy controls in the study. Patients had to have at least one well demarcated erythematous scaly psoriatic plaque that was not limited to the scalp. In those instances where there was only a single psoriatic plaque, the case was only considered if the plaque occupied more than 1% of total body surface area. Study subjects did not use any (a) systemic anti-psoriatic treatments for 2 weeks prior to biopsy or (b) topical anti-psoriatic treatments for 1 week prior to biopsy. Informed consent was obtained from all subjects, under protocols approved by the Institutional Review Board of the University of Michigan Medical School and was conducted according to the Declaration of Helsinki Principles. Subjects with failed gene expression profiling or failed genotyping were excluded from the analysis. The final analysis included 53 psoriatic patients and 57 normal controls.

Genotype Data. Subjects (as a subset of 1,409 cases and 1,436 controls) were genotyped by Perlegen Sciences using four proprietary, high-density oligonucleotide arrays in partnership with the Genetic Association Information Network (GAIN). Quality control filters, which are described in detail in Nair et al. [2009], were applied. Briefly, we excluded markers with <95% genotype call rates, with minor allele frequency <1%, with HWE P value $<10^{-6}$, with >2 mismatches among 48 pairs of individuals that were genotyped twice or with >2 mendelian inconsistencies among 27 trios; we also excluded samples with call rates <95% and with outlier heterozygosities. In total, 438,670 autosomal SNPs were genotyped successfully for 53 psoriatic patients and 57 controls. As previously described [Li, et al. 2010; Li, et al. 2009], we used information on patterns of haplotype variation in the HapMap CEU samples (release 21) to infer missing genotypes 'in silico'. We analyzed only SNPs that were genotyped or could be imputed with relatively high confidence (estimated r^2 between imputed SNP and true genotypes >0.3, so that patterns of haplotype sharing between sampled individuals and HapMap samples consistently indicated a specific allele; we use this r^2 based threshold, rather than one based on the posterior probability of each imputed genotype, because it naturally accommodates SNPs with different allele frequencies and because it is the same threshold used in many recent genomewide association studies, including for psoriasis [Nair, et al. 2009]).

Gene Expression Data. Two biopsies (one lesional skin, one uninvolved skin; 6 mm each) were taken under local anesthesia from each psoriatic subject while one 6 mm punch biopsy (normal skin) was taken from normal controls. Lesional skin biopsies were taken from psoriasis plaques, and uninvolved skin biopsies were taken from the buttocks, at least 10 cM away from the nearest plaque. The normal skin biopsies were also taken from the buttocks. RNA from each biopsy was isolated using the RNeasy kit (Qiagen, Valencia, CA). Samples were run on Affymetrix U133 Plus 2.0 arrays to evaluate expression of ~54,000 probes in accordance with the manufacturer's protocol. The raw data from 180 microarrays was processed using the Robust Multichip Average (RMA) method. Prior to analysis, we adjusted RMA expression values to account for batch and gender effects. Procedures for extracting RNA, controlling RNA quality and pre-processing of gene expression data are described in detail elsewhere [Gudjonsson, et al. 2009].

eQTL mapping. We tested SNP-gene expression associations separately in normal skin (n = 57), in uninvolved skin (n = 53), and in lesional skin (n = 53). Given the small sample size in each analysis (<60) and hence the relatively low statistical power, we tested only cis-associations between each transcript (i.e. probe) and those SNPs in its cis-candidate region (from 1Mb upstream of the transcription start site to 1Mb downstream of the transcription end site). Specifically, we used the score test in Merlin (fastassoc option) to test the association [Chen and Abecasis 2007]. For genotyped SNPs, the number of copies of one allele was modeled. For imputed SNPs, the dosage (i.e. the expected number of copies) of one allele was modeled. We used a p-value threshold of 9×10^{-7} as the significance threshold to originally identify *cis*-eQTLs, which corresponds to a false discovery rate (FDR) of approximately 0.01 in each of the three skin types. For gene expression traits that were significantly associated with more than one *cis*-SNP, we reported only the most significant *cis*-SNP and counted this SNP-gene pair as one independent *cis*-association.

Measuring cis-eQTL overlap among three skin types. To test if *cis*-associations identified in one skin type can be replicated in the other two skin types, we started with the significant SNP-gene pairs identified in one skin type and then tested their significance in other two skin types using a nominal p-value threshold of 0.05. We did not use the genome-wide p-value significance threshold of 9×10^{-7} because here we only tested a small number of SNP-gene pairs for replication. Ideally we should use the novel method we developed in Chapter 3 to better measure the *cis*-eQTL overlap among three skin types, but the small sample sizes of the three skin types prevented us from using the sample splitting strategy if we wanted to maintain adequate statistical power for the study.

Studying other features of skin cis-eQTLs

Relationship of skin eQTL SNPs to association signals in psoriasis genome-wide association studies (GWAS). We compiled a list of 9,462 eQTL SNPs that had at least one cis-association in normal, uninvolved or lesional skin that passed the significance threshold of 9×10^{-7} (corresponding to FDR at 0.01). From this list we selected all 389 skin eQTL SNPs that were nearly independent of each other (linkage disequilibrium $r^2 < 0.2$), favoring those SNPs with stronger cis-association p-values. Using a quantile-quantile (Q-Q) plot, we compared the distribution of psoriasis association p-values for these 389 eQTL SNPs against the null expectation. Disease association p-values were derived from a meta-analysis of two psoriasis GWAS: GAIN psoriasis GWAS [Nair, et al. 2009] and the Kiel psoriasis study [Ellinghaus, et al. 2010], where 472 psoriatic cases and 1,146 normal controls were genotyped by Illumina HumanHap 550 v1 (550k) SNP arrays, and then all HapMap SNPs were imputed in the same way as mentioned above. To further compare this distribution with the GWAS p-value distribution of non-eQTL SNPs, we first removed all skin eQTL SNPs from the GWAS SNP set, and then randomly picked 389 of the remaining SNPs 5,000 times to derive confidence intervals for the p-value distribution of non-eQTL SNPs. Because we were interested in testing if eQTL SNPs could reveal new psoriasis susceptibility loci, we removed from both skin eQTL and non-eQTL SNP lists those SNPs that were in the seven replicated loci from our recently published GWAS study (i.e. *HLA-C, IL12B, TNIP1, IL13, TNFAIP3, IL23A/STAT2*, and *IL23R*; the excluded region is 1Mb on both sides of the most significant SNP in GWAS).

Gene Ontology (GO) enrichment analysis of genes associated with cis-eQTLs. We searched for GO terms that were significantly enriched in each list of genes associated with eQTLs in the three skin types. This GO category enrichment analysis was performed using the publicly available software DAVID (Database for Annotation, Visualization and Integrated Discovery, http://david.abcc.ncifcrf.gov/, Bethesda, MD).

5.3 Results

Mapping *cis*-eQTLs in skin

As described previously [Gudjonsson, et al. 2009], we found that expression profiles for lesional skin were markedly different from those of normal and uninvolved skin. Using principal component analysis (PCA, Figure 1.2), we achieved near-perfect separation of lesional skin from normal and uninvolved skin, while the latter two skin types were intermixed. Here, we do not focus on a comparison of expression levels between the tissues (as in Gudjonsson et al. [2009]), but instead report on the *cis*-eQTLs in the different skin types.

Using a nominal p-value threshold of 9×10^{-7} (corresponding to an FDR for *cis*-association at approximately 0.01 for each of the three skin types), we identified 331, 275, and 235 independent *cis*-associations in normal, uninvolved, and lesional skin, respectively. We have created a publicly available database containing the catalogs of *cis*-eQTLs for each of the three skin types, which will allow researchers to interrogate their specific SNPs or genes of interest. Figure 5.1 gives two examples of *cis*-association between gene transcripts and their nearby SNPs: *ERAP2* has the most significant *cis*-association in normal skin peaking at rs2910686 and *cis*-associations at similar significant levels are observed in both uninvolved and lesional skin; *RPS26* has one of the most significant *cis*-associations in uninvolved and lesional skin peaking at rs11171739. Although the signal is less significant in normal skin, the same overall pattern of association is observed.

We then measured the overlap of *cis*-eQTLs among the three skin types by testing how many significant cis-eQTLs in one skin type were replicated in other two skin types at nominal p-value threshold of 0.05. The results are shown in Figure 5.2: 95.1%, 96.7%, and 98.7% of the significant *cis*-eQTLs in normal, uninvolved, and lesional skin, respectively, are also shared by the other two skin types. Furthermore, we observed only two *cis*-eQTLs in each set that were only observed in that skin type and this number is consistent with the false discovery rate of 0.01 (i.e. in a set of 200 signals, we expect to see two false positives). These results, consistent with the similar *cis*-association patterns observed in the three skin types (Figure 5.1), indicate that nearly all *cis*-eQTLs currently identified are shared by normal, uninvolved, and lesional skin. Therefore, the dramatic physiological changes that are apparent in psoriatic skin appear to have little impact on the identity of *cis*-eQTLs in skin.

Studying other features of skin cis-eQTLs

Relationship of skin eQTL SNPs to association signals in psoriasis GWAS

Out of a total of 9,462 SNPs that passed the eQTL significance threshold of 9×10^{-7} in normal, uninvolved, or lesional skin (FDR=0.01), we identified 389 independent skin eQTLs ($r^2 < 0.2$), and examined their potential importance in the context of psoriasis and other complex genetic disorders that have been subjected to GWAS. First, using the meta-analysis results for two psoriasis GWAS, we compared the distribution of disease association p-values for SNPs that define eQTLs and those that do not. For this comparison, we exclude SNPs within 1 Mb of regions known to be associated with psoriasis, so as to more directly evaluate the ability of eQTLs to suggest new loci. Figure 5.3 shows the quantile-quantile plot for the 389 independent eQTL SNPs in skin, with confidence intervals (CIs) estimated by sampling the same number of non-eQTL SNPs. The Q-Q plot clearly shows a trend for eQTL SNPs to be more strongly associated with psoriasis than non-eQTL SNPs; further, the majority of eQTL SNPs exceed the 75% confidence interval obtained by sampling non-eQTL SNPs and 6 of the top 8 ranked eQTL SNPs exceed the 95% CI determined by sampling non-eQTL SNPs. Table 5.1 lists the top 8 eQTL SNPs from the Q-Q plot, along with their cis-association and psoriasis GWAS results. While the overlap between eQTL signals and psoriasis associations is intriguing, we recognize that further follow-up genotyping will be required to confirm these signals. Still, examination of the genes in this list (*FUT2*, *RPS26*, *ARTS-1, and LRAP*) suggests several plausible biological connections, which are detailed in the discussion.

We also studied this list of skin eQTL SNPs in the context of other complex genetic diseases that have been subjected to GWAS. Among 1,482 significant ($p<10^{-5}$) SNP associations from 321 published GWAS curated by the National Human Genome Research Institute (http://www.genome.gov/gwastudies/), we found 14 skin eQTL SNPs (Supplementary Table 5.1), which are associated with 19 disease traits, while by chance we only expected to see five overlapping SNPs.

Enrichment of eQTLs for genes involved in MHC Class I antigen presentation

Using DAVID [Dennis, et al. 2003; Huang da, et al. 2009], we carried out a gene ontology analysis of biological processes enriched in eQTL transcripts from lesional, uninvolved, and normal skin, as well as LCLs. This analysis revealed significant enrichment for eQTLs regulating genes involved in the processing and presentation of endogenous peptide antigens via MHC class I in lesional skin (Supplementary Table 5.2). We also observed a similar but non-significant trend in uninvolved skin, normal skin, and LCLs. The skin eQTL associated genes observed to be enriched in this GO category included *ARTS-1* (also known as Endoplasmic Reticulum Aminopeptidase 1; *ERAP1*), *TAP2* (Transporter, ATP-Binding Cassette, Major Histocompatibility Complex, 2), *LRAP* (Endoplasmic Reticulum Aminopeptidase 2; *ERAP2*), and *TAPBPL* (TAP binding protein-like). These genes are intimately involved in the transport (*TAP2*, *TAPBLPL*) and processing (*ARTS-1*, *LRAP*) of peptides within the endoplasmic reticulum for subsequent presentation on the surface of cells within the antigen binding groove of MHC I class molecules [Peaper and Cresswell 2008]. These results provide further evidence for the genetic control of genes involved in MHC Class I antigen presentation in the skin.

Localization of cis-eQTLs with respect to the transcription start site of the transcripts they putatively regulate

We studied the localization of the most significant eQTL for each *cis*-association in normal, uninvolved, and lesional skin with respect to the transcription start site of the gene it putatively regulates. The most significant cis-eQTLs localize closely (most of them within 100kb) and roughly symmetrically around the transcription start site (Supplementary Figure 5.1). This localization pattern in the skin confirms previous observations in LCLs [Dixon, et al. 2007; Stranger, et al. 2007; Veyrieras, et al. 2008].

5.4 Discussion

We are the first to perform an eQTL mapping analysis in human skin and to identify eQTLs in normal skin, uninvolved skin from psoriatic individuals, and lesional psoriatic skin. Our results thus provide a useful resource for studying regulation of gene expression in skin. Furthermore, these results are useful in interpreting genetic susceptibility loci identified by GWAS of multiple complex traits, including skin diseases.

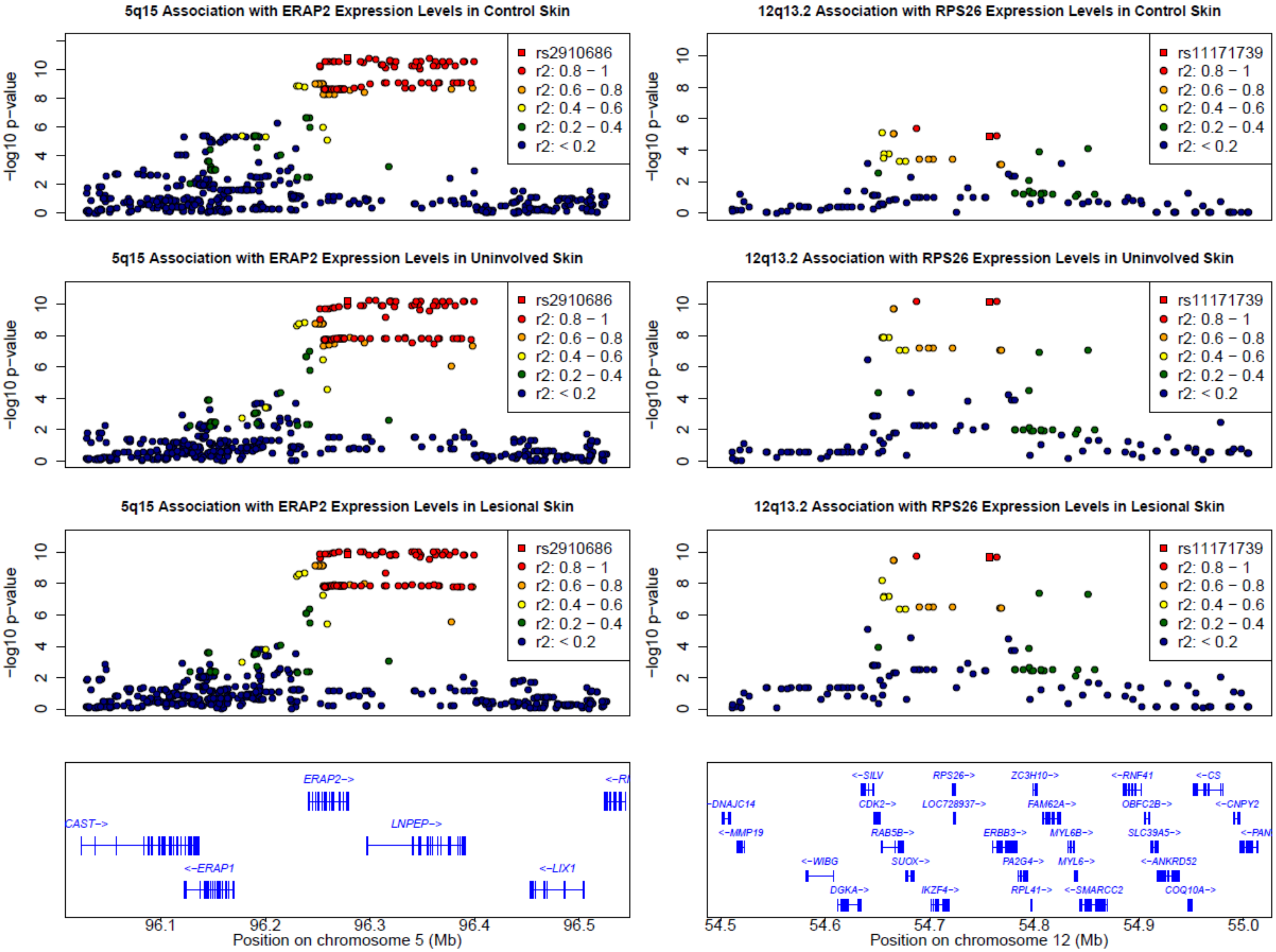
We examined whether eQTL SNPs were more likely to be associated with psoriasis. This is analogous to other analyses that might focus on SNPs that are likely to be functional because, for example, they encode non-synonymous SNPs. This focused analysis of eQTL SNPs identified the *FUT2* (rs492602, most significant *cis*-association p-value = 1×10^{-9} in lesional skin), *RPS26* (rs11171739, most significant p-value = 7×10^{-11} in uninvolved skin), *ARTS-1* (rs7063, most significant p-value = 6×10^{-10} in normal skin), and *LRAP* (rs2910686, most significant p-value = 6×10^{-11} in normal skin) loci as regions where eQTL SNPs are attractive for further analysis in psoriasis and other autoimmune diseases. The *FUT2* region has been associated with serum levels of vitamins functioning in the folate pathway [Tanaka, et al. 2009]. *FUT2* encodes a fucosyltransferase involved in the synthesis of blood group antigens [Kelly, et al. 1995], which are also involved in the fucosylation of cell-surface proteins on epithelia [Boren, et al. 1993]. SNP rs492602 is the most significant eQTL SNP associated with *FUT2* expression. Interestingly enough, it is among a set of SNPs in perfect linkage disequilibrium ($r^2 = 1$) that yielded the most significant p-values in the psoriasis meta-analysis for the *FUT2* region. *RPS26* encodes a ribosomal protein subunit and was bio-informatically linked to antigen processing and presentation and T-cell mediated immunity. eQTL SNPs near *RPS26* (e.g. rs2292239) have been associated with Type I diabetes [Schadt, et al. 2008], although their relevance as a direct disease determinant has been questioned recently [Plagnol, et al. 2009]. In our data, the same SNPs show suggestive association with psoriasis (p-value = 0.01) and with *RPS26* transcript levels in both uninvolved and lesional skin (p-value < 10^{-9} in both tissues). We also observed highly significant eQTL associations for *ARTS-1* (also known as *ERAP1*) and *LRAP* (also known as *ERAP2*). These eQTLs have also been noted in previous eQTL studies of LCLs [Morley, et al. 2004]. The products of these genes are associated in heterodimeric complexes within the endoplasmic reticulum and have been shown to be intimately involved in the process of trimming peptides in preparation for loading into MHC Class I molecules [Saveanu, et al. 2005]. Previously, variants residing in the gene *ARTS-1* have been significantly associated with ankylosing spondylitis [Burton, et al. 2007]. This is particularly interesting given the role of *ARTS-1* in MHC Class I antigen processing and presentation and the fact that psoriasis, psoriatic arthritis, and ankylosing spondylitis are the only major autoimmune diseases that are primarily associated with MHC Class I. Inspection of Q-Q plots suggests that eQTL SNPs in skin are enriched for psoriasis association signals, which provides support for using eQTLs as a potential filter for choosing candidate loci for replication. Although these eQTL SNPs do not currently attain genome-wide significance for association with psoriasis (Table 5.1), their cis-association with a gene transcript indicates that they tag functional/regulatory variants in skin tissue and hence may play a role in the disease.

Because of the limited sample size in each of the skin tissue types, our study is powered to identify only strong *cis* signals (i.e. signals with relatively large effect sizes). Our analysis has shown that the vast majority of strong *cis*-eQTLs are shared in the three skin tissue types, which indicates that the physiology of the disease does not change the identity of those strong *cis*-eQTLs. This finding does not preclude a role for cis-eQTLs in psoriasis or other skin diseases. First, although it appears that the same set of transcripts are cis-regulated in all three skin types, differences in genotype frequencies for the regulatory SNP between cases and controls can result in differences in expression levels for transcripts they regulate between psoriatic and normal skin. Second, it is possible that the same cis-eQTLs can have different effects on genes in normal and diseased skin tissues.

In summary, our work sheds light on transcriptional regulation by genetic variants in skin and provides an insight into the genetics of gene expression in different skin tissue types.

5.5 Figures and tables

Figure 5.1 Regional plots for evidence of cis-association between SNPs and gene *ERAP2* (left panel) or gene *RPS26* (right panel). P-values at the most significant SNP (the most significant SNP associated with *ERAP2* in normal skin on the left panel; the most significant SNP associated with *RPS26* in uninvolved skin on the right panel) are highlighted with a square. P-values for other SNPs are drawn as circles and color coded according to the degree of linkage disequilibrium with the most significant SNP.



5q15 Association with ERAP2 Expression Levels in Control Skin

12q13.2 Association with RPS26 Expression Levels in Control Skin

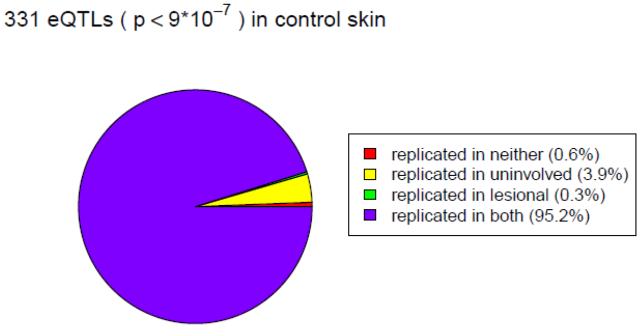
5q15 Association with ERAP2 Expression Levels in Uninvolved Skin

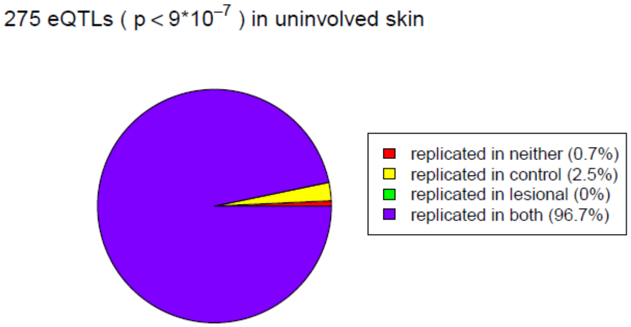
12q13.2 Association with RPS26 Expression Levels in Uninvolved Skin

5q15 Association with ERAP2 Expression Levels in Lesional Skin

12q13.2 Association with RPS26 Expression Levels in Lesional Skin

Figure 5.2 The sharing of cis-eQTLs in normal, uninvolved, and lesional skin with the other two types of skin (the p-value threshold for discovery is 9×10^{-7} and the p-value threshold for replication is 0.05).


331 eQTLs ($p < 9*10^{-7}$) in control skin


275 eQTLs ($p < 9*10^{-7}$) in uninvolved skin

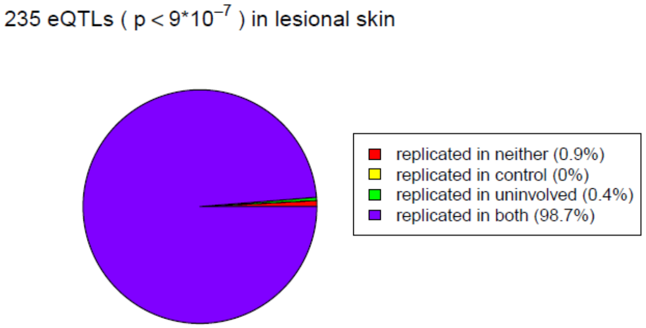
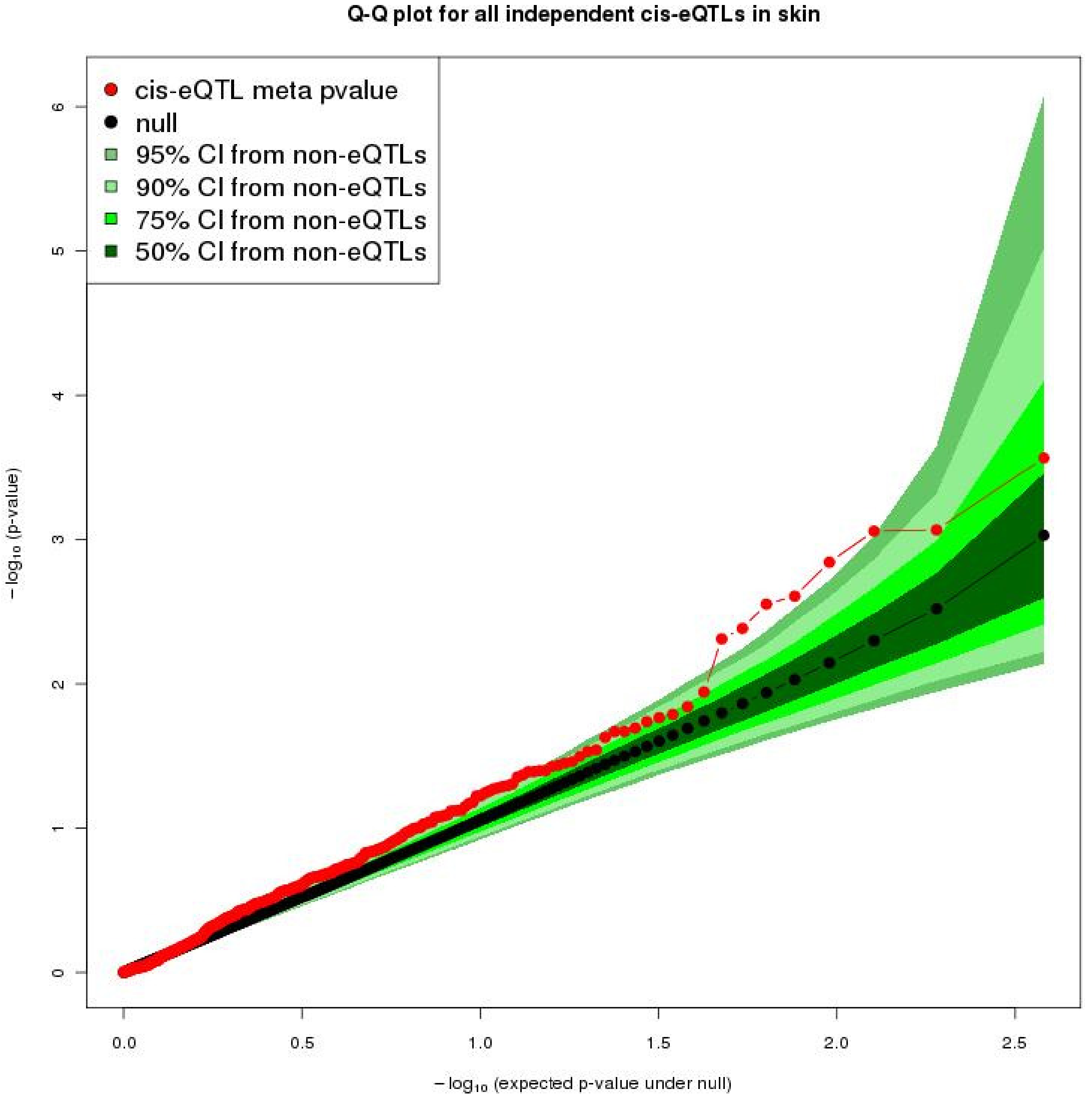

235 eQTLs ($p < 9*10^{-7}$) in lesional skin

Figure 5.3 Quantile-quantile plot of psoriasis GWAS p-values for 389 independent eQTL SNPs in skin, with confidence intervals defined by non-eQTL SNPs.



Q-Q plot for all independent cis-eQTLs in skin

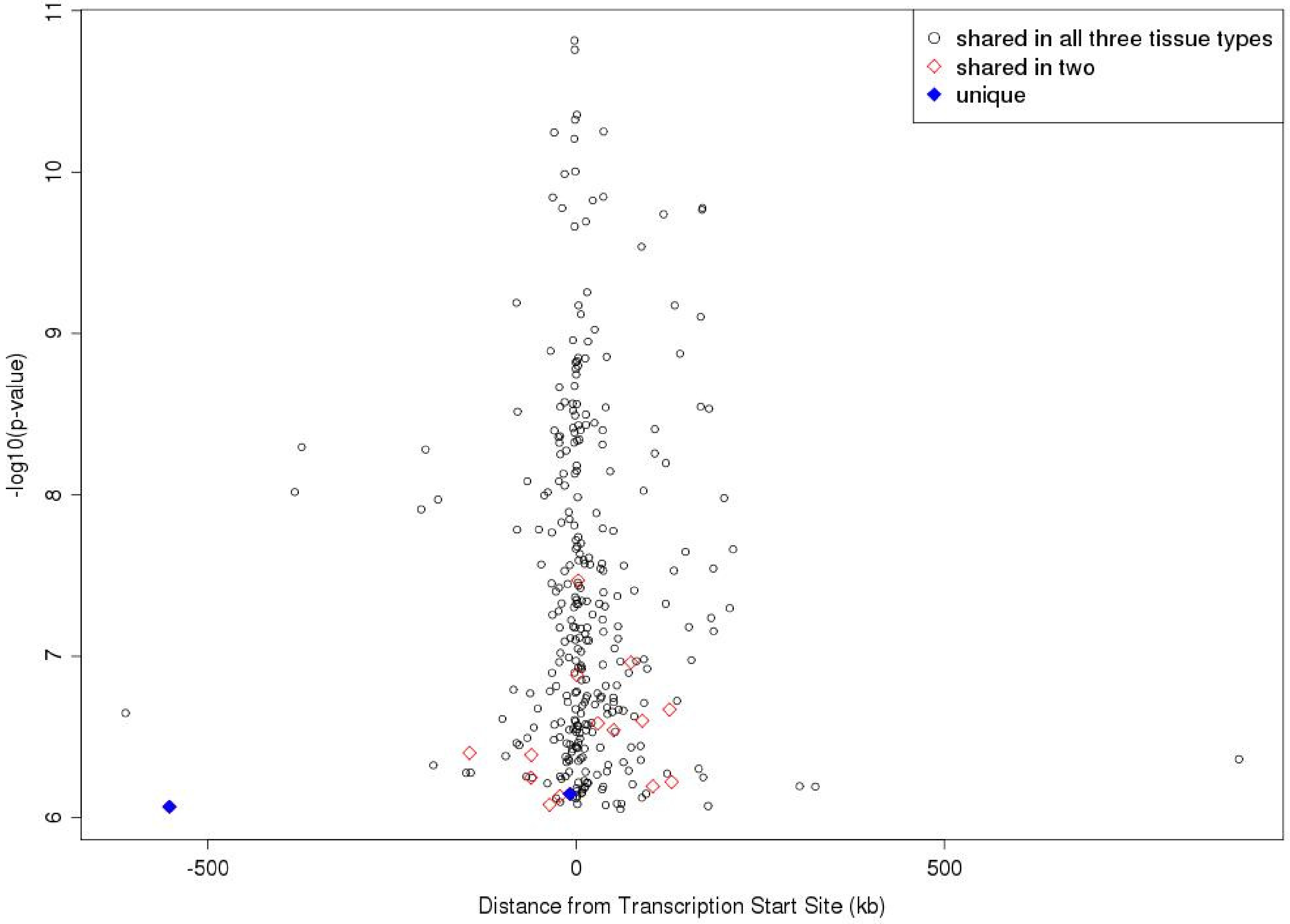
Table 5.1 *cis*-association and psoriasis association meta-analysis results for the 8 independent skin eQTL SNPs with most significant psoriasis association

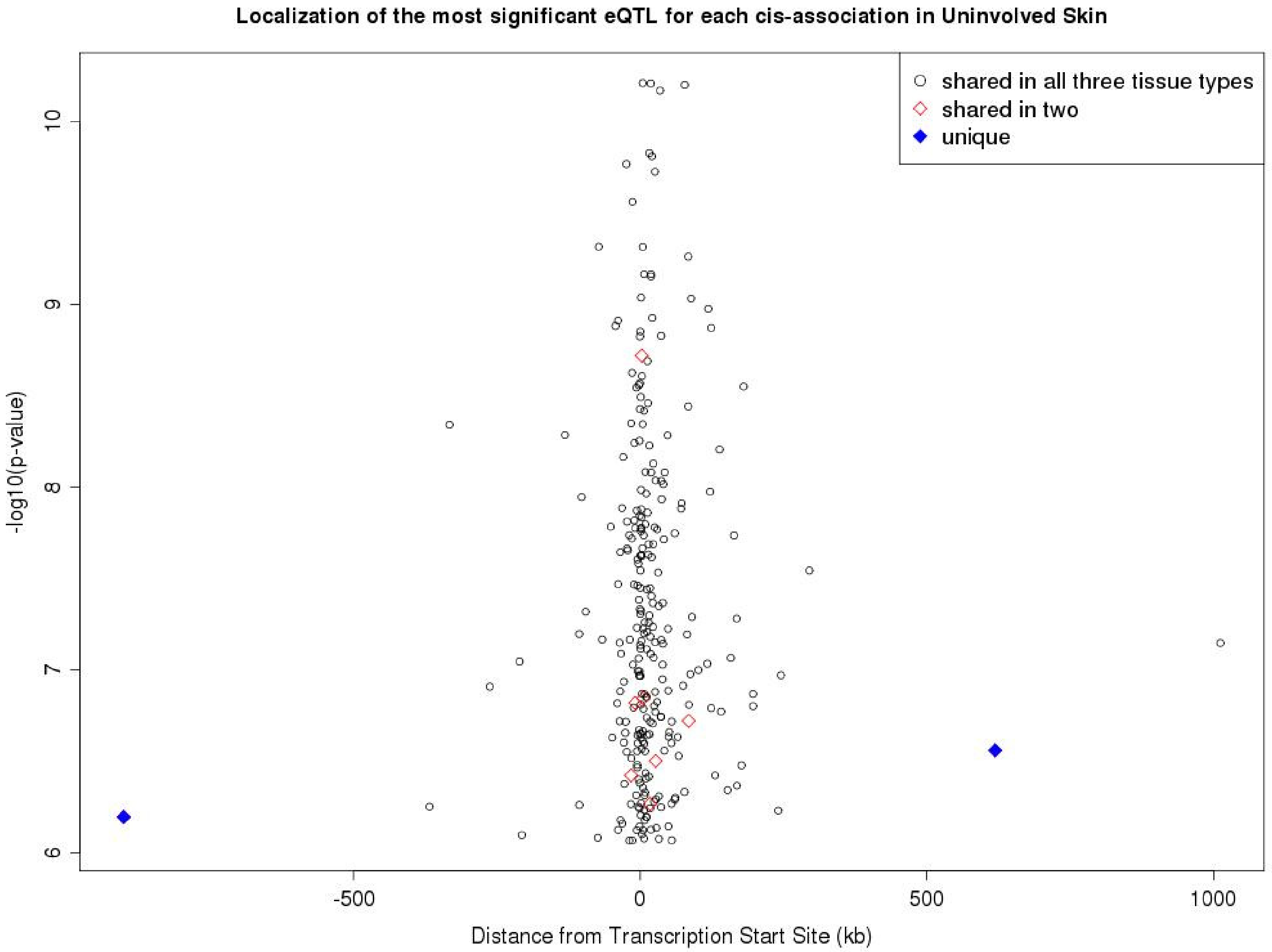
Marker	*cis*-association				Psoriasis GWAS			
	cis-associated gene	*cis*-association p-value			Alleles risk/nonrisk	Risk allele frequency	OR	Meta p-value[a]
		Normal	Uninvolved	Lesional				
rs492602	FUT2	0.00015	1.9E-06	1.0E-09	G/A	0.482	1.169	0.00027
rs12039309	TMEM77	3.2E-07	5.3E-06	2.7E-05	G/T	0.271	1.212	0.00085
rs11171739	RPS26	1.3E-05	6.8E-11	2.1E-10	C/T	0.418	1.174	0.00087
rs13008446	LOC348751	1.9E-06	0.00017	1.3E-07	A/G	0.655	1.191	0.0014
rs8082268	C17orf45	1.0E-07	6.7E-07	4.4E-05	C/T	0.651	1.188	0.0024
rs2910686	ERAP2	5.6E-11	8.5E-11	6.1E-11	C/T	0.427	1.144	0.0027
rs139910	TNRC6B	2.6E-07	0.0005	0.033	C/G	0.744	1.117	0.0041
rs503612	ENDOD1	2.9E-09	9.6E-09	6.9E-09	C/A	0.564	1.148	0.0048

[a] The p-value from the psoriasis GWAS meta analysis (GAIN study + Kiel study). For all 8 SNPs, the effect directions are consistent in the two studies of the meta analysis.

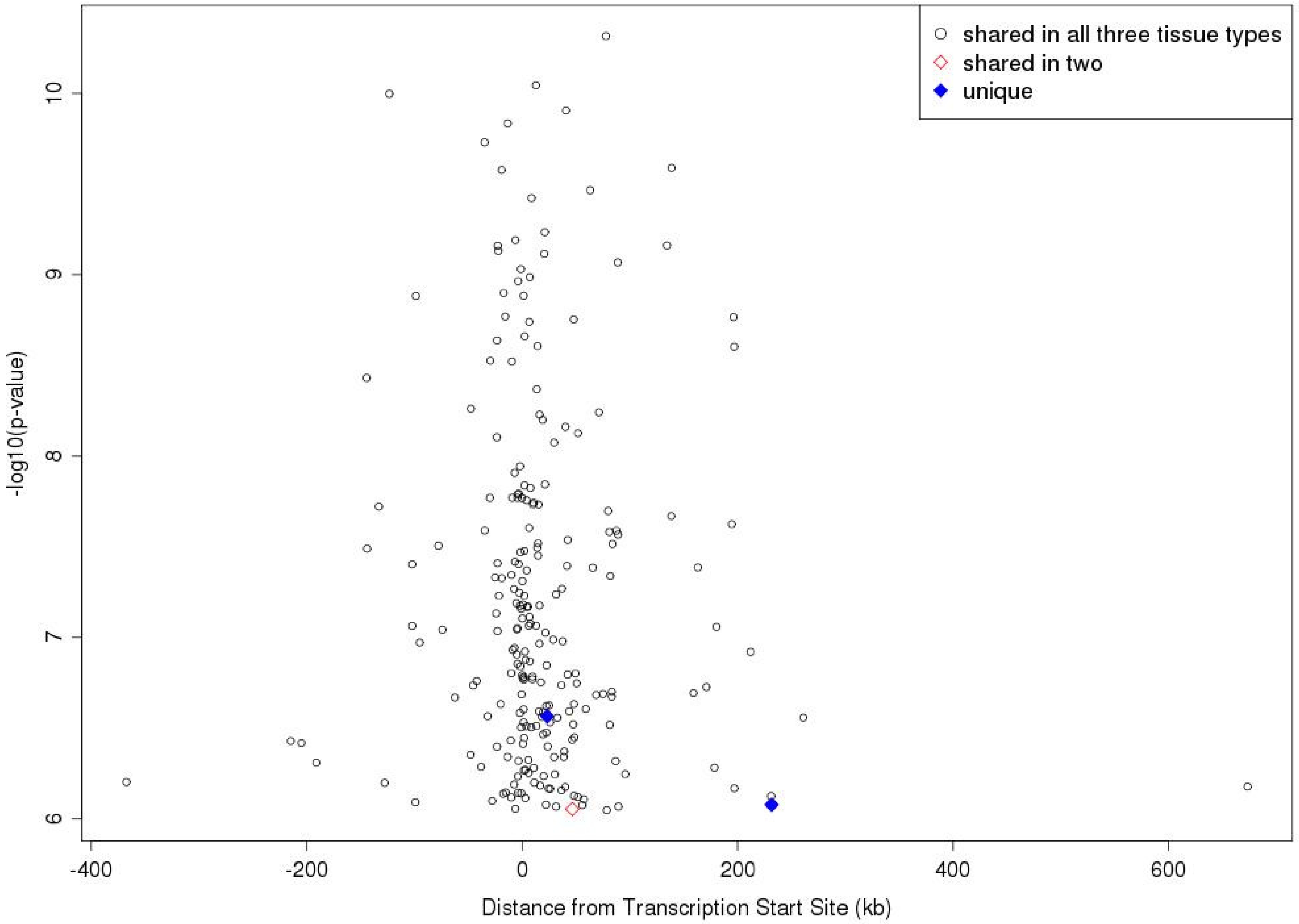
Supplementary Fig 5.1 Localization of the most significant eQTL for each *cis*-association in control, uninvolved, and lesional skin with respect to the transcription start site of the genes they putatively regulate.

Localization of the most significant eQTL for each cis-association in Control Skin




Localization of the most significant eQTL for each cis-association in Uninvolved Skin
○ shared in all three tissue types
◇ shared in two
◆ unique
-log10(p-value)
6
7
8
9
10
-500
0
500
1000
Distance from Transcription Start Site (kb)

Localization of the most significant eQTL for each cis-association in Lesional Skin



Supplementary Table 5.1 Fourteen skin eQTL SNPs are associated with 19 diseases/traits in a catalog of published genome-wide association studies curated by NHGRI.

Disease Trait	Reported Genes	SNPs	Strongest SNP Risk Allele	Risk Allele Frequency	GWAS p-value
Height	TMED10	rs910316	rs910316-?[a]	0.15	1.00E-07
Height	HLA-B	rs13437082	rs13437082-?	0.13	5.00E-08
Folate pathway vitamins	FUT2	rs602662	rs602662-A	0.53	3.00E-20
Body mass index	NEGR1	rs2568958	rs2568958-A	0.58	1.00E-11
Weight	NEGR1	rs2568958	rs2568958-A	0.58	2.00E-08
Body mass index	NEGR1	rs2815752	rs2815752-A	0.62	6.00E-08
Cholesterol, total	DOCK7	rs10889353	rs10889353-C	0.32	4.00E-12
LDL cholesterol	DOCK7	rs10889353	rs10889353-C	0.32	0.000008
Triglycerides	DOCK7	rs1167998	rs1167998-C	0.32	2.00E-12
Triglycerides	ANGPTL3	rs10889353	rs10889353-C	0.33	3.00E-07
Type 1 diabetes	ERBB3	rs2292239	rs2292239-A	NR[b]	3.00E-16
Type 1 diabetes	HLA	rs9272346	rs9272346-G	NR	6.00E-129
Plasma level of vitamin B12	FUT2	rs492602	rs492602-G	0.49	5.00E-17
Height	ANAPC13,CEP63	rs10935120	rs10935120-A	0.33	7.00E-08
Type 1 diabetes	RAB5B, SUOX, IKZF4, ERBB3, CDK2	rs1701704	rs1701704-C	0.35	9.00E-10
Triglycerides	ANGPTL3	rs1748195	rs1748195-C	0.7	2.00E-10
Type 1 diabetes	ERBB3	rs11171739	rs11171739-C	0.42	1.00E-11
Type 1 diabetes	MHC	rs9272346	rs9272346-G	0.61	0
Type 1 diabetes	ERBB3	rs2292239	rs2292239-A	0.34	2.00E-20

[a] ?: A risk allele not reported

[b] NR: not reported

Supplementary Table 5.2 Enrichment of GO term "antigen processing and presentation of endogenous peptide antigen via MHC class I" (GO:0019885) for genes that are associated with *cis*-eQTLs in lesional skin. Results are also shown for control and uninvolved skin, as well as for LCLs.

Tissue Type	List Hits	List Total	Population Hits	Population Total	Fold Enrichment	p-value	Benjamini FDR
Lesional	4	123	6	12954	70.2	1.6E-5	0.04
Control	3	163	6	12954	39.7	0.0022	0.99
Uninvolved	2	137	6	12954	31.5	0.061373	1.00
LCLs	4	960	6	12954	9.0	0.0068	0.78

Chapter 6

Conclusions and Discussion

Transcriptional regulation of gene expression is essential for almost every process in a cell and abnormal transcriptional regulation is likely to be involved in the etiology of many diseases. Therefore, it is important to understand the mechanisms of gene regulation and to dissect the effects of the genetic variants on quantitative levels of gene expression. My dissertation focuses on developing statistical methods for genome-wide association studies of global gene expression (i.e. eQTL analysis).

In Chapter 2, we develop a new statistical method to identify DNA variants that are associated with expression levels of multiple genes, so-called "master regulatory" single nucleotide polymorphisms (SNPs). Our method proposes a new statistic to summarize evidence for association between each SNP and all measured transcripts. This statistic summarizes not just the number of signals that exceed a particular threshold, but also the strength of these signals. In a genome-wide scan, we rank SNPs based on this summary statistic and determine significance by permutation. Simulation studies show that the new summary statistic is more powerful than conventional methods for detecting master regulatory SNPs. As an example, we apply our method to the gene expression and genotype data on 200 lymphoblastoid cell lines. We identify potential master regulators of gene expression, but are not able to replicate them in an independent new data set.

Our new summary statistic is essentially a weighted average of individual association strength (i.e. p-values on a logarithmic scale), with more significant individual associations having larger weights. By pooling the strength of individual associations together, we increase the statistical power of the study. As expected, simulation studies show that our new summary statistic is more powerful than the conventional counting methods in identifying master regulators. However, we note that our weighting scheme is not likely to be the optimal one that generates the most powerful statistic. The optimal weighting scheme for identifying a master regulator will depend on the alternative hypothesis (i.e. how a master regulator is regulating the expression of its associated genes). We will need to perform studies with larger sample sizes so that we are able to identify master regulators and then dissect the mechanisms through which the master regulators influence gene expression.

Estimating accurately the eQTL overlap between two tissues allows researchers to quantify the benefits of studying eQTLs in different tissues. In Chapter 3, we propose a novel method for estimating the eQTL overlap between two tissues. Our multi-step procedure generates a list of potential eQTLs and then uses unbiased estimates for eQTL effect sizes to estimate the expected number of replicating eQTLs for a specific sample size. We can then estimate the proportion of overlapping eQTLs in this context. When applied to compare *cis*-eQTLs detected in analyses of 57 skin biopsies and of 340 lymphoblastoid cell lines, our method shows that ~70% of eQTLs are shared between the two tissues, a much larger proportion than the naive estimate of 30-50%. We also apply the method to another independent data set of Dimas et al. [2009] and get consistent results on the two-tissue overlap percentage. The same method we describe here to compare eQTL sets between tissues could be used to compare eQTL sets between many different groups, including comparisons of eQTL lists between populations, sexes, and cases and controls.

In our effort to obtain unbiased estimates for eQTL effect sizes, we make two assumptions on the distribution of eQTL effect sizes. In Chapter 4, we use simulation studies to show that our method works well when the two assumptions both hold. We further show that even if the two assumptions are moderately violated, the overlap percentage estimation is still reasonably good.

We adopt a sample-splitting strategy for simultaneously identifying significant eQTLs and estimating their effect sizes. In theory, the full version of our method allows the samples to be split arbitrarily into two parts in terms of the sample proportions. From our simulation in Chapter 4, we show that there is usually a set of values for the splitting proportion that gives similarly good estimation of the overlap percentage with regard to the mean squared error.

I intend to pursue in my future work several extensions of the method for estimating overlap between two tissues. I will describe each of these extensions below.

One very natural extension of our method is to adapt it so that we can estimate the eQTL overlap among three tissues. Consistent with the definition of π in estimating the two-tissue eQTL overlap, in the case of three tissues, we define π as the proportion of true eQTLs in Study 1 that are also true eQTLs in Studies 2 and 3. Likewise, we attempt to estimate a power-adjusted expected overlap in significant eQTLs, $\pi_{adjusted}$. It turns out that directly applying in three tissues the same decision-tree idea from Chapter 3 will not work. This is because it will categorize eQTLs into too many groups based on their status in Study 1, Study 2, and Study 3. Instead, we can achieve our goal by doing a "sequential" estimation as follows. By definition,

$$\pi = \pi_{2|1} \times \pi_{3|1,2}$$

where $\pi_{2|1}$ is defined as the proportion of true eQTLs in Study 1 that are also true eQTLs in Study 2, and $\pi_{3|1,2}$ is defined as the proportion of true overlapping eQTLs in Study 1 and Study 2 that are also true eQTLs in Study 3. Similar to the procedure in Chapter 3, we are able to obtain power-adjusted $\hat{\pi}_{2|1}$, and $\hat{\pi}_{3|1,2}$. Therefore we can get:

$$\hat{\pi}_{adjusted} = \hat{\pi}_{2|1} \times \hat{\pi}_{3|1,2}$$

Our procedure will work as follows. We split samples in Study 1 into two halves (Study 1A and Study 1B). We use Study 1A to identify significant eQTLs in Study 1. Then we use Study 2 to get the observed raw overlap percentage between Studies 1 and 2; we use Study 1B to get unbiased effect size estimates for significant eQTLs identified in Study 1A. Based on the above parameter estimates, we can get $\hat{\pi}_{2|1}$ using the formula (2) from Chapter 3. We perform a similar procedure again to get $\hat{\pi}_{3|1,2}$. From above, we get a list of observed significant overlapping eQTLs in Studies 1 and 2. Then we test how many of them can be replicated in Study 3 and obtain the observed raw overlap percentage among Studies 1, 2, and 3. Likewise, we use Study 1B to get unbiased effect size estimates for the significant overlapping eQTLs identified in Studies 1 and 2, and hence the power estimator. Again, the same formula from Chapter 3 can be applied to get $\hat{\pi}_{3|1,2}$. When estimating the effect sizes for overlapping eQTLs among three tissues, we make two similar assumptions as we do in the case of two tissues: Assumption 1) In one study/tissue, the three-tissue overlapping, two-tissue overlapping, and tissue-specific (non-overlapping) eQTLs have the same effect size distribution; Assumption 2) For tissue overlapping eQTLs, their effect size distributions are the same across tissues.

This framework of estimating eQTL overlap among three tissues can be easily adapted to estimating the eQTL overlap among multiple tissues. For example, for four tissues, we can get power-adjusted overlap percentage as

$$\hat{\pi}_{adjusted} = \hat{\pi}_{2|1} \times \hat{\pi}_{3|1,2} \times \hat{\pi}_{4|1,2,3}$$

where each individual component is defined similar as above and can be estimated using procedures similar to the one described in the case of three tissues.

Another potential improvement to our method is to avoid the sample-splitting strategy in the procedure. The sample-splitting strategy applied to Study 1 provides unbiased eQTL effect size estimates, but it leads to the loss of power in identifying significant eQTLs in Study 1 because of the reduction of sample sizes. Instead, we can use all the samples in Study 1 to identify significant eQTLs, and then use resampling techniques (e.g. crossvalidation or the bootstrap) to get better estimation for the eQTL effect sizes. A similar idea was proposed by Sun and Bull [2005] in the context of reducing selection bias in genome-wide studies. Another probable solution is to jointly identify significant eQTLs and estimate their effect sizes using the same samples as in the "winner's curse" work [Xiao and Boehnke 2009; Zollner and Pritchard 2007]. Both Zollner and Prichard [2007] and Xiao and Boehnke [2009] have proposed similar ascertainment-corrected maximum likelihood methods to reduce the bias of effect size estimators in the context of genetic association studies. How these ideas can be applied for our purpose remains to be investigated.

Finally, in Chapter 5, we report the first eQTL map of human skin and identify eQTLs in normal skin, uninvolved skin from psoriatic individuals, and lesional psoriatic skin. Our analysis shows that the vast majority of strong *cis*-eQTLs are shared in the three skin tissue types, which suggests that the physiology of the disease does not change the identity of those strong *cis*-eQTLs. Consistent with recent studies and with the idea that control of gene expression can mediate relationships between genetic variants and disease risk, we found that eQTL SNPs are more likely to be associated with psoriasis than randomly selected SNPs. Our results provide a useful resource for studying regulation of gene expression in skin.

Our study is limited by the sample size (~60) such that only eQTLs with large effect sizes, the majority of which would be *cis*-eQTLs, can be identified. In the future, larger eQTL studies will enable us to study both *cis*- and *trans*- eQTLs and hence provide a more comprehensive profile for the genetics of gene expression in skin. In collaboration with our colleagues, we have started collecting gene expression data from skin tissues using RNA sequencing technology. This new technology is expected not only to provide us with higher quality data, but also to enable us to study the regulation of different isoforms of the same transcript and to obtain a more sophisticated picture of genetic regulation of transcription. It is also of interest to examine how eQTL results from RNA sequencing compare with the results presented here based on the microarray technology.

In summary, in this dissertation, I have addressed several methodological issues in the genome-wide association studies of gene expression, including identifying genetic variants associated with the expression of many genes, estimating the eQTL overlap between two tissues (and potentially extending to multiple tissues). These methods provide practical solutions to the problems rising from the studies on the genetics of gene expression. They can also be adapted to the analysis of other types of high-throughput data. I have also presented results from applying my methods and other available methods to real genetic studies, including the first eQTL map of human skin, which is a useful catalog for examining the functional impact of genetic variants associated with psoriasis. My dissertation work has the potential to provide scientists with useful statistical methods and tools to dissect efficiently the genetics of psoriasis and other complex diseases.

References

Benjamini Y, Hochberg Y. 1995. Controlling the false positive discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc (B) 57:289-300.

Boren T, Falk P, Roth KA, Larson G, Normark S. 1993. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. Science 262(5141):1892-5.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic dissection of transcriptional regulation in budding yeast. Science 296(5568):752-5.

Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ and others. 2007. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet 39(11):1329-37.

Chen WM, Abecasis GR. 2007. Family-based association tests for genomewide association scans. Am J Hum Genet 81(5):913-26.

Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA and others. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37(3):233-42.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. Nat Rev Genet 10(3):184-94.

de Cid R, Riveira-Munoz E, Zeeuwen PL, Robarge J, Liao W, Dannhauser EN, Giardina E, Stuart PE, Nair R, Helms C and others. 2009. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. Nat Genet 41(2):211-215.

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4(5):P3.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M and others. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325(5945):1246-50.

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M and others. 2007. A genome-wide association study of global gene expression. Nat Genet 39(10):1202-7.

Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, Daly MJ, Steinhart AH, Abraham C, Regueiro M, Griffiths A and others. 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314(5804):1461-3.

Ellinghaus E, Ellinghaus D, Stuart PE, Nair RP, Debrus S, Raelson JV, Ding J, Li Y, Weidinger S, Eberlein B and others. 2010. Genome-wide association study identifies a psoriasis susceptibility locus at TRAF3IP2. Nat Genet:in press.

Goring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG and others. 2007. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet 39(10):1208-16.

Goring HH, Terwilliger JD, Blangero J. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69(6):1357-69.

Gudjonsson JE, Ding J, Johnston A, Tejasvi T, Guzman AM, Nair RP, Voorhees JJ, Abecasis GR, Elder JT. 2010. Assessment of the Psoriatic Transcriptome in a Large Sample: Additional Regulated Genes and Comparisons with In Vitro Models. J Invest Dermatol.

Gudjonsson JE, Ding J, Li X, Nair RP, Tejasvi T, Qin ZS, Ghosh D, Aphale A, Gumucio DL, Voorhees JJ and others. 2009. Global gene expression analysis reveals evidence for decreased lipid biosynthesis and increased innate immunity in uninvolved psoriatic skin. J Invest Dermatol 129(12):2795-804.

Hakonarson H, Grant SF, Bradfield JP, Marchand L, Kim CE, Glessner JT, Grabs R, Casalunovo T, Taback SP, Frackelton EC and others. 2007. A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature 448(7153):591-4.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44-57.

Kelly RJ, Rouquier S, Giorgi D, Lennon GG, Lowe JB. 1995. Sequence and expression of a candidate for the human Secretor blood group alpha(1,2)fucosyltransferase gene (FUT2). Homozygosity for an enzyme-inactivating nonsense mutation commonly correlates with the non-secretor phenotype. J Biol Chem 270(9):4640-9.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I and others. 2002. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298(5594):799-804.

Li Y, Willer C, Ding J, Scheet P, Abecasis G. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol:in press.

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. Annu Rev Genomics Hum Genet 10:387-406.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9(5):356-69.

Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, Depner M, von Berg A, Bufe A, Rietschel E and others. 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 448(7152):470-3.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. 2004. Genetic analysis of genome-wide variation in human gene expression. Nature 430(7001):743-7.

Nair RP, Duffin KC, Helms C, Ding J, Stuart PE, Goldgar D, Gudjonsson JE, Li Y, Tejasvi T, Feng BJ and others. 2009. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat Genet 41(2):199-204.

Nichols T, Hayasaka S. 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. Stat Methods Med Res 12(5):419-46.

Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, Roberts RG, Nimmo ER, Cummings FR, Soars D and others. 2007. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39(7):830-2.

Peaper DR, Cresswell P. 2008. Regulation of MHC class I assembly and peptide binding. Annu Rev Cell Dev Biol 24:343-68.

Pesarin F. 2001. Multivariate permutation tests : with applications in biostatistics. Chichester ; New York: J. Wiley.

Plagnol V, Smyth DJ, Todd JA, Clayton DG. 2009. Statistical independence of the colocalized association signals for type 1 diabetes and RPS26 gene expression on chromosome 12q13. Biostatistics 10(2):327-34.

Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW and others. 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39(5):596-604.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. Nat Rev Genet 7(11):862-72.

Saveanu L, Carroll O, Lindo V, Del Val M, Lopez D, Lepelletier Y, Greer F, Schomburg L, Fruci D, Niedermann G and others. 2005. Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. Nat Immunol 6(7):689-97.

Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ and others. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316(5829):1331-6.

Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C and others. 2008. Mapping the genetic architecture of gene expression in human liver. PLoS Biol 6(5):e107.

Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU and others. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316(5829):1341-5.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D and others. 2007. Population genomics of human gene expression. Nat Genet 39(10):1217-24.

Sun L, Bull SB. 2005. Reduction of selection bias in genomewide studies by resampling. Genet Epidemiol 28(4):352-67.

Tanaka T, Scheet P, Giusti B, Bandinelli S, Piras MG, Usala G, Lai S, Mulas A, Corsi AM, Vestrini A and others. 2009. Genome-wide association study of vitamin B6, vitamin B12, folate, and homocysteine blood concentrations. Am J Hum Genet 84(4):477-82.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F and others. 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 39(7):857-64.

Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4(10):e1000214.

Wayne ML, McIntyre LM. 2002. Combining mapping and arraying: An approach to candidate gene identification. Proc Natl Acad Sci U S A 99(23):14903-6.

Wille A, Hoh J, Ott J. 2003. Sum statistics for the joint detection of multiple disease loci in case-control association studies with SNP markers. Genet Epidemiol 25(4):350-9.

Wittkopp PJ. 2007. Variable gene expression in eukaryotes: a network perspective. J Exp Biol 210(Pt 9):1567-75.

Xiao R, Boehnke M. 2009. Quantifying and correcting for the winner's curse in genetic association studies. Genet Epidemiol 33(5):453-62.

Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L. 2003. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35(1):57-64.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G and others. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 40(5):638-45.

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM and others. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 316(5829):1336-41.

Zhang XJ, Huang W, Yang S, Sun LD, Zhang FY, Zhu QX, Zhang FR, Zhang C, Du WH, Pu XM and others. 2009. Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. Nat Genet 41(2):205-210.

Zollner S, Pritchard JK. 2007. Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80(4):605-15.