BLOOD EQTL DETECTION IN STRUCTURED POPULATIONS AND ITS APPLICATION TO INTERPRETATION OF GENETIC ASSOCIATION STUDIES

A Dissertation
Presented to
The Academic Faculty

by

Biao Zeng

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biological Sciences

Georgia Institute of Technology
December 2018

BLOOD EQTL DETECTION IN STRUCTURED POPULATIONS AND ITS APPLICATION TO INTERPRETATION OF GENETIC ASSOCIATION STUDIES

Approved by:

Dr. Greg Gibson, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Patrick McGrath
School of Biological Sciences
Georgia Institute of Technology

Dr. Joe Lachance
School of Biological Sciences
Georgia Institute of Technology

Dr. Jingjing Yang
School of Medicine
Emory University

Dr. Annalise Paaby

School of Biological Sciences
Georgia Institute of Technology

Date Approved: October 31, 2018

Dedicated to my family

ACKNOWLEDGEMENTS

It has been a great and fruitful journey for me in the last four years at Georgia Tech. During this journey, I'm given a huge amount of supports and want to express a lot of gratitude to my mentors and friends. Firstly, I would like to give my sincerest appreciation to my advisor, Prof. Greg Gibson, for his tutorship, guidance in my research career. I have a fantastic time working with Prof. Gibson, and the inspiring discussion with him results in valuable ideas in this thesis. His influences on me will be my forever treasure. Without him, this dissertation would have never been possible.

I would then extend my gratitude to my committee members: Dr. Joe Lachance, Dr. Annalise Paaby, Dr. Patrick McGrath, and Dr.Jingjing Yang. Thanks very much for their valuable comments and suggestions.

Gratitude should also be expressed to members in Gibson's group. In particular, I would like to thank Dr. Urko Marigorta, Dalia Gulick, Swetha Garimalla, Ruoyu Tian, Meixue Duan, Angela Mo, Kiera Berger, Sini Nagpal, for the help, discussions, and collaborations during my Ph.D. study. In particular, I would like to thank Dr. Urko Marigorta for his helping hand at the beginning year when I first join Gibson lab. I thank all my friends at Georgia Tech for their help over the past years.

Finally, I want to express my sincere gratefulness to my family, especially to my wife Qiankun Niu, who have always been there with endless love and support. It is a great fortune for me to be with her in this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

CAGE: Consortium for the Architecture of Gene Expression

CLPP: Co-localization posterior probability

CNV: Copy number variation

DHS: DNaseI hypersensitive site

eQTL: expression quantitative trait locus

FHS: Framingham Heart Study

GTEx: Genotype-Tissue Expression project

GWAS: genome-wide association study

LCL: Lymphoblastoid cell line

LD: Linkage Disequilibrium

PIP: Posterior Inclusion Probability

QTL: Quantitative Trait Locus

SNP: single nucleotide polymorphism

TFBS: Transcription factor binding site

TRS: Transcriptional risk score

SUMMARY

Expression QTL (eQTL) detection has emerged as an important tool for unravelling the relationship between genetic risk factors and disease or clinical phenotypes. Most studies focus on analyses predicated on the assumption that only a single causal variant explains the association signal in each interval. This greatly simplifies the statistical modeling, but is liable to biases in scenarios where multiple linked causal-variants are responsible. Here in this thesis, my primary goal was to address the prevalence of secondary cis-eQTL signals regulating peripheral blood gene expression locally, utilizing two large human cohort studies, each greater than 2,500 samples with accompanying whole genome genotypes. The CAGE dataset is a compendium of Illumina microarray studies, and the Framingham Heart Study (FHS) is a two-generation Affymetrix dataset. I firstly describe performing simulation to reveal the potential interference of causal variants in LD regions. I then also describe a Bayesian co-localization analysis of the extent of sharing of cis-eQTL detected in both studies as well as with the BIOS RNA-seq dataset. Stepwise conditional modeling demonstrates that multiple eQTL signals are present for ~40% of over 3,500 eGenes in both microarray datasets, and that the number of loci with additional signals reduces by approximately two-thirds with each conditioning step. Although fewer than 20% of the peak signals across platforms fine-map to the same credible interval, the co-localization analysis finds that as many as 50%~60% of the primary eQTL are actually shared. Subsequently, co-localization of eQTL signals with GWAS hits detected 1,349 genes whose expression in peripheral blood is associated with 591 human phenotype traits or diseases, including enrichment for genes with regulatory functions such as protein kinase

activity and DNA binding. Just one quarter of these co-localization signals are replicated, further highlighting the technological and methodological barriers to reconciliation of GWAS and eQTL signals. My results are provided as a web-based resource for visualization of multi-site regulation of gene expression and their association with human complex traits and disease states. In addition to the cis-eQTL study, as a member of the eQTLgen consortium, I also conduct trans-eQTL detection in multiple cohorts, including FHS, which contains related individuals, and performed cis-trans eQTL mediation analysis, which I will report as a side project. This thesis provides novel insights into the complexity of gene regulation and the low consistency of fine mapping across studies, and introduces new software, PolyQTL, for co-localization of genetic signals in structured populations.

CHAPTER 1.

INTRODUCTION

1.1 Background of eQTL analysis

Despite low levels of nucleotide diversity in humans, and less than 0.5% amino acid sequence divergence for orthologous proteins between human and chimpanzee (International Human Genome Sequencing Consortium, 2005), there are obvious profound differences across a broad spectrum of phenotypes between these two species. These and other observations highlight the importance of gene regulation (Jacob and Monod 1961), instead of protein function, for phenotypic evolution. Moreover, heritability analysis with twin-studies in the past half century has demonstrated that for many human traits, half or more of the phenotype variance can be explained by genetic factors (Polderman et al. 2015) (Gusev et al. 2014). Over 90% of GWAS hits locate in non-coding regions, indicating that these regions likely manifest their effects through regulation of gene expression (Manolio et al. 2009), consistent with further evidence that variants in the vicinity of DNaseI hypersensitive sites (DHS) capture most of the heritability. In parallel, evolutionary studies of cis-regulatory regions have illustrated that regulatory elements seem to contribute substantially to both adaptive substitutions and deleterious polymorphisms. Thus, understanding the mechanisms that regulate human gene expression is not only crucial for basic biology but also for the interpretation of which polymorphisms at human disease loci are causal (Torgerson et al. 2009).

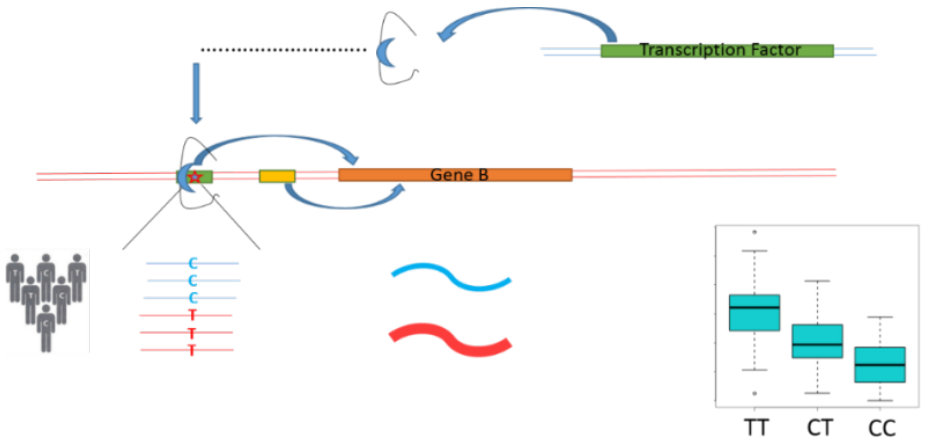

Figure 1.1 Diagram of eQTL and eQTL detection.

Currently, the most direct genome-wide approach to dissect the effect of genetic variation on gene expression is expression Quantitative Trait Locus (eQTL) analysis (GTEx Consortium 2015), as illustrated in Figure 1.1. Expression of gene B is regulated by two regulatory elements, one of which (green rectangle) is a transcription factor binding site, which is bound by a transcription factor in a sequence-specific manner. In a population of individuals, if a site in the TFBS (red star) has two alleles, C and T, then individuals with T, on average, may for example have a higher expression than C. Once the genotype and expression levels for samples from the population have been obtained, we can perform association or linkage analysis to verify the relationship between phenotype and biomarker.

With the advance of next generation sequencing technology, personalized medicine or precision medicine is becoming increasingly prevalent in human health studies. These approaches assume that patients are unique, having their own characteristics and distinct responses to disease or drugs. The goal is to assign patients into different groups according to genetic or genomic biomarker information, which is then used to guide doctors' medical recommendations. I hypothesize that the application of eQTL studies will not only greatly broaden our understanding of personal transcriptomes, but can also be used to improve the

accuracy of personalized medicine. Although current application of personalized medicine is still immature, some studies have already demonstrated the usefulness in personalized medicine by integrating eQTL with GWAS to predict phenotype trait and disease. TWAS predicts gene expression with detected eQTL sites, and regression of these predictions onto phenotype traits may provide greater accuracy than previous genetic risk score methods. Also, in the Gibson lab, an approach named Transcriptional Risk Score (TRS) analysis has been developed in which causal genes for the phenotype trait are first evaluated for joint GWAS and eQTL associations, and then expression levels for the significant genes are summed in order to measure the risk of disease. It has been shown to provide greater prediction power than previous methods for autoimmune diseases such as Crohn's disease (Marigorta et al, 2017).

1.2 Definition of eQTL

eQTL are genomic regions which contain variants contributing to variance in gene expression. Heritability studies partition the sources of phenotype variance in a population into contributions from two broad categories, environmental and genetic factors. Environmental factors, like food resource, life style, development stage, influence all individuals, while micro-environmental stochastic effects are also recognized. Genetic variation is due to all of the DNA sequence polymorphisms occurring on one individual genome. There are multiple kinds of genetic variants, including single nucleotide polymorphism (SNP), copy number variation (CNV), and structural variation (SV). Recently, the importance of epigenetic modifications, namely heritable factors that do not change the DNA sequence but do alter chromatin function, have been recognized as a third important source of variability. Interactions between these categories also contribute, but

are more difficult to detect. eQTL are polymorphisms or mutations that affect regions on the chromosome that are crucial for control of transcript abundance, and hence are identified as regulatory variants. They alter regulatory elements such as promoters and enhancers, influencing the rate of transcription, or affect splice sites leading to alternative splicing, and some mutations change the folding characteristics of transcripts and hence mRNA stability.

1.2.1 Classification of eQTL

Based on biological characteristics, eQTLs can be further classified. According to the distance to the target gene, they may be trans-eQTL or cis-eQTL. Trans-eQTL are located at a different locus, operationally defined as being on a different chromosome than the target gene, or on the same chromosome but a long distance (>1 Mb) from the target gene. cis-eQTL lie in the neighborhood of the target gene, where they are thought to affect gene expression by directly modulating RNA transcription and processing. Technically, cis-eQTL are on the same chromosome of a diploid pair, so the definition based on location within 1 Mb of the Transcription Start Site (TSS) more correctly defines local-eQTL, but the term cis-eQTL is more commonly used in the literature.

A limitation of eQTL detection is the study design. Although the expense of next generation sequencing has greatly reduced, most studies only consist of several hundred individuals. By collecting samples from different cohorts, statistical power can largely be increased, and reduce the incidence of false positives. However, it is almost unavoidable that samples are included from different populations, including perhaps some from European and some from Asian populations. The existence of population structure results

in false positives and in some cases may suppress signals. Control for population structure has become a standard statistical procedure utilizing principal component analysis. Mixed-linear modeling is widely used, and has been shown to be statistically optimal (Loh et al., 2018).

Considering statistical properties of the influence of SNPs on gene expression, there are several ways to classify eQTLs. Most common are additive eQTL, which function in an additive manner where each of the two allele increases or decreases transcript abundance by the same amount. For example, suppose that an eQTL A, which for simplicity I assume to be a bi-allelic variant with alleles M and m, where the M allele increases the gene expression, affects a neighboring gene B. The average increase due to each M allele carried by an individual is fixed, and is called the allelic effect size, or substitution effect. Homozygotes MM have twice as much expression as heterozgygotes relative to mm, irrespective of other alleles. Dominant or recessive eQTL effects are thought to be rare, since each chromosome contributes independently to the total gene expression. Recently, another form of eQTL has also been explored: variance or v-eQTL, in which, instead of increasing mean expression, the allele changes the variance of expression among genotypes. For example, the expression variance of individuals with MM is different from that of individuals with Mm and/or mm genotypes (Metzger et al. 2015; Gusev et al. 2016; Yang et al. 2016). There may be different sources of v-eQTL effects, one of which is epistasis, where the effect of a specific allele is conditional on the allele type at other sites, and this is now thought to be a potential source of unexplained human phenotypic variation. Another one is that variants may work independently to affect the variability of gene expression, perhaps by disrupting the stability of the transcription process. Metzger et al

(2015) investigated how mutation contributes to variance of gene expression in yeast by experimentally determining the effects of polymorphisms segregating in a gene promoter. They found that selection on expression noise resulting from v-eQTL has as large an impact on allele frequency variation as selection on mean expression level.

1.2.2 Detection of eQTL

Current application of cis-eQTL for GWAS interpretation generally requires that the individuals who are genotyped and phenotyped are from the same selected population. Based on the genotyping data (e.g. obtained with SNP arrays) one has to select informative markers, i.e. markers that are polymorphic in the study population. Expression data from DNA microarrays should undergo pre-processing (including background estimation and correction, probe set summary, and normalization) to be suitable for use in the subsequent steps. If other platforms for the expression measurements are used, appropriate pre-processing and data summarization procedures should also be performed prior to eQTL mapping. For the defined cis-region of a specific gene, all variants are associated against the gene's expression in a univariate fashion, and the statistical signals are then compared with GWAS. In cis-eQTL detection, the number of explored variants is hundreds to thousands, and the burden of multiple test correction is relatively small, so the statistical power is still high even with the most conservative correction method, Bonferroni correction. However, trans-eQTL detection requires testing of millions variants, so multiple test correction may obscure all signals, especially if the data set is limited to a few hundred individuals because of the expense and logistical difficulty of acquisition. So, most studies to date have concentrated on mapping cis-acting eQTLs (local to the gene region).

1.3 History of eQTL analysis in humans

The idea of eQTL detection, merging expression with genomic variation, dates back to 2001 (Jansen and Nap, 2001) who called the strategy “genetical genomics”. Just one year later, in 2002, the first eQTL study in yeast was published in Science (Brem et al. 2002), and in the same year, the first eQTL on human lymphoblastoid cell lines (LCL) was also published (Yan et al. 2002). A year later, Schadt et al. and Cheung et al. explored the genetic regulation of gene expression using variance component decomposition, demonstrating significant expression similarity among family members and inferring the existence of widespread contributions of genetic variation (Cheung et al. 2003; Schadt et al. 2003). The breakthrough paper in the human eQTL field was published in 2005 by Cheung et al., in which genome-wide high density SNP variants were linked to gene expression levels measured with microarray technology (Cheung et al. 2005). The first true genome-wide eQTL association study was reported by Stranger et al in 2007, and included comparison of multiple populations.

Since then, eQTL analyses have become prevalent in human genetics and by 2017, more than 100 eQTL studies have been published on human tissues. As shown in Table 1, the main tissues have been whole blood or LCL. In the early stage of eQTL analysis, the studied subjects were gathered from public data repositories such as CEPH, HapMap, or 1000 Genomes, and usually limited samples were available for analysis (usually <300 individuals). Eager to obtain more detailed information about gene expression and human complex traits, researchers have recently conducted more and more eQTL studies in non-blood tissues, including liver and brain, or under different environmental circumstances and across diverse populations. Since 2015, RNA-seq and whole-genome sequencing

technology have enabled new levels of resolution. Here I review eleven breakthrough studies detecting eQTL in human samples, highlighting the main results. All of the studies support the conclusion that genetic components account for a substantial proportion of gene expression variation in humans, and allow me to extrapolate some trends to be expected in human eQTL studies in the coming years.

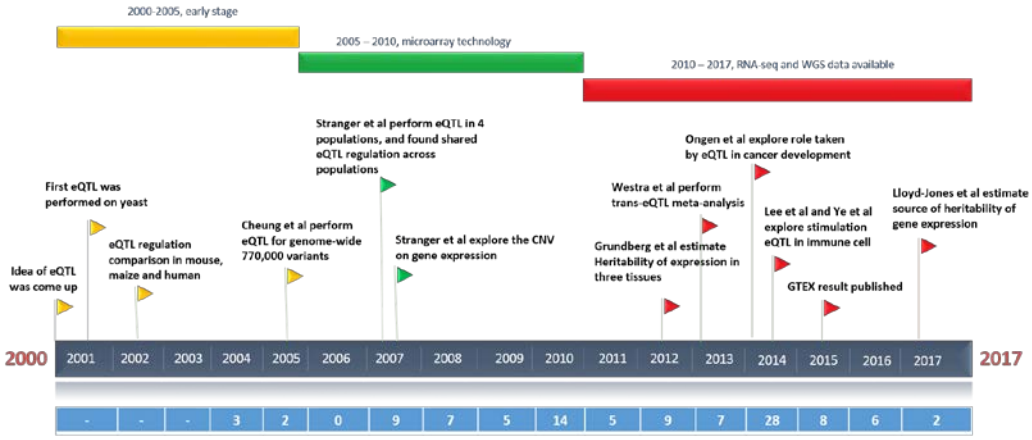

Figure 1.2 Breakthrough eQTL studies in human tissues

1. In the year that the idea of eQTL was proposed, the first eQTL linkage study in yeast was published (Brem et al. 2002). In this study, the genetic basis of expression differences between two strains was explored for 6,215 genes, and although only one quarter of the genes were found to be under overall genetic control, 570 genes were regulated by at least one eQTL. Furthermore, a dozen hot-spots regulating many genes in a trans-manner were described. Results from this study clearly revealed that at the transcriptome level, gene expression is globally regulated by genetic variants.

2. The first study in humans to explore local genetic contributions to gene expression variation was conducted in LCL with the CEPH family data (Yan et al. 2002). These are pedigrees of European-ancestry living in Utah with grandparents, parents, and

up to eight children. Yan et al focused on 13 target genes for which 17-37 heterozygous individuals were available, and measured the relative expression of two alleles in the same individual. Their results showed significant differences in allelic variation for 6 of the 13 genes, revealing expression patterns in humans consistent with Mendelian inheritance, and anticipating later allele-specific expression (ASE) research.

3. In 2003, Schadt et al. performed a comparative analysis of gene expression genetics in mouse, maize and human (Schadt et al. 2003). Linkage analysis was performed to detect eQTL for 23,574 transcripts measured in livers of 111 mice from the F2 generation of a laboratory cross, and approximately 100 microsatellite markers were used to discover that 9-16% of the explored genes have eQTLs with LOD scores > 4.3. Similarly, eQTL analysis of the maize ear leaf identified 26% of transcripts harboring ≥ 1 eQTL with a LOD score > 3.0. Although genotypic data was lacking, Schadt et al. also studied a small number of human LCLs of 56 subjects through variance components analysis and identified differential expression for 11% of the genes assayed, of which about a third had detectable heritability. Overall, these findings provided the first hint of the complexity of the genetic architecture of gene expression across species.

4. Up to this point, human eQTL studies had been performed with only limited genetic markers (≤1000), and were mainly based on linkage analysis. Cheung et al carried out association analysis with dense sets of single-nucleotide polymorphism (SNP) markers from the HapMap Project. For 27 of 374 molecular phenotypes, analysis of genome-wide association was performed with >770,000 SNPs. The association analysis confirmed previous results from linkage studies, and narrowed down the candidate regions, greatly increasing the fine-mapping resolution (Cheung et al. 2005).

5. To explore the characteristics of gene regulation in different human populations, Stranger et al. (Stranger et al. 2007b), investigated LCL from four populations in the HapMap project: 30 Caucasian (CEU) trios, 30 Yoruba (YRI) trios from Nigeria, 45 unrelated Chinese (CHB) and 45 unrelated Japanese (JPT). Analysis of 14,456 transcripts revealed that 10% and 13% of molecular phenotypes in CEU and YRI had heritability > 0.2, respectively, and 958 of the eQTL were discovered in both populations. Furthermore, 6% of explored transcripts had significant eQTLs in at least 1 population at $p < 0.001$; 2% in at least two populations, and 0.4% in all four populations. In order to further characterize the population specificity of gene expression regulation, Spielman et al compared gene expression in three major population groups, and found that at least 25% of the gene were differentially expressed among populations (Spielman et al. 2007), although this result was later found to be largely confounded by batch effect, so the divergence estimate had been much inflated (Akey et al. 2007). Idaghdour et al evaluated the eQTL regulation in whole blood from Arab and Amazigh, and found that most of the eQTLs had consistent signals, verifying the shared *cis*-genetic influences on gene regulation (Idaghdour et al. 2010b). Despite the small sample size, these eQTL results coherently demonstrated the sharing of genetic factors in gene expression regulation across different human populations.

6. Revealing the genetic component of gene expression across human tissues, Grundberg et al (2012) presented a comprehensive analysis of gene expression in LCL, skin, and adipose. They calculated heritability and evaluated the genetic and non-genetic contributions to expression with a large sample of mono- and dizygotic twins, concluding that at least 40% of the total heritable cis effect on expression may originate from rare

variants, while a substantial proportion of gene expression heritability must be due to trans-acting genetic factors.

7. In contrast to previous eQTL studies mainly exploring contributions of SNPs, Stranger et al investigated the effects of another type of genetic variation, copy number variants (CNVs), on gene expression levels. They performed association for 14,925 transcripts with CNVs in the HapMap populations and determined that there were significant CNV associations that replicate across ethnic groups as well as some that are unique to single populations (Stranger et al. 2007a). Most CNV associations were independent of proximal SNPs, highlighting the importance of structural variants in addition to SNPs in the regulation of gene expression.

8. Recently, several groups have begun to investigate the genetic basis for differences among individuals in gene expression in different contexts, such as the immune response to stimulation. These effects have been termed response-eQTL, and were first explored in the nematode, *Caenorhabditis elegans* (Li et al. 2006). The first human response-eQTL study described by Lee et al using dendritic cells (DCs) derived from monocytes of healthy individuals, with 295 Caucasians, 122 African Americans, 117 East Asians. The DC were stimulated ex vivo with lipopolysaccharide (LPS), influenza virus, or the cytokine interferon-β (IFN-β) (Lee et al. 2014). Common variants were genotyped and correlated with gene expression from each condition, and then synthetic promoter constructs and genome engineering were applied to experimentally confirm some of the detected associations. A nanostring array was used to measure 1,598 molecular phenotypes, of which 264 were shown to be regulated by genetic variants associated with gene expression in human DCs. Signals from 121 of the loci were uniquely influenced by

stimuli. Co-localization analysis also revealed that 35 of the eQTLs were likely to affect autoimmune phenotypes through alteration of gene expression.

9. Most large human eQTL studies have been performed with healthy subjects rather than directly in the context of disease. Ongen et al conducted an RNA-sequencing experiment involving 103 matched tumor and normal colon mucosa samples from Danish colorectal cancer (CRC) patients, of which 90 were germline-genotyped (Ongen et al. 2014). Correlation of genotypes with gene expression found 1,693 and 948 eQTLs in normal samples and tumor samples, respectively. They estimated that 36% of the tumor eQTLs are cancer-specific, partially driven by altered expression of specific transcription factors and changes in methylation patterns. The authors also found that tumor-specific eQTLs were more enriched for low CRC genome-wide association study (GWAS) P values than shared eQTLs, which implies that some of the GWAS variants are tumor specific regulatory variants. Importantly, compared to genes with shared eQTLs, genes with tumor-specific eQTL tended to accumulate more allele-specific expression, indicating that somatically-derived mutations may constitute cancer regulatory drivers.

10. To assign the contribution of cis- and trans-eQTL to overall expression variance, Lloyd-Jones et al analyzed the mRNA levels for 36,778 probes in 2,765 individuals, and investigated the genetic architecture of gene expression in peripheral blood (Lloyd-Jones et al. 2017). 11,204 cis and 3,791 trans independent expression quantitative trait loci (eQTL) were detected. For expressed probes (15,966), 66% had a non-zero narrow-sense heritability, the mean estimate of which was 0.192, 31% of which was assigned to detected eSNPs, while 69% remained missing. The evidences suggests that approximately half the genetic variance for gene expression is not tagged by common

SNPs, possibly indicating a crucial role for rare variants on gene expression. This also has implications for the likely evolutionary origin of the variance that is tagged by common SNPs, a large proportion of which can be attributed to identifiable eQTL of large effect, typically located in cis. Limited by the small statistical power and great burden of multiple test correction, trans-eQTL analysis remains challenging. Westra et al (2013) performed expression quantitative trait locus (eQTL) meta-analysis in 5,311 individuals with replication in 2,775 individuals. 233 were SNPs associated with complex traits from 103 independent loci and replicated as trans-eQTL. Among these trans effects, an excess of variants associated with cholesterol metabolism and type 1 diabetes were found to alter the expression of multiple genes known to be associated with traits, revealing likely regulatory mechanisms for the downstream effect of many trait-associated variants.

11. Limited by the availability of tissues from humans, most studies have explored only two or three, if not a single, tissue. To have a broader understanding of transcriptome regulation across human tissues, and to elucidate the functional consequences of genetic variation associated with complex human disease and quantitative traits, the Genotype-Tissue Expression (GTEx) project was launched. Ardlie et al. (2015) presented the first analysis of RNA sequencing data from 1641 samples derived from 43 biopsy tissues from 175 recently deceased individuals. In this analysis, it was found that an average of 20,940 genes were expressed in the explored tissues, highest in testis and lowest in whole blood. A U-shaped distribution of the number of tissues expressing each gene was observed, meaning that most genes are expressed either in most or just a few tissues. eQTL analysis was performed on 9 tissues with largest sample size. A total of 10,130 eQTLs were detected, 50% of which were shared by all of the 9 tissues, showing high concordance of

effect directions. The detected eQTLs were enriched for trait associations, many taking function in a tissue-dependent manner. Co-localization analysis demonstrated that ~6% GWAS-SNPs were in the same LD block as the detected eSNP, suggesting that their influence on the trait is mediated through the eQTL effect.

Table 1.1. eQTL studies in human tissues since 2004.

Cohort	Tissue	Year	Sample size	Reference	Note
CEPH	blood	2004	210	Monk et al(MONKS *et al.* 2004)	346 markers explored
CEPH	LCL	2004	94	Morley et al(MORLEY *et al.* 2004)	2,756 autosomal SNPs
CEPH	LCL	2005	57	Cheung et al(CHEUNG *et al.* 2005)	
Hapmap	LCL	2005	60	Stranger et al(STRANGER *et al.* 2005)	
Hapmap	LCL	2007	238	Stranger et al(STRANGER *et al.* 2007a)	
Hapmap	LCL	2007	270	Stranger et al(STRANGER *et al.* 2007b)	
SAFHS	LCL	2007	1,280	Göring et al(GÖRING *et al.* 2007)	
Hapmap_CEU	LCL	2007	30	Stranger et al(STRANGER *et al.* 2007b)	
Hapmap_CHB	LCL	2007	45	Stranger et al(STRANGER *et al.* 2007b)	
Hapmap_JPT	LCL	2007	45	Stranger et al(STRANGER *et al.* 2007b)	
Hapmap_YRI	LCL	2007	30	Stranger et al(STRANGER *et al.* 2007b)	
BR	Brain_Cortex	2007	193	Myers et al(MYERS *et al.* 2007)	
SIGN	LCL	2007	206	Dixon et al(DIXON *et al.* 2007)	
Sorbs	blood	2008	948	Tönjes et al(TÖNJES *et al.* 2010)	
IFA/IFB	blood/subcutaneous adipose	2008	673/1002	Emilsson et al(EMILSSON *et al.* 2008)	
Hapmap	LCL	2008	269	Choy et al(CHOY *et al.* 2008)	
Hapmap_CEU/YRI	LCL	2008	209	Price et al(PRICE *et al.* 2008)	
LV	Liver Cell	2008	427	Schadt et a(SCHADT *et al.* 2008)	
Hapmap_CEU	LCL	2008	60	Duan et al(DUAN *et al.* 2008)	
Hapmap_YRI	LCL	2008	69	Duan et al(DUAN *et al.* 2008)	
InChianti	blood	2009	705	Tanake et al(TANAKA *et al.* 2009)	
LOAD	Brain	2009	765	Webster et al(WEBSTER *et al.* 2009)	
3C	LCL	2009	75	Dimas et al(DIMAS *et al.* 2009a)	
3C	Fibroblast	2009	75	Dimas et al(DIMAS *et al.* 2009a)	
3C	T_cell	2009	75	Dimas et al(DIMAS *et al.* 2009a)	
Morocco	blood	2010	203	Idaghdour et al(IDAGHDOUR *et al.* 2010a)	
DILGOM	blood	2010	631	Inouye et al(INOUYE *et al.* 2010)	
Hapmap	LCL	2010	69	Pickrell et al(PICKRELL *et al.* 2010)	
GHS	Monocyte	2010	1490	Zeller et al(ZELLER *et al.* 2010)	

Hapmap_CEU	LCL	2010	60	Montgomery et al(MONTGOMERY *et al.* 2010)	
Hapmap_YRI	LCL	2010	69	Pickrell et al(PICKRELL *et al.* 2010)	
BR2	Cerebellum	2010	150	Gibbs et al(GIBBS *et al.* 2010)	
BR2	Frontal_cortex	2010	150	Gibbs et al(GIBBS *et al.* 2010)	
BR2	Temporal_cortex	2010	150	Gibbs et al(GIBBS *et al.* 2010)	
BR2	Pons	2010	150	Gibbs et al(GIBBS *et al.* 2010)	
Psoriatic	Skin	2010	57	Ding et al(DING *et al.* 2010)	
	Liver	2010	960	Musunuru et al(MUSUNURU *et al.* 2010)	rs12740374
	Subcutaneous fat	2010	433	Musunuru et al(MUSUNURU *et al.* 2010)	rs12740374
	Omental fat	2010	520	Musunuru et al(MUSUNURU *et al.* 2010)	rs12740374
Fehrmann	blood	2011	1469	Fehrmann et al(FEHRMANN *et al.* 2011)	
LV2	Liver cell	2011	266	Innocenti et al(INNOCENTI *et al.* 2011)	
	trabecular bone	2011	113	Grundberg et al(GRUNDBERG *et al.* 2011)	
	Liver	2011	266	Innocenti et al(INNOCENTI *et al.* 2011)	
	brain	2011	269	Colantuoni et al(COLANTUONI *et al.* 2011)	
SHIP-Trend	blood	2012	653	Mehta et al(MEHTA *et al.* 2013)	
BSGS	blood	2012	962	Powell et al(POWELL *et al.* 2012)	
IM	Monocyte	2012	288	Fairfax et al(FAIRFAX *et al.* 2012)	
IM	B_cell	2012	288	Fairfax et al(FAIRFAX *et al.* 2012)	
MuTHER	LCL	2012	160/856	Grundberg et al(GRUNDBERG *et al.* 2012)	
MuTHER	Skin	2012	160/856	Grundberg et al(GRUNDBERG *et al.* 2012)	
MuTHER	Adipose	2012	160/856	Grundberg et al(GRUNDBERG *et al.* 2012)	
	Lung	2012	1111	Hao et al(HAO *et al.* 2012)	
	Cortex, cerebellum	2012	400	Zou et al(ZOU *et al.* 2012)	
Rotterdam	blood	2013	881	Hofman et al(HOFMAN *et al.* 2013)	
EGCUT	blood	2013	734	Metspalu et al	
KORA F3/F4	blood	2013	322/740	Mehta et al(MEHTA *et al.* 2013)	
GEUVADIS	LCL	2013	462	Lappalainen et al(LAPPALAINEN *et al.* 2013)	
MRCA, MRCE	LCL	2013	405/950	Liang et al(LIANG *et al.* 2013)	
E-GEUV	LCL	2013	373	Lappalainen et al(LAPPALAINEN *et al.* 2013)	
E-GEUV	LCL	2013	89	Lappalainen et al(LAPPALAINEN *et al.* 2013)	
Cardiology	blood	2014	338	Kim et al(KIM *et al.* 2014)	
Bangladeshi	blood	2014	1800	Pierce et al(PIERCE *et al.* 2014)	
CHDWB	blood	2014	189	Preininger et al(PREININGER *et al.* 2013)	
SIGN	blood/neutrophil	2014	114	Andiappan et al(ANDIAPPAN *et al.* 2015)	
ALSPAC	blood	2014	869	Bryois et al(BRYOIS *et al.* 2014)	
NTR-NESDA	blood	2014	2752	Wright et al(WRIGHT *et al.* 2014)	
BLD	blood	2014	1387	Tigchelaar et al(TIGCHELAAR *et al.* 2015)	
DGN	blood	2014	922	Battle et al(BATTLE *et al.* 2014)	
ImmVar	CD4+	2014	407	Raj et al(RAJ *et al.* 2014a)	

ImmVar	Monocyte	2014	401	Raj et al(RAJ *et al.* 2014a)	
ImmVar	$CD14^{+}CD16^{lo}$ monocytes	2014	534	Lee et al(LEE *et al.* 2014)	
ImmVar	LPS induced $CD14^{+}CD16^{lo}$ monocytes	2014	534	Lee et al(LEE *et al.* 2014)	
ImmVar	induced by influenza virus	2014	534	Lee et al(LEE *et al.* 2014)	
ImmVar	$CD14^{+}CD16^{lo}$ monocytes induced by IFN-b	2014	534	Lee et al(LEE *et al.* 2014)	
ImmVar	$CD4^{+}$ 4h α328	2014	348	Ye et al(YE *et al.* 2014)	236 gene explored
ImmVar	$CD4^{+}$ 48h α328	2014	348	Ye et al(YE *et al.* 2014)	
ImmVar	$CD4^{+}$ 4h IFNβ	2014	348	Ye et al(YE *et al.* 2014)	
ImmVar	$CD4^{+}$ 48h IL-6, TGFβ	2014	348	Ye et al(YE *et al.* 2014)	
	colorectal cancer	2014	103	Ongen et al(ONGEN *et al.* 2014)	
	breast cancer	2014	415/407	Li et al(LI *et al.* 2013)	
HGVD	blood	2014	298	Narahara et al(NARAHARA *et al.* 2014)	
	Dendritic Cell	2014	534	Lee et al(LEE *et al.* 2014)	1598 transcriptional profile
	Stimulated monocytes	2014	432	Fairfax et al(FAIRFAX *et al.* 2014)	
	Heart	2014	129	Koopmann et al(KOOPMANN *et al.* 2014)	
	10 brian regions	2014	134	Ramasamy et al(RAMASAMY *et al.* 2014)	
	skeletal muscle	2014	45	Lindholm et al(LINDHOLM *et al.* 2014)	
	islets	2014	89	Fadista et al(FADISTA *et al.* 2014)	
PRAX1	platelet	2014	154	Simon et al(SIMON *et al.* 2014)	
YoungFinns	blood	2015	1428	Turpeinen et al(TURPEINEN *et al.* 2015)	
LIFE	blood	2015	2107	Burkhardt et al(BURKHARDT *et al.* 2015)	
Framingham	blood	2015	5626	Huan et al(HUAN *et al.* 2015)	
GTEX	blood/other tissue	2015	420	GTEx Consortium(CONSORTIUM 2015a)	
	prostate	2015	565	Thibodeau et al(THIBODEAU *et al.* 2015)	
CartaGene	blood	2015	521	Hussin et al(HUSSIN *et al.* 2015)	
	Islets	2015	118	Bunt et al(VAN DE BUNT *et al.* 2015)	
	6 immune cell type	2015	91/46/43	Peters et al(PETERS *et al.* 2016)	
	skeletal muscle	2016	267	Scott et al(SCOTT *et al.* 2016)	
TwinsUK	LCL/adipose/skin	2016	845	Hore et al(HORE *et al.* 2016)	
	macrophage	2016	168	Nédélec et al(NÉDÉLEC *et al.* 2016)	
	Monocyte	2016	200	Quach et al(QUACH *et al.* 2016)	
	Whole blood	2016	377	Walsh et al(WALSH *et al.* 2016)	
	LCL	2016	786	Peterson et al(PETERSON *et al.* 2016)	
METSIM	Adipose	2017	770	Civelek et al(CIVELEK *et al.* 2017)	
	CD4+/CD8+	2017	293/283	Kasela et al(KASELA *et al.* 2017)	

1.4 Interpretation of eQTL

The primary detection of eQTL is methodologically straightforward: gene expression is measured from hundreds of individuals usually by microarray or RNA-seq, genotypes are obtained and imputed by reference to an appropriate population of known haplotypes, and then statistical analysis, either association or linkage studies, are conducted to evaluate the relationship between the variance of phenotype (expression) and the identities of the genotypes. For a given statistical confidence interval, either based on conservative Bonferroni adjustment or at a False Discovery Rate threshold, a variant is determined to be associated with the phenotype (Figure 1). However, the interpretation of eQTL results is generally not so simple. In biology, a cis eQTL is usually expected to be a regulatory element that contributes to gene expression, while, for real data, what is discovered is just an interval on the chromosome within which it is inferred that a causal variant exists. Because of the existence of extensive LD structure in human populations, it is hard to distinguish which variant is causal, or whether or not there are multiple causal variants in a single region. It should be appreciated that eQTL performs association analysis, and association doesn't always imply causality. Additionally, whole blood, the most widely used tissue, is actually a mixture of various cell types, and the estimation of cell-type-specific eQTL is unavoidably highly biased.

The general and parsimonious assumption is that functional variants are sparsely distributed, and hence that their precise localization or estimation of effect sizes is not affected by interference due to confounding of statistical signals. However, as genome-wide association studies have increased in size it has become clear that multi-site effects are not uncommon. For example, the latest meta-analysis of height suggests that over one

third of the more than 400 identified loci have multiple independent signals (Wood et al. 2014), and similarly in the transcriptome literature, the expression of a large proportion of genes in lymphocyte cell lines has been shown to be regulated by two or more locally-acting variants (cis-eQTL) (Liang et al. 2013). Gusev et al (Gusev et al. 2013) observed that all SNPs at known GWAS loci can explain 1.29-fold more heritability than GWAS-associated SNPs on average, and Lloyd-Jones et al (2017) revealed that in peripheral blood ~23% transcripts are regulated by multiple independent eQTLs. The same situation exists in eQTL analysis as well, and lies behind the strategy of predicting gene expression from all SNPs within 1 Mb of each gene (Mancuso et al. 2017). Thus, determining the effect of these multiple variant SNPs on target transcript levels gives considerably more detail concerning the complex regulatory interactions at a locus. Including epigenetic markers and enhancer-gene interaction information, Corradin et al (2014) revealed specific cases where several variants in LD simultaneously affect gene expression. Applying Bayesian methods and incorporating genomic annotations, Wen et al (2015) identified multiple cis-eQTL signals for ~12% genes with eSNP. Since linkage disequilibrium within a locus can be extensive, the potential for mis-estimation of eQTL effects due to interference between signals from tightly linked polymorphisms is high.

Researchers have also found enrichment for cell type-specific eQTLs among disease susceptibility alleles. Dimas et al (2009b) detected 69 to 80% of regulatory variants operating in a cell type–specific manner. Raj et al found ~40% eQTL functioning in a cell-specific manner, and documented over-representation of T cell–specific eQTLs among susceptibility alleles for autoimmune diseases and of monocyte-specific eQTLs among Alzheimer's and Parkinson's disease variants (Raj et al. 2014b). Using gene expression as

a proxy for cell count percentage, Westra et al (2015) inferred cell-type specific effects from whole blood data, also demonstrating that SNPs associated with Crohn's disease preferentially affect gene expression within neutrophils.

1.5 Interpreting GWAS with eQTL results

Since the first GWAS results were published in 2005 (Klein et al. 2005), several thousand genetic regions on human chromosomes have been found to be associated with human phenotypes including disease states. Since it is now assumed that the majority of SNP-trait associations identified by GWAS can be attributed to effects on gene expression, precise estimation of the location and effect sizes of regulatory polymorphisms has become important for understanding the relationship between genetic and phenotypic variation (Maurano et al. 2012a; Farh et al. 2015). Expression quantitative trait locus analysis and related functional genomic strategies are thus now a standard component of genetic fine mapping (Nicolae et al. 2010). The minimal expectation is that they can identify the gene within a locus that accounts for a GWAS signal, although even this is a far from trivial undertaking (Chung et al. 2014; Pickrell 2014). Many investigators make the stronger assumption that co-localization of eSNP and GWAS signals to a tight linkage disequilibrium interval implies the ability to define if not the causal variant, then at least a credible set of SNPs that include the causal site (Trynka et al. 2013; Gaulton et al. 2015; Kichaev and Pasaniuc 2015; Liu et al. 2015). The strong enrichment of chromatin marks such as DNAse Hypersensitive Sites (DHS) in the vicinity of eQTL validates this assumption (ENCODE Consortium, 2012; Roadmap Epigenomics Consortium, 2015).

Table 1.2. Experimentally verified GWAS hits

Disease/Trait study	Gene	eQTL	tissue	Reference
Asthma	ORMDL3	YES	LCL	Moffatt et al, 2007(MOFFATT *et al.* 2007)
Blood lipid level	SORT1	YES	Liver	Musunuru et al, 2010(MUSUNURU *et al.* 2010)
Blood lipid level	PPP1R3B	YES	Liver	Teslovich et al, 2010(TESLOVICH *et al.* 2010)
Blood lipid level	TTC39B	YES	Liver	Teslovich et al, 2010(TESLOVICH *et al.* 2010)
Breast cancer	RRP1B	YES	PyMT-induced primary tumours	Crawford et al, 2007(CRAWFORD *et al.* 2007)
Chronic lymphatic leukaemia	ASPM	UNKNOWN		Horvath et al, 2006(HORVATH *et al.* 2006)
ventricular conduction system	SCN10A	UNKNOWN		Sotoodehnia et al, 2010(SOTOODEHNIA *et al.* 2010)
Parkinson	SNCA	YES	Postmortem frontal cortex	Soldne et al, 2016(SOLDNER *et al.* 2016)
T2D	IRX3/IRX5	YES	Adipocyte	Claussnitzer et al, 2015(CLAUSSNITZER *et al.* 2015)

However, high resolution fine mapping eQTL results aligned with GWAS studies for diverse phenotypes has as yet provided only a few instance of site-specific evidence that variants affecting human complex traits and diseases function through their effect on gene expression. Table 1.2 lists nine experimentally verified GWAS hits which have been validated to affect the phenotype. There are two prominent examples. The first one is a common SNP at 1p13, a locus associated with the risk of myocardial infarction (Musunuru et al. 2010). This SNP is found to be located in the 3′ untranslated region of a gene, and the minor allele creates a binding site for a transcription factor (TF) that is preferentially expressed in the liver, as a consequence of which, the target gene sortilin 1 (SORT1) is upregulated specifically in the liver. Knockdown studies in mouse liver confirmed that higher expression of the sortilin protein results in lower levels of low-density lipoprotein cholesterol (LDL-C), which is associated with higher risk of myocardial infarction. The second example is the SNCA gene. Using the CRISPR/Cas9 genome editing method, Soldner et al (2016) identified a common Parkinson's disease (PD)-associated risk variant

in a non-coding distal enhancer element that regulates the expression of alpha-synuclein (SNCA), a key gene implicated in the pathogenesis of PD. The results suggest that the transcriptional deregulation of SNCA is associated with sequence-dependent binding of a brain-specific TF. Both of these GWAS hits were found to be associated with gene expression, and affect TF binding ability, which complies with biological expectations.

On the other hand, several recent studies have begun to question the presumed identity of eQTL and GWAS hits: even though there is a highly significant overlap at the level of the locus (Maurano et al. 2012), it is not so clear that the precise variants are the same. Farh et al (2015) integrated regulatory elements and GWAS results, and estimated that only ~10% of the GWAS hits function as eQTL, and a more recent comprehensive study of autoimmune disease also argued that only one quarter of examined GWAS loci may act as discovered eQTL (Chun et al. 2017). Similarly, work based on GTEx gene expression aiming to integrate GWAS and eQTL results concluded that only a minority of GWAS loci match eQTL (Hormozdiari et al. 2016). These results from statistical analysis raise the question of why there are so many instances of discordant fine localization: are we simply limited by the low statistical power to detect association signals (Udler et al. 2010); is there mis-estimation of signal strength and location in the case of multiple eQTL per transcript; or are regulatory effects so cell-type and context-specific that true co-localization is often missed? From the opposite perspective, regulatory sites may often be selectively neutral due to small probability of affecting phenotypes, and hence do not appearing in low-powered GWAS scans.

1.6 Constraints on fine-mapping resolution due to LD

Linkage disequilibrium (LD) is a phenomenon whereby alleles at closely linked sites tend to be inherited together (Pritchard and Przeworski 2001). Several evolutionary factors, including demography, population structure, recombination, mutation, and natural selection, create, shape and modify the rate of decay of LD. Consider two loci, A and B, in an ancestral population, where site A is a bi-allelic variant, with M and m are the major and minor allele, respectively, and there is only a single allele N for B. At some specific time or place, a mutation happens in site B in a person with M in site A, and this allele spreads into the population. Initially the new site n is only found on the M chromosome, so there is complete LD between M and n. As time goes by, recombination occurs, forming the haplotype with m and n. Eventually, n is just as likely to be on the M and the m chromosomes, at which point linkage equilibrium is reached. Also, admixture between two populations creates temporary LD at loci throughout the genome. By definition, a straightforward method to measure the LD strength is to evaluate $D = P_M P_N - P_{MN}$, where P_M is the allele frequency of allele M in site A, and P_N the allele frequency of allele N in site B, P_{MN}, the haplotype frequency of MN. In the case of linkage equilibrium, alleles in different loci segregate randomly, resulting in $P_{MN} = P_M P_N$, and D=0. In the presence of LD, D does not equal zero. Since D calculated with the above formula has the disadvantage that it is largely affected by allele frequencies, making it difficult to compare the level of linkage disequilibrium between different pairs of loci, two alternative measures of LD have been devised to correct this: D' and r^2. D' is a normalized D divided by the theoretical maximum difference between the observed and expected allele frequencies as follows: $D' = D/D_{min}$, $D_{min}=\max(-P_M P_N, -(1-P_M)(1-P_N))$, when $D<0$, and $D_{min}=\min(P_M(1-P_N), (1-P_M)P_N)$, when $D>0$. Before recombination, D' is always 1, and in the presence of recombination,

D’ decays gradually. An alternative to D' is the correlation coefficient between pairs of loci, $r=D/sqrt(P_M(1-P_M)P_N(1-P_N))$ (Hill and Robertson 1968).

In the human genetics field, the Out of Africa hypothesis is the most prominent model used to explain the evolutionary history of anatomically modern humans. According to this model, modern humans dispersed from East Africa to the Eurasian landmass and other continents starting 100,000~60,000 years ago (Mellars 2006; Mallick et al. 2016). Consequently, populations of European- and Asian-ancestry are much younger than African ones. There has been less time to break down LD resulting from sub-sampling of a fraction of human diversity during the population bottleneck that occurred during dispersal. Consequently, Europeans and Asians tend to have longer segments of LD than Africans, and non-causal variants typically locate in long LD blocks along with causal ones. Consider the scenario where a causal variant, C, affects a continuous phenotype, and assume a minor allele frequency P_C, and allelic effect size β_C. When performing an association study to evaluate the relationship between each genetic variant and the phenotype, for non-causal variants in LD with a causal variant, the effect size estimate is $\dot{\beta}=r^* \beta_C$, where r is the correlation coefficient between causal and non-causal variants. In the presence of extensive LD in the study population, r tends to be large for tens to hundreds of linked sites, and consequently estimation for causal and non-causal variants is similar, greatly reducing fine-mapping resolution. Even more disturbing, when multiple causal variants exist, interference between these variants will bias the estimation. Considering another scenario, where two causal variants affect the phenotype together, estimation for any variant will be $\dot{\beta}= r_1^* \beta_{C1}+ r_2^*\beta_{C2}$, where r_1 is the genotypic correlation of the explored variant to causal variant 1, and r_2 to causal variant 2. When the causal variants function in

opposing directions, there is a strong chance that $\dot{\beta}$ equals or is near to 0, which means the causal variant will be mis-labelled as non-causal. When the causal variants function in the same direction, it is possible that $\dot{\beta}$ for the non-causal site is greater than the true βc for either causal site, and the incorrect inference will be made that there is a single eSNP, whose location and effect size will be mis-specified.

1.7 Comparison of Frequentist and Bayesian tools for eQTL estimation

Assuming that multiple eQTL contribute to expression variance at a locus, diverse methods have been developed to detect multiple independent eQTLs. Conditional analysis is the most widely used method to detect independent signals in GWAS (Yang et al. 2012). It performs step-wise detection, namely, the novel signal is found conditioning on the effects of previously detected signals. In the scenario that multiple causal variants locate in different LD blocks, conditional analysis selects one variant as a tag for each LD block. In contrast, Bayesian-based methods are designed to perform association and fine-mapping simultaneously. In the Bayesian framework, to select variables as causal, the approach is to comprehensively survey the causal status space under a set of prior assumptions, and then use posterior probability values to evaluate the importance of each combination of variants. Several Bayesian methods have been developed, and there are two categories: one is used when individual genotype and phenotype are available, and the other one is applied when only marginal summary results are available. FMQTL (Wen et al. 2015) and DAP (Wen et al. 2016) use each individual's genotype and phenotype to detect multiple eQTL. FMQTL applies the Metropolis Hasting algorithm in the MCMC method to sample the causal states, and calculates a posterior inclusion probability from the marginal sampled

causal states. To reduce the computational burden, Wen et al developed DAP, in which only the causal status associated with high-probability variants were explored instead of sampling. Another package, CAVIAR (Hormozdiari et al. 2014), was developed to deal with summary statistics. In CAVIAR, the marginal statistical association results and LD structure are used to explore the probability of each causal status. CAVIAR-BF (Chen et al. 2015), PAINTOR (Kichaev et al. 2014), and FINEMAP (Benner et al. 2016) were developed with similar logical models. One disadvantage of most of these Bayesian methods is that they output a list of SNPs with a specific confidence that causal variants are included or a set of causal statuses with high probability, but there is no estimation for the parameters of the underlying causal variants, such as effect size betas. To estimate these parameters, we need to choose the causal status with the largest probability, and perform multiple-variable regression to estimate the allelic effect sizes.

In this thesis, I first describe simulation studies which reveal that in the scenario of multiple eQTL regulation, interference between causal variants results in greatly biased estimation, and then develop a statistical model to identify multiple regulatory variants affecting gene expression, combining both frequentist and Bayesian methods. To integrate eQTL signal to interpret GWAS results, I then develop a new joint mapping method to evaluate the co-localization of eQTL-GWAS signal, and use it to identify causal genes and causal variants for human complex traits and disease. In addition, as a member of Prof Lude Franke's eQTLgen consortium (University of Groningen, The Netherlands), I also report on methods that I used for trans-eQTL detection, comparing results across gene expression platforms.

1.8 Thesis Structure

Specific Aims: Statistical Dissection of the Regulation of Gene Expression and its application in interpretation of GWAS findings

In this thesis, on the assumption that multiple regulatory variants are present per locus, I conduct a systematic evaluation of whether and how transcription is affected by combinations of SNPs in two or more regulatory intervals. Based on the resultant regulatory signals, I then describe how to use this information to interpret GWAS results, and reveal potential biological mechanisms. My approach is to perform a Bayesian statistical assessment of causal variants and causal genes for human complex phenotypes and disease states. Two large eQTL datasets have been analyzed in order to develop statistical methods that control for population and family structure, and perform simultaneous multi-site eQTL detection in the presence of variable levels of linkage disequilibrium. With the available ~1,300 GWAS summary results, I also conducted co-localization analysis to evaluate the potential causal variants, and causal genes.

Aim 1. Establish limits to fine-mapping imposed by interference among linked sites at a single locus.

The first objective of my thesis was to explore sources of error in estimating joint eQTL effects. Empirical analysis of the CAGE dataset of 2,800 whole blood profiles generated statistical evidence that 2, 3 or even more sites regulate the expression of a gene is common. I used simulations to explore the effect of untagged variants and multi-site regulation on the localization of, and effect size estimation of, statistical peaks. I also showed that under plausible parameters there is a non-trivial likelihood that discovered eQTL credible intervals do not actually include the causal variant. To reveal the limitations

of current frequentist and Bayesian methods, I then performed simulations to compare sequential conditional and Bayesian joint mapping methods for eQTL detection in scenarios where multiple eQTL regulation is prevalent, and compared the advantages and pitfalls of the two methods.

Aim 2. Development of a novel pipeline, PolyQTL, for fine mapping eQTL effects.

In Aim 1, it became apparent that conditional mapping is fast and efficient for detection of independent eQTL in low LD, while the Bayesian methods are more computationally demanding but have increased resolution of multiple eQTL within high LD blocks. In this Aim I present a new pipeline, PolyQTL, which combines the two approaches while also accounting for population and pedigree structure and can incorporate functional and evolutionary information into the fine mapping. The method was applied to contrast multi-eQTL profiles in the CAGE and Framingham Heart Study (FHS) whole blood datasets. The results indicate that the Illumina and Affymetrix platforms yield similar numbers of eQTL, but have very different fine mapping results for more than half of all expressed genes. I discuss the reasons for the discrepancy, and have generated a public database that provides joint mapping profiles and summary statistics suitable for co-localization studies.

Aim 3. Application of eQTL to interpret GWAS results.

Since in previous studies, only a limited proportion of GWAS hits were demonstrated to influence traits through gene expression, I inferred that there are two major limitations: first, the assumption of single-causal variant in one locus may not hold since there are multiple local causal variants influencing regulation of the transcript, and second,

sample size. In this aim, I tested whether additional secondary eQTLs can be used to discover novel co-localization signals. I extended the previously developed PolyQTL method to evaluate whether or not the eQTL statistical signals overlap with GWAS signals. A significant degree of overlap indicates a co-localization, and my method can be used to refine both causal variants, and causal genes.

Aim 4. Trans-eQTL detection in structured populations.

Although eQTL analysis has become prevalent in human genetics and is widely used to interpret GWAS results, most eQTL studies focus only on cis-eQTL effects. The detection of trans-eQTL may provide additional, more detailed information regarding gene expression networks, and elucidate potential biological mechanisms for the regulation of transcript abundance. Dr. Lude Franke launched the eQTLgen Consortium in order to focus on trans-eQTL, by applying meta-analysis with more than 30,000 samples collected from labs around the world. As a member of this consortium, I was responsible for detection of trans-eQTL in the highly family-structured Framingham Heart Study, and for comparison, using a pipeline provided by eQTLgen, of results for three cohorts collected in Professor Gibson's lab. In this aim, I provide a detailed description of my method of trans-eQTL detection, and explore the likely relationship between trans- and cis-eQTL.

CHAPTER 2

Interference between causal variants in LD Constrains eQTL fine mapping

ABSTRACT: In this chapter, I explore the interference between causal variants when they locate in a region of high linkage disequilibrium. Widely used to interpret genetic risk factors associated with disease or clinical phenotypes by GWAS, most of which locate in non-coding regions, expression quantitative trait locus (eQTL) detection has emerged as an important tool for elucidating detailed biological mechanisms. Most eQTL studies apply univariable linear regression to discover primary signals, and then conduct sequential conditional modeling to detect additional genetic variants affecting gene expression. However, this approach assumes that functional variants are sparsely distributed and that close linkage between them has little impact on estimation of their precise location and the magnitude of effects. Here, I describe a series of simulation studies designed to evaluate the impact of linkage disequilibrium (LD) on the fine mapping of causal variants with typical eQTL effect sizes. In the presence of multisite regulation, even though between 80 and 90% of modeled eSNPs associate with the normally distributed traits, up to 10% of all secondary signals could be statistical artifacts, and at least 5% but up to one-quarter of credible intervals of SNPs within $r^2 \geq 0.8$ of the peak may not even include the causal site. The Bayesian methods eCAVIAR and DAP (Deterministic Approximation of Posteriors) provide only modest improvement in resolution. With the results from simulations, I conclude that fine mapping of causal variants needs to be adjusted for multisite influences, but ultimately experimental verification of individual effects is needed. Presumably similar conclusions apply not just to eQTL mapping, but to multisite influences on fine mapping

of most types of quantitative trait loci. Contents in this chapter have already been published in the journal *G3*, as Zeng et al., 2017.

2.1 Background

Biological and statistical evidence imply that many SNP-trait associations identified from genome-wide association studies can be attributed to effects on gene expression. Consequently, fine-mapping to precisely estimate the effect size and define the location of causal variant(s) has become crucial for our understanding of the relationship between genetic risk and phenotypic variation (Maurano et al. 2012; Farh et al. 2015), and eQTL analysis now has become a standard component for GWAS studies (Nicolae et al. 2010). However, investigators tend to expect that it is simple to identify the causal genes at each locus, but even this is a far from trivial undertaking (Chung et al. 2014; Pickrell 2014). Further, when integrating eQTL signals into the GWAS interpretation, many researchers make the stronger assumption that co-localization of eSNP and GWAS signals to a tight LD interval implies the ability to define if not the causal variant, then at least a credible set of SNPs that include the causal site (Trynka et al. 2013; Gaulton et al. 2015; Kichaev and Pasaniuc 2015; Liu et al. 2016). Studies across a wide range of organisms including yeast, mice, and several plant species, reviewed by Albert and Kruglyak (2015) and Cubillos et al. (2012), show that individual regulatory substitutions can be experimentally defined and linked to visible phenotypes. Similarly, the SORT1 example in humans (Musunuru et al. 2010) showed how dissection of the path from regulatory variant to tissue-specific expression can define causal influences on (heart) disease. However, this is painstaking work that relies on strong prior statistical or functional prediction of likely credible intervals. The enrichment of chromatin marks such

as DNAse hypersensitive sites in the vicinity of eQTL validates the assumption that many credible intervals encompass regulatory SNPs (ENCODE Project Consortium 2012; Roadmap Epigenomics Consortium 2015), but conversely raises the question of why there are so many instances of discordant fine localization (Huang et al. 2015; Chun et al. 2017); does this reflect biochemistry (regulatory sites do not always map to ENCODE elements), or simply limits to the statistical resolution of association signals (Udler et al. 2010)?

It is generally assumed that functional variants are sparsely distributed across genomic loci, and that the statistical estimation of effect size and localization isn't hindered by the confounding of statistical signals stemming from LD. However, as GWAS have increased in size, it has become clear that multisite effects are not uncommon. For example, a recent meta-analysis of height suggests that over one-third of the 400 identified loci have multiple independent signals (Wood et al. 2014), and that the expression of a large proportion of genes in lymphocyte cell lines is regulated by two or more locally acting variants (cis-eQTL) (Liang et al. 2013). Human chromosomes have extensive LD often over more than 100kb, and thus the potential for mis-estimation of eQTL effects due to interference between signals from tightly linked polymorphisms could be high. In this chapter I describe a combination of simulation studies designed to address this concern.

Heritability analyses have shown that, on average, up to half of the variance of phenotypic traits, or of transcript abundance, can be explained by genetic factors, mostly acting in an additive manner (Powell et al. 2013; Wright et al. 2014). In contrast to GWAS for visible phenotypic traits, in which the identified genetic variants only explain a subtle proportion of phenotypic variance (<1%), an important difference in eQTL is that one or a few SNPs are often found to explain a large proportion (>10%) of the genetic variance.

These variants usually lie within 1 Mb of the transcript and are defined as cis-acting regulatory polymorphisms. Expression heritability results by Lloyd-Jones et al. (2017) showed that 35% of all expressed genes in peripheral blood have narrow sense heritability >0.1, with a median of 0.3, and the primary cis-eQTL, which has the strongest association signal at a locus, typically explains 85% of the locally acting variance, which is two-thirds of that attributed to all detected eQTL. The majority of the genetic variance is generally actually due to trans-acting polymorphisms of small effect.

The largest blood eQTL study reported to date, assembled from meta-analysis of over 5,000 individual Illumina microarray samples (Westra et al. 2013), reported single site local associations that are genome-wide significant for 6418 genes (44% of those tested) with a 5% false discovery rate. However, the blood eQTL browser only provides single site (unconditional) estimates for all local SNPs at each locus. A more powerful cross-population Bayesian method (Gusev et al. 2014), applied to just 420 lymphocyte cell lines in the Geuvadis dataset (Lappalainen et al. 2013), found a very similar number of genes with evidence for regulation by a local eQTL (eGenes), 14% of which had strong evidence for secondary association signals in a multisite analysis (Wen et al. 2015).

One of the major factors constraining our understanding of eQTL regulation is the limited availability of human samples. Increased sample size should not only reveal more specific instances of multisite regulation, but also provide the opportunity to more accurately define effect sizes in the presence of multiple sites that have varying degrees of LD. Figure 2.1 illustrates the reasoning for a hypothetical case. Five true eSNP effects are indicated by red lines with increasing effects above the horizontal and decreasing below it,

while single site unconditional estimates at common variants across the locus are indicated as black lines.

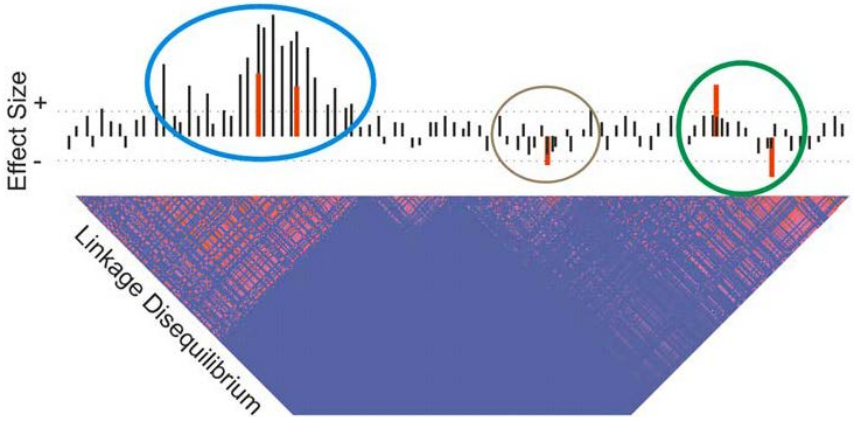

Figure 2.1 Schematic of multi-site regulation of gene expression. Black bars indicate univariate estimates of allelic effects of minor alleles increasing (above the horizontal) or decreasing (below the horizontal) gene expression without conditioning on other sites. Red bars show the actual effects at 5 SNPs in this locus, which has a linkage disequilibrium profile with two large and one small block of elevated LD (pink squares). Dotted horizontal lines indicate a statistical significance threshold, which is only exceeded in the univariate modeling by the two left-hand sites (blue circle). Since these two sites act in the same direction, they reinforce one another, leading to over-estimation of their effect sizes, whereas the two at the right (green circle) interfere with one another antagonistically, leading to under-estimation of their effects. The effect at the fifth site (brown circle) may only be identified following conditional analysis.

Where two sites in high LD have effects in the opposite direction (green circle), they will either cancel each other out or substantially bias the effect size estimates. Where two sites act in the same direction (blue circle), their effects will tend to be added together, and hence the strongest association will overestimate the effect while the secondary site will be underestimated or not detected. Weaker associations (brown circle) may be undetected if they are influenced by even low levels of LD. In theory, if the location of the functional sites is known a priori, these difficulties can be resolved by multisite linear regression simultaneously fitting all of the identified SNPs. In practice, the identities of the functional sites are unknown, and exhaustive multisite modeling is impractical, so

sequential conditional analyses are used to find secondary, tertiary, and so forth associations that are independent of the primary signal. In the presence of strong LD, this approach is expected to miss independent associations, which will remain confounded with the primary signal.

Several Bayesian methods have recently been introduced to improve localization of linked causal variants. CAVIAR (Hormozdiari et al. 2014) enumerates all possible causal states for one or more sites in a short interval of 100 SNPs, but to control the computational burden, the maximum number of causal variants is typically set to two. It is claimed to improve identification of causal variants by 20%–50% over existing methods such as BIMBAM (Servin and Stephens 2007). The eCAVIAR extension for combined eQTL and GWAS analysis (Hormozdiari et al. 2016) uses a greedy method to find a subset of SNPs with a specific confidence (95% by default) that causal variants are identified as candidates. PAINTOR (Kichaev et al. 2014) uses a similar algorithm, whereas FM-QTL (Wen et al. 2015) applies an MCMC algorithm to explore the causal status space, utilizing a posterior inclusion probability to choose the causal variant credible interval. DAP software was then developed (Wen et al. 2016) to explore high probability causal intervals with reasonable runtime. FINEMAP (Benner et al. 2016) uses a logical schema that is similar to that of CAVIAR, but adopts a Shotgun Stochastic Search method to restrict the search space and focus on combinations of high probability intervals.

In this study, I aimed to explore the sources of error in estimating eQTL effects by using simulations to ask how multisite regulation influences (i) the number of independent peaks detected by stepwise conditional analysis, (ii) the accuracy of localization of true causal variants, and (iii) the effect size estimation of discovered causal variants. I also

address the question of what proportion of discovered peaks may be driven by undocumented variants in LD with the genotyped sites, and conclude with a comparison of the performance of two recently developed Bayesian joint localization methods, eCAVIAR and DAP. Only minor improvements in detection of linked causal variants was obtained, and this has little impact on fine mapping, particularly in regions of high LD or if sample sizes are small.

2.2 Material and methods

2.2.1 Consortium for the Architecture of Gene Expression (CAGE) dataset

My simulations utilize genotypes obtained from the CAGE dataset, which consists of Illumina HT12 microarray-based gene expression profiles, as well as whole-genome genotype information from five research studies: the Brisbane Systems Genetics Study (BSGS, N = 926) (Powell et al. 2012), the Atlanta-based Centre for Health Discovery and Well-Being (N = 439) (Wingo and Gibson 2015) and Emory Cardiology Genebank (N = 147, Kim et al. 2014), the Estonian Genome Centre, University of Tartu study (N = 1065, Schrammet al. 2014), and the Morocco Lifestyle study (N = 188, Idaghdour et al. 2010), for a total of 2765 individuals. Since the BSGS sample includes twins, it was removed to avoid complications of relatedness, leaving a set of 1839 European- ancestry unrelated individuals. IRB approval was obtained for the combination of data into a mega-analysis, both by the University of Queensland and for each participating site.

Genotype imputation for the CAGE cohort was performed jointly on the five contributing studies by collaborators at the University of Queensland, to ensure uniformity of assignment of strand identities of SNPs. It was described in detail in Lloyd-Jones et al.

(2017) and at https://github.com/CNSGenomics/impute-pipe. Briefly, the pipeline involved: (1) pre-imputation quality control and data consistency checks; (2) imputation to the 1000G reference panel with Impute2 (Howie et al. 2012); (3) post imputation quality control (filtering on various data features); and (4) merging datasets on common SNPs.

2.2.2 Simulation studies

I conducted four different simulation studies. In all cases, the terminology uni-site (univariable) is used to refer to models where a single causal variant is modeled as a fixed effect, and multisite (multivariable) where two or more variants are modeled, also as fixed effects. The term multi-variate modeling is used for situations where there are two or more dependent variables, whereas in these models I assess the joint effects of two or more causal variables, hence perform multivariable modeling. Some models also incorporate random effects of covariates such as a genetic relationship matrix.

The first set of simulations assessed the power and accuracy of two site regressions assuming that the identities of the two causal variants are already known. I modeled the influences of effect size, minor allele frequency (maf), LD, and sample size. Environmental variance was randomly generated as a z-score (mean 0 and SD 1) and genotype effects (b) were added in SD units (sdu) multiplied by 0, 1, or 2 according to genotype so as to account for from 2% to 30% of the phenotypic variance, computed as $2P(1-P)\beta^2$. Thus, an allele with $\beta = 0.8$ is expected to explain 20% of the variance if maf P = 0.2, or 32% of the variance if P=0.5.The influence of LD was assessed at $r^2 = 0.1$, 0.5, or 0.9, noting that as LD increases, high r^2 values are not obtained for combinations of a rare and a common

allele. For each combination of parameter values, I generated 1000 randomizations of the environmental variance, and assessed (i) the univariate estimate at each genotype, (ii) the mean conditional estimate of the second SNP, and (iii) the joint effect estimates with both SNPs. From these values, I computed the mean absolute value of the deviation between the observed estimate and the true effect size from the univariate, conditional, and joint (two site) models. The univariate estimates agree extremely well with expectations from the analytical solution described in the Results.

The second simulation study asked whether unimputed variants influence the localization of eSNP signals. Since non-imputed SNPs are not present in the CAGE data, I approximated their identities by randomly sampling from a set of CAGE-imputed SNPs weighted to have the same frequency distribution shown in Figure 2.3C and assigning effect sizes from 2% to 10% of the variance explained for normally distributed pseudogene expression traits using the CAGE (without BSGS samples) genotypes. I then removed the SNP and all other SNPs with $r^2 \geq 0.8$, and performed stepwise conditional regression, documenting instances of primary and secondary signals at $P \leq 10^{-5}$, as plotted in Figure 2.3E. The cumulative proportion of spurious secondary signals was computed by summing the detection rate by the size of the maf bin of the unimputed SNPs.

The third set of simulations were performed to evaluate the difference in effect size estimates using the multisite linear regression method for parameter estimation from data representative of the LD structure in the CAGE dataset. For each of 500,000 iterations, four sites were chosen at random from a window extending from 200 kb upstream of the transcription start site to 200 kb downstream of the transcription termination site of a randomly picked gene in the CAGE cohort (excluding the BSGS data, since it includes

twins), and assigned an effect size from a uniform distribution of variance explained (VE) relative to environmental noise ranging between 0.02 and 0.1. The effect size β for an allele with maf p is computed as $\sqrt{\mathrm{VE}/[2\mathrm{p}(1-\mathrm{p})]}$. Subsequently, each phenotype was simulated as β_i* $geno_i$ + N(0,1), whereβ_i is the simulated allelic effect size for a SNP i, and $geno_i$ is the dosage of minor allele at the simulated SNP for a given sample. The significance threshold for sequential conditional detection of the variants in a sample of 1839 CAGE individuals was set at $P \leq 10^{-5}$, since simulations indicated a ≤ 1% false discovery rate for null variants at this level. I evaluated (i) how many of the four SNPs were significant in sequential conditional modeling, (ii) the mean LD between each SNP and the other three SNPs in the model, (iii) the effect size estimates from the conditional and joint multisite models, (iv) the difference between these two estimates as a function of the mean LD, and (v) the rank of the discovered SNP for each peak eSNP and the modeled sites, which were assumed to be the causal variants for some trait.

The fourth set of simulations was performed to evaluate the influence of two Bayesian methods for fine mapping that is sensitive to the LD structure at a locus. First, eCAVIAR was used to also assess the accuracy of co-localization of eQTL and GWAS signals. Summary statistics were generated for normally distributed traits where either one, two, or three sites chosen at random from contiguous intervals of 100 SNPs in the full sample of 1835 CAGE genotypes were assigned to explain between 2% and 10% of the variance. Effects were assigned in the same direction for each minor allele. Marginal single site estimates were generated by uni-variable regression, and then eCAVIAR was used to combine the Posterior Probabilities, which were multiplied together to yield the Co-

Localized Posterior Probabilities (CLPP) with a significance threshold of 0.001 as recommended (Hormozdiari et al. 2016). Owing to the high computational burden, only 4000 simulations were performed. GWAS variants are in general unlikely to explain this amount of variance, but the statistical evidence is approximately equivalent to that expected for typical trait associations where a SNP explains ≤ 0.1% of the variance in a sample of 20,000 individuals. The effect of sample size was evaluated by fitting a single eQTL effect to just 200 individuals in each simulation. Second, the DAP simulations were performed using the adaptive algorithm, which estimates the number of causal variants from the data and also generates a list of possible sites that could explain the effect(s). Again, owing to the high computational burden, only 130 simulations were performed, using the same parameters as for the sequential conditional analyses with four assigned causal variants. A final set of simulations designed only to fine map three causal sites in a single moderate to high LD block extracted contiguous sets of 100 SNPs, and randomly assigned effects only on the condition that three sites selected from the set each had $r^2 \geq 0.3$ with one another.

2.3 Results

2.3.1 Underestimation of allelic effects by sequential conditional analysis

My basic simulation framework utilizes the current standard mapping approach of sequential conditional analysis, in which the residuals from discovery of each SNP are taken forward as the dependent variable in a new scan for an independent SNP (Yang et al. 2012). To explore the performance of this strategy in the context of four causal regulatory variants in the vicinity of a typical gene, 500,000 simulations were carried out by randomly picking four SNPs within 200 kb up- or downstream of the TSS and TES ends of a

randomly chosen gene, from the imputed whole-genome genotypes of 1839 unrelated European ancestry individuals. I assigned each SNP an allelic effect size so as to explain between 2 and 10% of the variance of a trait otherwise uniformly distributed with a mean of 0 and SD of 1. Power to detect individual univariate effects of this magnitude is close to 100% at the significance level $P \leq 10^{-5}$. The sampling was performed across all genes so as to sample from the typical LD structure in the European ancestry human genome. Furthermore, effects were randomly assigned under three scenarios, with either four positive (4:0), three positive and one negative (3:1), or two positive and two negative (2:2) effects of the minor allele on the trait. For the eQTL detection, once the sequential conditional detection was completed, I determined which of the four causal variants was in high LD ($r^2 \geq 0.8$) with one of the discovered sites. If a peak was in high LD with more than one causal variant, it was assumed that it tagged the highest effect site.

Table 2.1. Tagging Efficiency of Detection of Causal Variants with r^2 cutoff 0.8

Number of Detected Causal Variants	Scenario (Positive: Negative effects)*		
	4:0	**3:1**	**2:2**
1	0.6% (0.84)	1.4% (1.00)	1.2% (0.62)
2	5.7% (0.69)	10.4% (0.91)	12.8% (0.86)
3	28.2% (0.78)	24.3% (0.85)	22.5% (0.76)
4	55.5% (0.89)	59.0% (0.92)	60.2% (0.90)
>4	10.0% (0.63)	5.0% (0.66)	3.3% (0.56)

* % indicates the percent of cases with the indicated number of independent discovered variants; numbers in brackets are the proportion of discovered variants that are in LD ($\mathbf{r}^2 > 0.8$) with one of the simulated causal variants.

Table 2.1 summarizes the "tagging efficiency", namely the percent of simulations in which the indicated number of significant independent sites was detected, as well as the proportion of the discovered variants that are in high LD ($r^2 \geq 0.8$) with one of the simulated causal variants. Across all three scenarios, at least three independent peaks are detected in 90% of the simulations, and at least four independent peaks in two thirds of the simulations. Notably, in the scenarios where all four minor alleles influence expression in the same direction, 10% of the simulations detected five or more independent peaks, at least one of which must be a spurious association, despite a $\leq$ 1% false discovery rate for univariate associations of the same magnitude. The proportion of credible intervals ($r^2 \geq 0.8$ regions around each discovered variant) that contain the actually simulated site ranges between 85% and 90% in each scenario, again indicating relatively poor localization of the causal variant.

Table 2.2. Detected true causal variants in simulations with 4, 3 and 2 causal variants

Detected	4 SNP Scenario			3 SNP Scenario		2 SNP Scenario	
variants	4:0	3:1	2:2	3:0	2:1	2:0	1:1
0	1.1%	1.2%	1.4%	2.1%	1.9%	5.5%	6.3%
1	5.5%	6.0%	6.0%	12.1%	15.1%	9.4%	9.1%
2	18.4%	22.3%	23.7%	17.6%	17.9%	85.1%	84.6%
3	23.7%	23.4%	22.1%	68.2%	65.0%		
4	51.3%	47.1%	46.7%				

Similar results are reported in Table 2.2 for the reciprocal measure of what fraction of simulated variants is captured by discovered variants. It includes results for simulations with just two or three causal variants, and reports the percentage of cases where the causal variant was within the $r^2 \geq 0.8$ credible interval for a discovered peak. Across the sets of 500,000 simulations, at least two variants are detected > 85% of the time, but the power to detect all of the multiple eSNPs is a function of the number of sites operating in the same direction. It is highest for the case where the minor alleles for all four variants have effects in the same direction and least where two are in one direction and two in the opposite direction. In the 4:0 scenario, three or more of the four eSNPs are detected three-quarters of the time, whereas this proportion drops toward two-thirds with the simulations for 2:2. No variants are detected in just over 1% of the simulations, and just one variant in ~6% of them, while 80% of the variants are detected overall. This proportion rises to 90% for the two-variant simulations, illustrating how multisite interactions reduce the discovery of independent eQTL peaks.

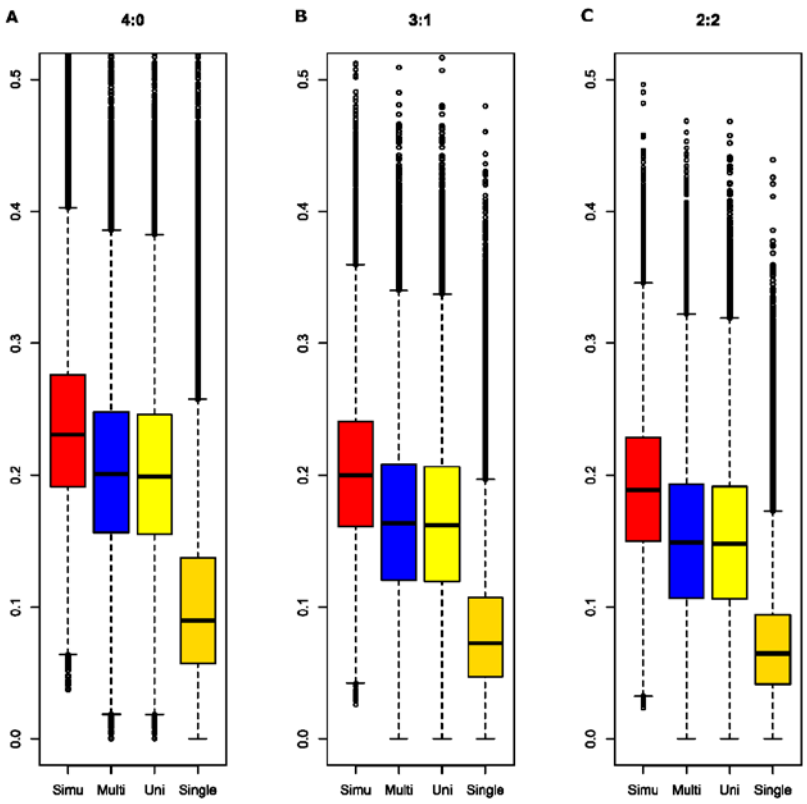

Figure 2.2 Proportion of variance explained by detected eSNPs in simulations. Box and whiskers show median, interquartile range, and 95% C.I. for the proportion of variance explained under three scenarios for 500,000 simulations of four sites affecting gene expression. From left to right in each simulation, Simu is the variance explained by the known sites, Multi is the result fitting discovered eSNPs jointly, Uni is the result of summing the effects from sequential conditional modeling, and Single is the effect of the peak detected eSNP. The y-axis shows the proportion of variance explained. Scenarios are 4:0, all four minor alleles with effects in same direction; 3:1, one minor allele effect in the opposite direction; 2:2, two minor alleles on one direction and the other two in the opposite direction. eSNP, expression single nucleotide polymorphism.

Another way to consider the power of multisite detection is to ask how much of the variance explained by the four SNPs is captured by the discovered variants. Box and whisker plots in Figure 2.2 shows that, under all three scenarios, on average 85%–90% of the variance is captured, namely in these simulations ~15%–20% of the transcript abundance. Although effect sizes of all SNPs were drawn from the same distribution, the first discovered SNP (rightmost box in each panel) typically explains between one-third

and one-half of the expected variance, suggesting that it often tags some of the effect of another site. Since the primary SNP captures on average more than two thirds of the heritability at each locus in actual peripheral blood data (Lloyd-Jones et al. 2017), it is likely that secondary and tertiary SNP effects are, in reality, smaller than primary SNP effects. As expected, summation of the independent contributions from the sequential conditional models, or fitting all of the discovered variants simultaneously in a multivariable model, explains very similar proportions of the variance overall. Similar results are seen with two or three simulated causal variants.

2.3.2 Estimation of the proportion of secondary associations that are false positives

The detection of more association peaks than the number of simulated sites implies that some fraction of peaks are false positives that arise due to sampling artifacts in the presence of high LD, whereby an imperfectly tagged site is split into two or more spurious signals. An example is shown in Figure 2.3A and B, contrasting the local Manhattan

profiles for a single causal variant that splits into two associations when the peak SNP, rs9806753, is excluded from analysis.

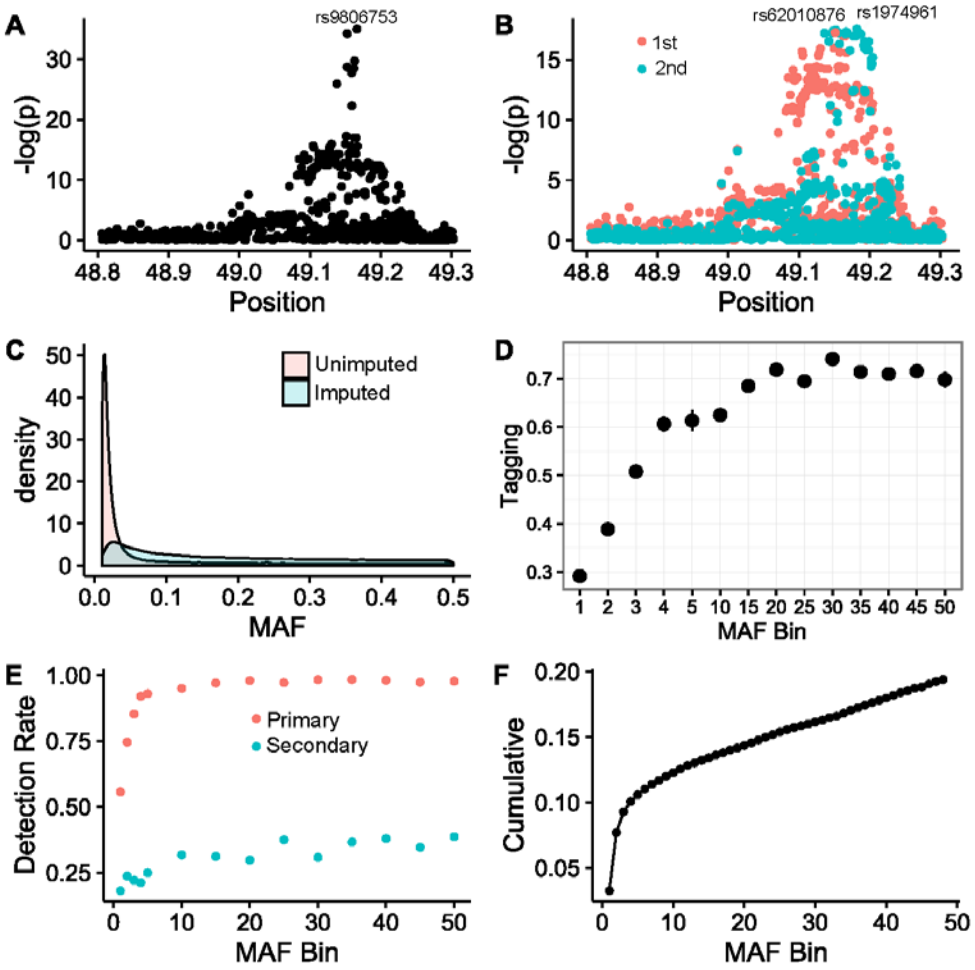

Figure 2.3 False multiple eQTL detection due to unimputed variants. (A, B) example showing multi-eQTL due to poor tagging. (A) SNP rs9806753 (European-ancestry maf = 0.23) was simulated to generated an eSNP effect, but removal of this variant and all SNPs within $r^2 > 0.5$ in analysis, (B) results in the effect captured by a primary (rs62010876, maf = 0.10) and secondary (rs1974961, maf = 0.13) signals. (C) Empirical maf distribution of 1.4 M unimputed 1000G (red) and 8.3 M imputed 1000G variants in CAGE (blue), showing shift to lower frequencies for variants not tagged. (D) Tagging efficiency as a function of maf based in mean r^2 for the strongest correlated SNP for 10,000 randomly selected variants in the CAGE. (E) Corresponding signal detection rate at $P < 10^{-5}$ for randomly assigned effect sizes, explaining between 2% and 8% of a simulated gene expression trait for primary (red) and secondary (blue) signals when the simulated variant is excluded from the analysis. (F) Cumulative proportion of sites expected to generate a false multiple eQTL detection. Multiplication of this proportion by the number of unimputed SNPs in genic regions and the actual proportion of SNPs that have effects (unlikely to be >1%) yields up to 400 possible false positive secondary associations.

To explore the frequency with which this occurs, I conducted simulations assuming a single causal variant based on the genotypes measured in 1839 European ancestry samples from the CAGE cohorts (see Methods). Randomly assigning causal effects resulted in the appearance of a secondary signal at $P \leq 10^{-5}$, conditioned on the causal site, at 0.3% of the loci. This is approximately as expected given 8.3 million imputed SNPs at 22,000 loci, and is also the same as the false discovery rate of primary signals in the absence of any simulated causal locus; that is to say, the random expectation is for 0.3% of transcripts to have a false eQTL discovery at $P \leq 10^{-5}$ in the CAGE dataset. However, this ignores the possibility that the causal variants are not present in the imputed genotypes. Of the 9.7 million SNPs, indels, and CNV with maf ≥ 0.01 in the European populations in the 1000G database, 1.9 million are not imputed in the CAGE samples. Figure 2.3C shows that the maf distribution for these variants is strongly shifted toward rarer alleles relative to the imputed SNPs, and is centered at a maf ≥ 0.02. Consistent with Yang et al. (2015), the average tagging efficiency (r^2 value) of these SNPs is a function of maf, being >0.7 for maf >0.05, but dropping to <0.5 for maf = 0.01, as seen in Figure 2.3D.

Since I cannot simulate effects at nonimputed SNPs, I approximated such alleles by randomly simulating a causal variant from the CAGE SNPs with the same frequency distribution, but excluding it from the analysis along with all variants that would tag it at the typical level observed for the nonimputed SNPs of the same maf. I then asked how often the effect is captured by multisite signals, as a function of residual tagging efficiency. I allowed for increased effect sizes with lower maf by simulating effects in the constant range of 2%–10% of the variance explained. The proportion of such pseudo unimputed

SNPs that generate primary signals is reduced with lower maf, due in part to the smaller proportion of variance explained by the less common variants that partially tag them. For common variants, there is almost always a second site in high enough LD to capture most of the causal signal in the absence of genotypes at the causal variant, but rare variants are insufficiently tagged to generate a signal at all, 90% of the time. In the presence of tagging SNPs with $r^2 \geq 0.5$ to the "unobserved" causal variant, false secondary associations are observed ~40% of the time. At the other end of the spectrum, rare variants (maf ≤ 0.01) that produce a primary signal at a tagging SNP with $0.1 \leq r^2 \leq 0.3$ also produce a secondary signal but less frequently. The blue curve in Figure 2.3E indicates the inferred fraction of unimputed variants that could induce secondary signals as a function of maf, and Figure 2.3F shows that the cumulative proportion of such spurious eQTL weighted by observed maf proportions approaches 20%. Approximately 14% of the 1.9M unimputed variants are located within 200 kb of a gene, and assuming that 0.1% of these actually have an eQTL effect, this suggests the potential for ~250 such effects.

These computations argue that up to 10% of the observed 2300 multisite associations reported by Lloyd-Jones et al. (2017) have the potential to be false signals driven by inefficient tagging of unimputed variants in CAGE. The proportion could be greater if the fraction of functional SNPs is higher, as suggested for example by Tewhey et al. (2016), who used a very sensitive MPR assay to implicate 3% of regulatory sites in 3642 eQTL regions (842/32,373 tests) as capable of modulating transcript abundance. However, the proportion of sites with detectable signals capable of explaining > 2% of the variance is certainly lower, and 1 in 1000 (0.1%) is a reasonable estimate given that there are of the

order of 1300 documented variants in the vicinity of each gene and no more than 30% of expressed genes have a secondary eQTL signal.

Synthetic associations due to even rarer variants may be expected to generate split associations as well (Dickson et al. 2010; Zhu et al. 2012). Yang et al. (2015) found that ~20% of the variance for height can be explained by SNPs with maf ≤ 0.1, in part due to larger effect sizes of prevalent very rare SNPs, many of which are likely secondary associations. We also found that there is an excess of rare variants (maf ≤ 0.01) influencing extremes of gene expression, also with a slightly larger distribution of effect sizes than common variants (Zhao et al. 2016). Too many unknown parameters need to be evaluated to give a good estimate of the number of false positive secondary associations due to synthetic effects of very rare alleles, but it may be another few percent.

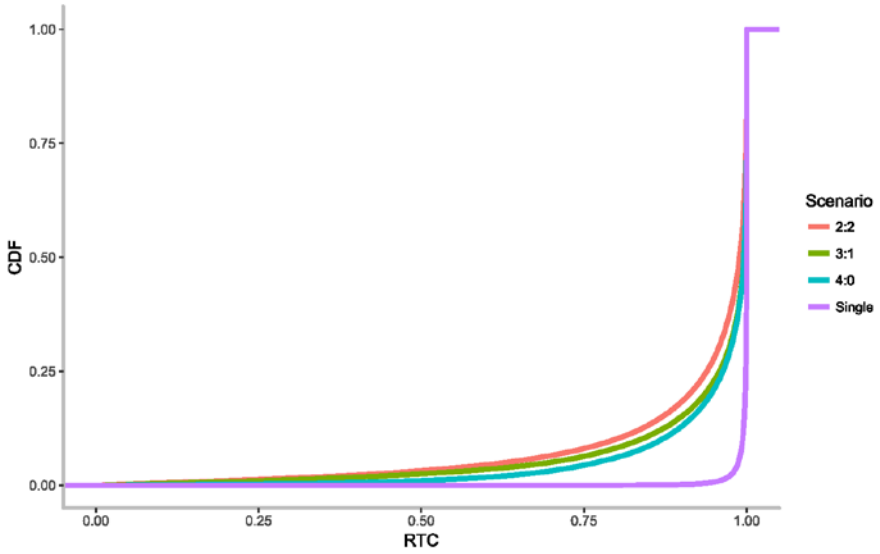

Figure 2.4. The single causal variant assumption biases fine mapping of causal variant locations. Each curve represents the cumulative probability distribution for RTC scores for the primary causal variants under a model with a single causal variant (purple), or with four causal variants under scenarios 4:0, 3:1, and 2:2 (blue, green, red curves, respectively). RTC scores close to one imply equivalence of the significance values of the eQTL and causal variant. eQTL, expression quantitative trait locus; RTC, Regulatory Trait Concordance.

2.3.3 Effect of multisite modeling on accuracy of localization of associations

A possibly more important measure of estimation bias is the location of the peak SNP relative to the causal site. The most straightforward measure of co-localization is the Regulatory Trait Concordance (RTC) score (Nica et al. 2010), which is intuitive and easily implemented on the scale of the simulations. It is essentially a ranking of the significance of the detected eQTL P-value relative to the casual site, RTC = (N_{SNPs} - $Rank_{causal\ SNP}$)/N_{SNPs}, where a value of one indicates identity, and zero that the two sites are in the same locus but highly unlikely to be capturing the same signal. Figure 4 plots the cumulative frequency distribution for RTC scores for the primary eQTL signals relative to the largest effect causal variant in each simulation, contrasting the 4:0, 3:1, and 2:2 scenarios. For comparison, under a single variant model, RTC is always close to one as expected [note that, since we do not simulate the GWAS signal as well, these values are inflated relative to data where the identity of the actual causal variant is unknown (Nica et al. 2010). For 10% of simulations in the presence of multiple regulatory variants, the RTC score of the primary SNP drops below 0.9, again with greater tendency toward mis-estimation of the eQTL location in models with opposing effect directions of the minor alleles. This analysis confirms the results in Table 2.2, indicating that up to 15% of all detected SNPs are not in high LD ($r^2 > 0.8$) with a simulated variant in the imputed panel of SNPs.

Localization of tertiary and quaternary signals is affected more strongly, but intriguingly, considering just associations within $r^2 > 0.8$ of a simulated SNP, the secondary signal is slightly more likely to be the first or second ranked SNP for one of the causal

variants than is the primary signal. This is true under all three scenarios, which have very similar profiles to that shown for the 3:1 scenario in Figure 2.5 (since there is wide variance in the number of SNPs in each region, we simplified the analysis by reporting just the SNP ranks in this figure, rather than RTC). It should be noted that there is not strong concordance between the relative proportion of variance explained by the causal variants and whether they are the primary through quaternary association, since LD has a strong influence on detection power. Although the vast majority of discovered sites are within three or four SNPs of at least one of the four causal variants when they are in high LD with one of them, it cannot be concluded that the order of discovery corresponds to the true order of effect sizes.

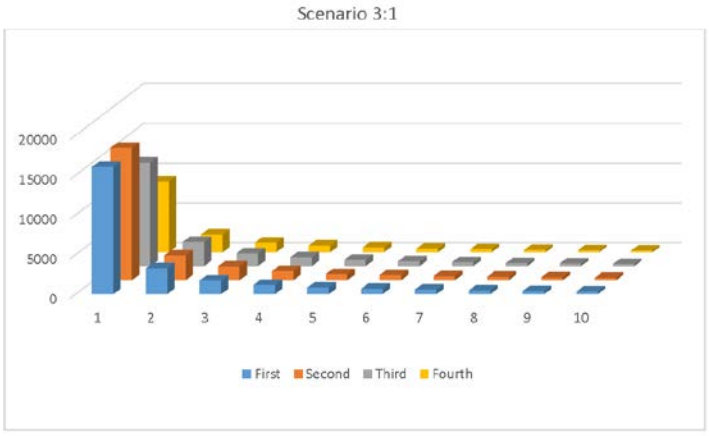

Figure 2.5. Signal ranks of simulated causal variants. The number of simulations in which a discovered variant was the indicated rank (left to right) for the first through fourth discovered variant (front to back). Rank refers to the number of SNPs with a smaller p-value than the modeled causal variant, where 1 implies the causal and discovered are the same SNP, 2 that one other SNP in the LD region had a smaller p-value, and so forth. Only cases where the discovered variant was within r^2 >0.8 of the causal variant are shown.

2.3.4 Joint fitting pairs of known causal variants accurately estimates effect sizes

Before addressing the accuracy of effect size estimation following stepwise conditional analysis, it is worth noting that in the case where the identities of two causal

variants are known a priori, joint fitting of the two SNPs in a single regression on transcript abundance always results in more accurate effect size estimates given a large sample size. The bias in estimation due to linkage disequilibrium between pairs of SNPs is a function of the two effect sizes (β_1 and β_2), the correlation between the SNPs (r), and the ratio of the square root of the product of their allele frequencies (Yang et al. 2012):

$$\hat{E}(\beta_1) - \beta_1 = r\beta_2 \sqrt{p1(1-p1)/[p2(1-p2)]}$$

This is maximized for pairs of SNPs at the same frequency, increases with high LD, and can be either positive or negative depending on whether the signs of the minor allele effects are coupled or not. Figure 2.6 provides a visual summary of the biases, compared with the effect of jointly fitting the two SNPs with a sample size of 2,000, which uniformly improves the effect size estimates.

Three results deserve highlighting. First, for each combination of allele frequencies, increasing the allelic effect size results in more severe biases, in the most severe cases over- or under-estimating the effects by as much as 50% of the variance explained. Second, the first picked SNP (with the larger effect size) has the greater deviation between the estimated and true effect size. This makes intuitive sense as the larger effect will generally be the first detected one and absorbs much of the effect of the other SNP, which will typically be under-estimated in the conditional analysis, but to a lesser extent. If the deviation is computed simply on the uni-site unconditional values, the opposite result is obtained: the deviation is greatest for the smaller effect site. Third, the mis estimation is greatest for lower allele frequencies, which is particularly noteworthy since most eQTL have maf in the range of 0.1 to 0.3.

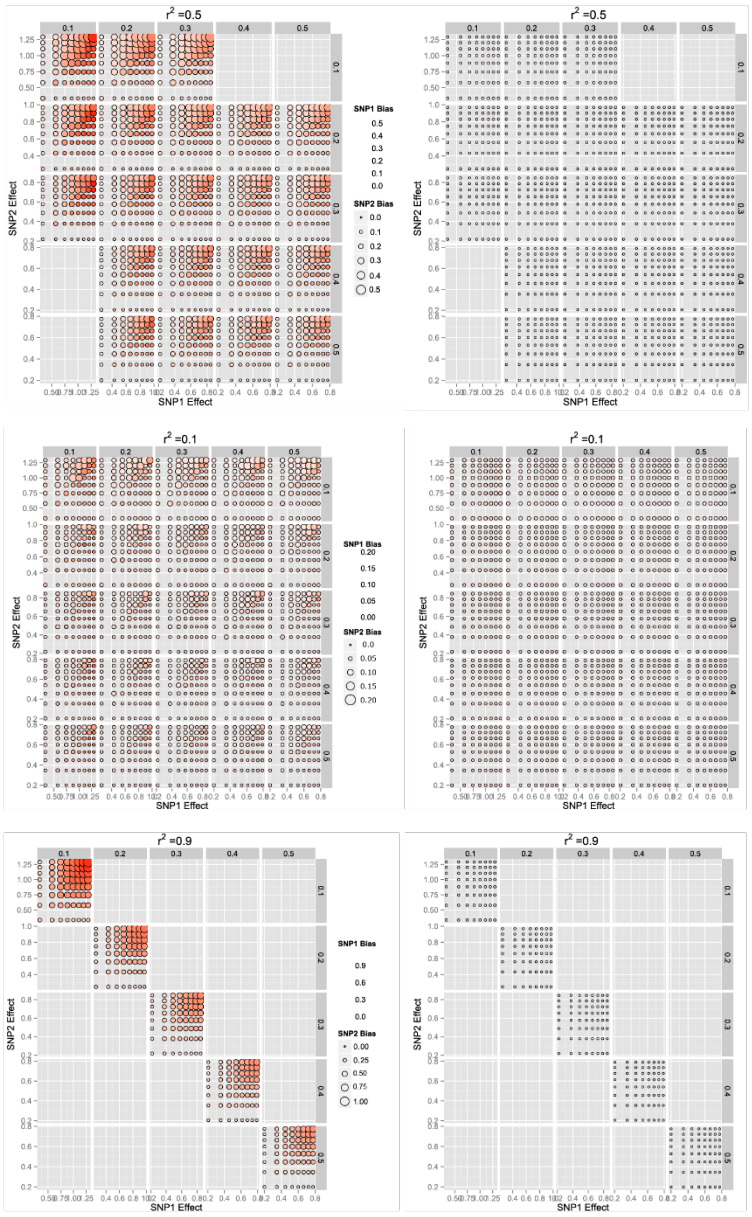

Figure 2.6. Simulation of the influence of minor allele frequency and β on allelic effect size estimation. Left side, sequential conditional modeling; right side joint multi-site modeling, with n= 2,000. Top row LD r^2 = 0.5, Middle row, LD r^2 = 0.1, Bottom row LD r^2 = 0.9. The 25 sectors on each panel show results for combinations of two alleles with maf from 0.1 to 0.5 (left to right, top to bottom), within which each circle represents the average of 100 replicate simulations for allelic effects explaining from 2% to 30% of the expression variance (left to right, bottom to top). Circle size and color is proportional to the absolute value of the deviation between the estimated and actual effect size β in standard deviation units. Transition from white to red represents greater deviation for allele 1; larger circles represent greater deviation from the simulated β for allele 2. Empty fields arise because the indicated level of LD is not possible for the corresponding allele frequencies.

I also considered the power to detect joint effects in the presence of LD. As the p value cutoff for detection becomes more stringent, the bias in estimation becomes more severe, since the first picked SNP absorbs the effect of both alleles into the same estimate, leaving the statistical power of the other allele, conditioned upon the first one, close to zero. With a sample size of 2,000 and intermediate LD, when both alleles are modeled jointly, power to detect both effects remains high across the plausible parameter space once the effect size exceeds 5% variance explained, and the estimation for the beta value is still accurate. Down-sampling suggests that in order to estimate effects within 0.1 standard deviation units, for pairs of variants with LD $r^2 \sim 0.9$, each explaining 10% of the variance (namely having effect sizes of at least a half a standard deviation unit), a sample size of at least 900 is required.

Consequently, most small sample eQTL studies will fail to resolve linked sites into two effects. These results indicate how the typical assumption that an eQTL effect is due to a single variant in a set of credible SNPs in high LD is potentially highly biased. Similar conclusions apply to the situation where two SNPs operate in opposite directions, with the additional dilemma that they will not be detected at all and consequently strongly under-estimate the regulatory variance at a locus.

2.3.5 Mis-estimation of allelic effects sizes by sequential conditional analysis

Even though the sequential conditional and multisite models capture essentially equivalent proportions of the variance tallied across sites, biases in estimation of individual site effects ought to be reduced by the multisite modeling. To quantify this difference, I

computed the deviation between the observed and true simulated effect sizes (β in sdu) for each discovered peak located within r^2 >0.8 of an independent causal variant, and evaluated the absolute value of these deviations as a function of the mean LD between the causal variant SNP and the other three causal variants in the model. Figure 2.7A plots the difference between the models in Figure 2.7C and D, which show the average absolute value of the deviation for SNPs with the indicated true effect size and LD, for the sequential conditional estimates, and for multivariable models fitting all discovered variants jointly, respectively. The figure shows results for the 4:0 scenario where all minor alleles operate in the same direction. The scale from dark blue to yellow indicates mis-estimation of effect sizes ranging from less than 0.5 sdu to more than three units, where each pixel is averaged over the number of simulations with the indicated allele effect sizes and average LD in Figure 2.7B.

Several results are noteworthy. First, for causal variants in low LD, as expected, neither model results in appreciable estimate bias, but once the average LD rises above 0.5, effects can be misestimated by more than the effect size. For example, for β = 0.5, the absolute value of the difference between the observed and true effect is typically between 1 and 2 sdu, which depending on the allele frequency may correspond to at least 2% of the total gene expression variance. Second, for large effect alleles, the mis-estimation is appreciable even at intermediate levels of LD, and it is not unusual for estimates to be off by as much as 4 sdu under either model.

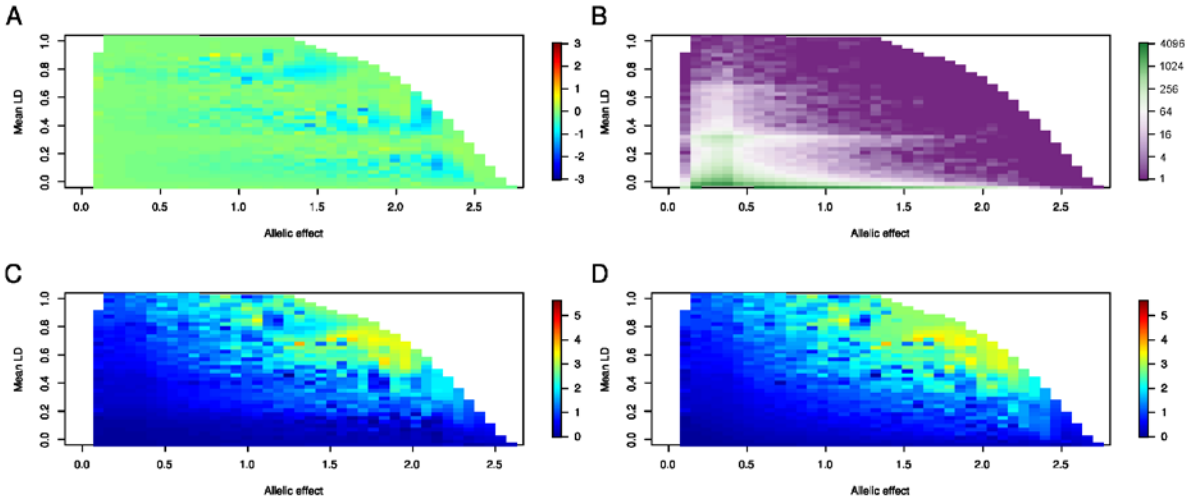

Figure 2.7. Biases in effect size estimation from conditional and joint analysis. All panels refer to 500,000 simulated data points where effect sizes were sampled from a uniform distribution to explain from 2% to 10% of the expression at a locus for each of 4 SNPs picked at random from 400kb intervals of the CAGE genotype data. Panel (A) compares estimates from joint and conditional modeling, as a heatmap of the average difference in panels C and D, where yellow indicates that joint modeling produces a larger estimated effect size, and blue a lower estimate with three bands of negative values indicating greater bias in the conditional estimates. Panel B shows the density distribution on the log2 scale of the number of simulations with alleles for each pixel with the indicated b (in standard deviation units, sdu) on the x-axis, and average LD with the other 3 sites at the locus on the y-axis. Panels C and D show the average absolute value of the deviation between the observed and known effect size for sites under the multi-site model where all discovered sites are fit jointly (C) or from single site estimates after each step of sequential conditional analysis (D), for the 4:0 scenario where all minor alleles have effects in the same direction.

Third, overall, the multisite modeling corrects some of the sequential conditional analysis bias. The difference in performance of the two estimation procedures shown in Figure 2.7A, where most values are pale green, indicates close similarity of the estimates, but bluish-tinged bands imply that the multisite model gives a better approximation to the true effect size for LD centered 0.1, 0.4, and 0.8. Mis-estimation without multivariate estimation can be twice as severe for very large effect alleles, although these only account for a very small fraction of all simulated alleles.

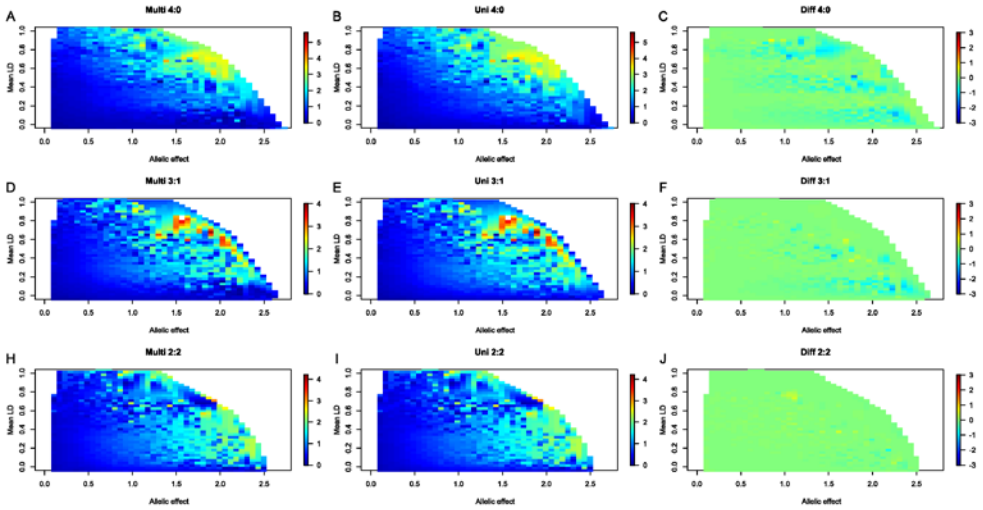

Figure 2.8. Effect size estimation bias under the three scenarios with 4 causal variants. Each scenario includes 90,000 simulations and panels from left to right in each row show the average absolute value of the difference between the estimated multisite β and true β, the absolute value of the difference between the estimated sequential conditional β and true β, and the average difference between these values. Top row is results for four sites where the minor alleles have the same sign of effect, bottom row for two sites each with the same sign, middle row a mixture of three in one direction and one in the other.

Similar trends are seen for the 3:1 scenario, where one of the minor alleles operated in the opposite direction to the other three, as well as in the 2:2 scenario (Figure 2.6), and simulations with just two or three causal variants yield similar conclusions (Figure 2.9 and 2.10 respectively). The advantage of joint modeling is reduced in the presence of opposing allelic effects, but still prevalent in the region with high allelic effect size and low LD; and, as noted above, a large proportion of causal sites are not discovered in the 2:2 scenario, so are not included in the estimation.

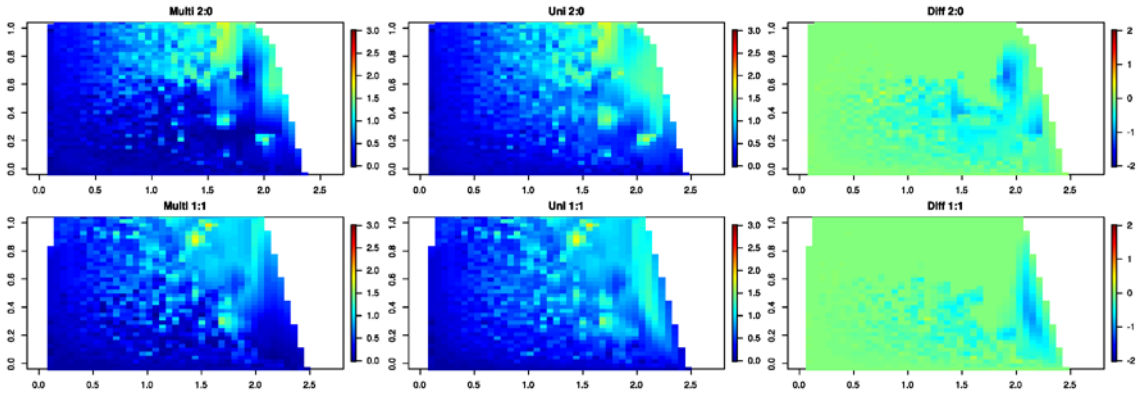

Figure 2.9. Effect size estimation bias under the two scenarios with 2 causal variants. Each scenario includes 90,000 simulations and panels from left to right in each row show the average absolute value of the difference between the estimated multisite β and true β, the absolute value of the difference between the estimated sequential conditional β and true β, and the average difference between these values. Top row is results for two sites where the minor alleles have the same sign of effect, bottom row for two sites with the same sign.

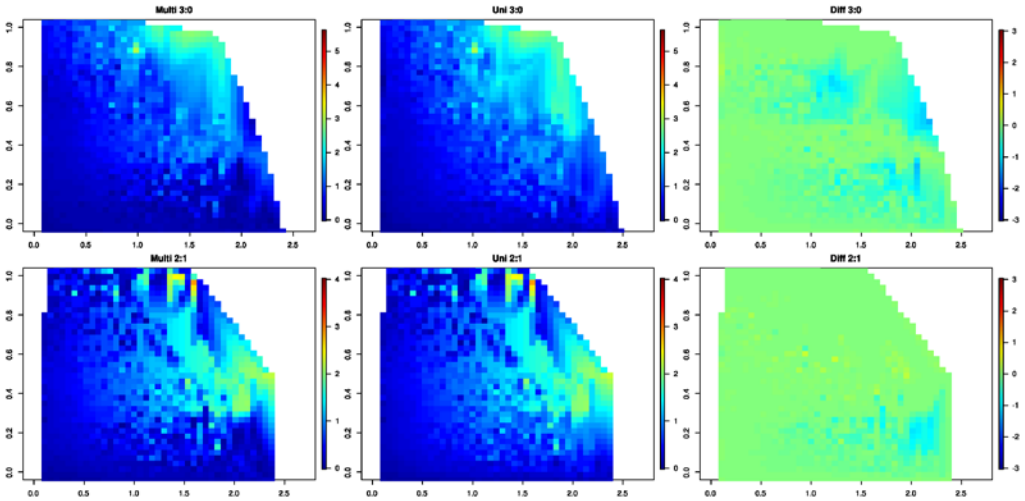

Figure 2.10. Effect size estimation bias under the two scenarios with 3 causal variants. Each scenario includes 90,000 simulations and panels from left to right in each row show the average absolute value of the difference between the estimated multisite β and true β, the absolute value of the difference between the estimated sequential conditional β and true β, and the average difference between these values. Top row is for three sites, all with the same sign, and bottom row is one site operating in the opposite direction to the other two.

In summary, stepwise conditional eQTL discovery is expected to discover between 70% and 80% of eQTL within realistic effect size ranges typical of those reported in the literature. Once discovered, multi-locus estimation of effect sizes provides slightly more

accurate estimates than the estimates from sequential conditional models, but for large effect alleles in high LD the corrections can be substantial.

2.3.6 Bayesian modeling only slightly improves mapping of multisite associations

Recently, a number of Bayesian approaches have been introduced that are designed to improve fine-mapping of eQTL effects (Giambartolomei et al. 2014, Zhou et al. 2013). One of these is eCAVIAR (Hormozdiari et al. 2016), which reports a Colocalization Posterior Probability (CLPP) based on the combined likelihood that a variant influences both the abundance of a transcript and a phenotype given the LD structure at a locus. The authors proposed a CLPP cutoff of 0.001 (for example, a posterior probability of 0.1 for the eQTL and 0.01 for a disease association), which corresponds in the simulations (Table 2.3) to discovery of 80.7% of single variants sampled at random from contiguous blocks of 100% SNPs in the CAGE European ancestry cohort genotype data. The computational burden of evaluating all possible four site combinations is too large for this model to be applied in genome-wide scans. Instead, I performed 4000 simulations of 1835 individuals in the presence of two or three regulatory variants, as well as a normally distributed phenotype, and evaluated the CLPP distributions. In the case of two causal variants, just 94.7% generated CLPP > 0.001, and for three causal variants, 84.9%. Figure 2.11 shows the cumulative distribution functions of the CLPP scores as a function of the number of causal variants, clearly documenting the trend for reduced confidence in joint localization as the degree of multisite regulation increases. Similar trends were seen with a more conservative CLPP cutoff of 0.01, confirming that interference among tightly linked sites reduces the power to detect independent causal variants. The upper red curve also indicates that the power to detect co-localization is greatly reduced with sample sizes of just 200 for

the eQTL sampling: in fact, just 60% of simulations with a single causal site yielded a CLPP > 0.001.

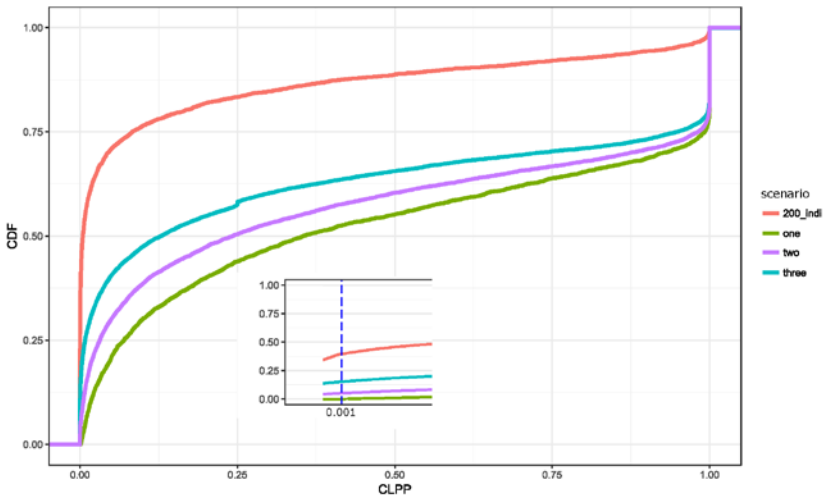

Figure 2.11. Co-localization with eCAVIAR in the presence of multiple regulatory sites. Cumulative Distribution Functions summarize CLPP scores for 4000 simulations each with one, two, or three assigned causal sites within a contiguous block of 100 SNPs. Green, purple, and teal curves show progressive degradation of evidence as the number of modeled causal variants increases, for simulations with 1835 subjects. The red curve shows that more than half of the simulations with just 200 subjects for the eQTL component have CLPP < 0.01, and just one-quarter >0.1, compared with one quarter >0.98 with 1835 subjects. CLPP, Combined Likelihood Posterior Probabilities; eQTL, expression quantitative trait locus; SNP, single nucleotide polymorphism.

Since eCAVIAR is not designed to cover intervals encompassing all of the regulatory regions of a typical gene and hence is not directly comparable with the stepwise conditional regression, I also evaluated the DAP algorithm (Wen et al. 2016). DAP is designed to identify independent credible intervals and report candidate SNPs across a locus, incorporating priors that weight likely functional or evolutionary evidence. Using the adaptive DAP procedure, which does not make any prior assumptions about the number of causal variants at a locus, I performed 130 simulations of four variants with effect sizes drawn as before to explain between 2 and 10% of the variance, all operating in the same

direction. The average number of detected variants was 3.76, which included 3.52 of the four simulated sites (88%) in the candidate list. Due to different statistical thresholds, it is difficult to compare this result with the stepwise conditional model, but it appears to be an improvement on the 80% reported in Table 2.2. As expected, DAP fails to detect true causal variants in the presence of high LD.

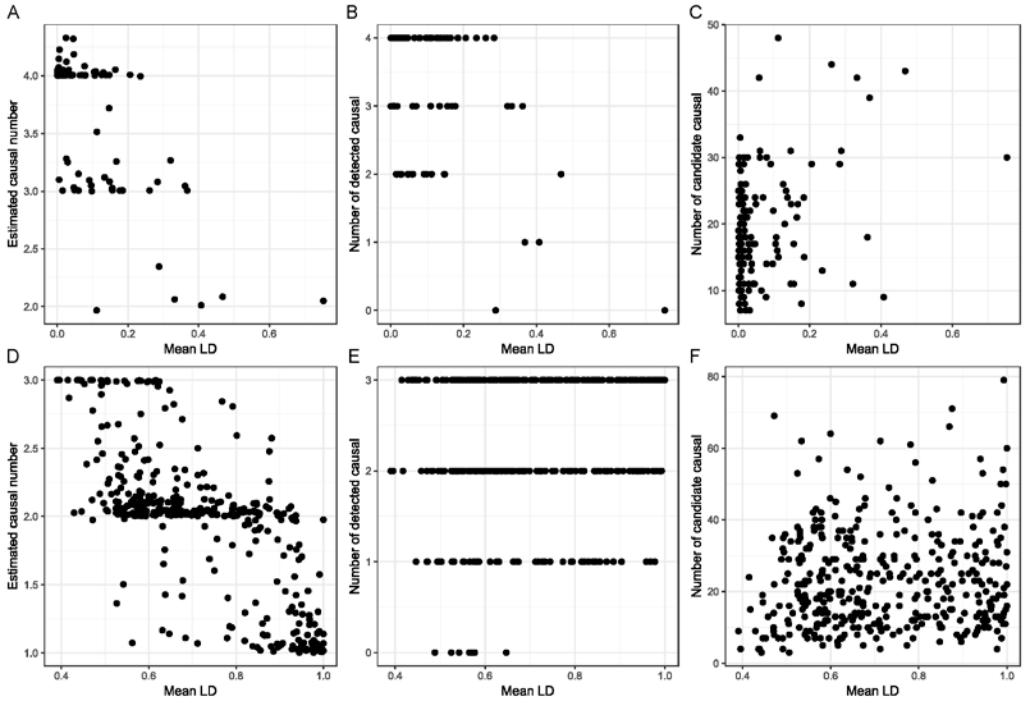

Figure 2.12. Fine mapping with DAP in the presence of multiple regulatory sites. (A–C) Results for simulations with four causal sites drawn at random from 200 kb upstream and downstream of each gene. (D–E) Results for simulations with three causal sites drawn from 100 continuous SNPs, each in LD with $r^2 > 0.3$. (A and D) show the estimated number of sites as a function of the mean LD between the sites, showing that as LD increases, detection of independent intervals decreases. (B and E) show the number of modeled (true) causal sites in the candidate lists, as a function of mean LD, which in this case increases for high LD. (C and F) show that the number of candidate sites increases with high LD, sometimes with 20 or more candidates defining a credible set for each true site. DAP, Deterministic Approximation of Posteriors; LD, linkage disequilibrium.

Figure 2.12A shows the dependency of the number of discovered variants on the mean LD between the four simulated sites in each simulation, while Figure 2.12B shows how many of the true causal variants are detected. Notably, if all four variants are in a

single block of high LD, no sites are detected since the posterior probabilities are dispersed across all of the variants. Across the full extent of up to 500 kb at most loci there are usually multiple LD blocks, so DAP, like stepwise conditional modeling, is quite efficient at detecting independent credible intervals. However, it consistently over reports the number of candidate variants and Figure 2.12C shows that this number also increases with LD. To confirm that DAP is still able to resolve multiple causal variants in the presence of high LD, I also ran 400 simulations with the constraint that three sites must be within $r^2 > 0.3$, using the DAP-k algorithm with k = 3 (assuming three sites), showing the results in Figure 2.12D–F. In this case, the number of detected independent associations dropped to 1.95, namely 65% of the simulated number. Although 79% of the simulated variants were among the candidate lists, these can become very large with a ratio of ten-to-one candidates for each true causal variant. Stepwise conditional analyses on the same simulations with a cutoff of $P \leq 10^{-5}$ discovered on average 1.5 (52%) of the three simulated effects and included 2.29 (76%) of the sites within the credible interval. Consequently, DAP does appear to improve performance, at the cost of a considerably higher computational burden.

2.4 Discussion

Studies of the genetic regulation of gene expression are making a meaningful contribution to the interpretation of GWAS results, as they provide functional insight into the nature of the causal genes. Efforts to fine map causal variants are, however, complicated by the limits of statistical resolution as it is not uncommon for tens, if not hundreds, of polymorphisms in a credible set to have similar statistical support (Gaulton et

al. 2015; Kichaev and Pasaniuc 2015). Inclusion of experimental evidence from epigenetic marks or signatures of evolutionary conservation into scores such as CADD ((Kircher et al. 2014) and CATO (Maurano et al. 2015) may improve resolution, along with methods such as RTC (Nica et al. 2010) and PICS (Farh et al. 2015) which prioritize variants based on the structure of linkage disequilibrium at a locus. In general these approaches assume parsimony, namely that there is a single variant that is responsible for the major GWAS or eQTL signature. Although it has become increasingly clear that many loci harbor multiple independent regulatory variants, I argued above that if the parsimony assumption is relaxed and it is assumed that multiple sites in strong linkage disequilibrium commonly account for a signature that is compounded into a single significant association, then the estimates from sequential conditional analysis can be highly biased.

To summarize, I find that over 5% of primary sites and more than a quarter of all causal sites are unlikely to be tagged at all; that in the presence of multi-site regulation at least 15% of all mapped sites are not in strong LD with any of the multiple imputed causal variants at a locus; and that another 10% of the associations are plausibly due to splitting of the signal due to an unimputed site. Taken together with the confirmation that over a third of all eGenes have two or more independent eSNPS, these results suggest that at least 5% and perhaps as many as a quarter of mapped credible intervals may not include the actual causal variant.

When choosing methods to perform multiple eQTL detection, I first used a conditional analysis method. Although multiple methods, like FMQTL, DAP, CAVIAR, PAINTOR, FINEMAP (Wen et al., 2015; Wen et al., 2017; Hormozdiari et al., 2014; Kichaev et al. 2014; Benner et al., 2016) have been developed to detect multiple QTLs,

most of these methods have been used to explored the confidence intervals for the causal variants, but there has not been any estimation of effect size parameters, such as β. CAVIAR, which is based on summary statistical association results, uses a greedy method to find a subset SNPs with a specific confidence (95% by default) that causal variants are included. PAINTOR uses a similar algorithm. CAVIAR enumerates all possible causal status, and to reduce the computational burden, a maximum number of causal variants must be chosen (default is 2, according to the newest version). FMQTL applies the MCMC method to explore the causal status space, and uses a posterior inclusion probability to choose the confidential interval containing possible causal variants. To reduce the computational burden, DAP was developed to explore high probability causal sites. FINEMAP uses similar a logical framework as CAVIAR, but uses a Shotgun Stochastic Search method to only explore the causal status with high probability.

Theory and simulation both indicate that if two linked sites both influence a trait, including gene expression, then multi-site models will uniformly outperform sequential uni-site ones with regard to estimation of the true effect size. When the identities of the variants are known, sample sizes of several thousand individuals are sufficient to jointly estimate their effects with high accuracy even in the presence of high levels of linkage disequilibrium with r^2 up to or even exceeding 0.9. The problem is that the identities of the variants are generally not known, and there are no established methods for comprehensive screening transcriptome-wide for localization of multi-locus local eQTL effects. The two exhaustive search algorithms, PAINTOR (Kichaev et al. 2014) and CAVIARBF (Chen et al. 2015) hold promise for detailed dissection of multi-site models at individual loci, along with the Bayesian shotgun stochastic search algorithm, FINEMAP

(Benner et al. 2016). This and the deterministic approximation of posteriors (DAP) approach to iterative refinement of multi-SNP models (Wen et al. 2016) should be evaluated for their ability to identify and estimate more of the multi-site effects than those obtained with the sequential conditional approach. I also caution that it is likely that single site effects may sometimes be artificially split into two or more linked contributions under each of these strategies. Consequently, for this study I adopted the existing standard approach of sequential conditional analysis.

I then estimated the bias in the estimates from the conditional analysis, by fitting multi-locus linear models to all of the discovered eSNPs at each locus. This revealed only modest improvements in accuracy for most of the discovered sites, but the modesty is in part an artefact of the discovery bias introduced by the sequential conditional process. The simulations assuming 2 to 4 effective sites per locus across a wide and representative range of linkage disequilibrium, show that in a sample size of 2,000, in general no more than 85% of the simulated causal sites are tagged by discovered associations that explain typically observed magnitudes of effect and would almost always be detected if a single site explained the variance. Similarly, Lloyd Jones et al. (2016) estimated that in the CAGE dataset, on average between 50% and 75% of the heritability due to locally acting regulatory polymorphism can be attributed to discovered variants. Multi-site modeling re-adjusts the remaining estimates typically by between 0.1 and 0.5 standard deviation units, which depending on the allele frequency accounts for between 2% and 5% of the variance explained, and only rarely more than 10%.

However, any variants with effect sizes greater than 1 sdu, and whose average r^2 with the other 3 SNPs is greater than 0.9, will be mis-estimated in both the single site and

joint models, typically by 1.5 sdu or more. The mis-estimation is on average the greatest where all of the effects are in the same direction, but is consistently observed also in the presence of associations with alternate signs. Since the number of loci with 2, 3, 4 and so forth discovered variants approximately halves for each additional variant, it seems reasonable to infer that at least 10% of loci actually have four or more eSNPs (two or three of which are discovered), and that perhaps one quarter have three or more eSNPs (of which one or two are discovered). I conclude that the sequential conditional estimates of eQTL effects are actually highly biased for a considerable proportion of variants.

Although this does not impact the total amount of variance explained by the discovered variants, it is likely to greatly impact fine mapping efforts, particularly where two or more effects are collapsed into one site in a credible interval. Thus, while large datasets have very good power for detection of complex regulatory contributions for individual genes, there are a host of technical and statistical reasons why fine mapping of causal variants remains a challenge. There are two immediate strong implications of these results. One is that even though the majority of identified eSNPs are expected to map to credible intervals that include the causal variant (Gusev et al. 2014; Finucane et al. 2015), there will also be many instances where incongruence between the statistical interval and chromatin or other functional evidence (Huang et al. 2015) is to be expected. The causal variant may simply be poorly mapped due to interference among linked functional sites. This effect may also influence the fine mapping of pleiotropic associations (Fortune et al. 2015).

The second implication of the high frequency of multi-site regulation is to emphasize caution in using univariate statistical support for an eSNP effect as sufficient

evidence that an association between a SNP and a trait is evidence for causation. At a minimum it is imperative that the full spectrum of eSNP effects across the locus be evaluated to confirm that the site is not simply in LD with higher likelihood eSNPs that are not themselves associated with the trait. Experimental validation of individual sites seems warranted in situations where establishment of the identity of the causal variant(s) is desired.

CHAPTER 3

PolyQTL: Bayesian multiple eQTL detection with control for population structure and sample relatedness

ABSTRACT: Expression quantitative loci (eQTL) are being used widely to annotate and interpret GWAS hits. Recent studies have demonstrated that individual gene expression is often regulated by multiple independent cis-acting eQTL. Diverse methods, frequentist and Bayesian, have already been developed to simultaneously detect and fine-map such multiple eQTL, but most of these ignore sample relatedness and potential population structure. This can result in false positives and disrupt the accuracy of fine-mapping. Here I introduce PolyQTL software for identifying and estimating eQTL effects. The package incorporates a genetic relatedness matrix to remove the influence of population structure and sample relatedness, while utilizing a Bayesian multiple eQTL detection pipeline to identify the most plausible candidate causal variants at one or more independent loci influencing abundance of a transcript. Most of contents in this chapter have published in *Bioinformatics*, as Zeng and Gibson, 2018 "PolyQTL: Bayesian multiple eQTL detection with control for population structure and sample relatedness".

3.1 Background

The idea of performing eQTL detection by genome-wide association combining genomic biomarkers and gene expression measurement was first reported in 2005 (Cheung et al., 2005). Stimulated by the great success of genome-wide association studies (GWAS) (Visscher et al., 2017; Albert and Kruglyakk, 2015), more than 100 genome-wide eQTL

studies have been performed on a variety of human tissues. Despite its great success in identifying regulatory effects, and widespread application in interpretation of GWAS results, new evidence reveals that only a limited proportion of GWAS hits seem to fine map to credible eQTL intervals. Part of the reason is that most eQTL studies are performed with a limited number of samples, usually fewer than 1,000, resulting in constrained statistical power. Owing to this low statistical power, only a limited number of genes will have significant signals, the winner's curse will lead to over-estimations of effect sizes, and non-causal variants will often have the strongest signal. To overcome these limitations, some studies, such as the eQTLGen consortium, have collected study samples from many labs around the world, and performed meta-analysis to discover cis- and trans-eQTL signals. In this type of analysis, a common procedure is to conduct principal component analysis on the genotypes and extract several PCs (usually 4), which are treated as covariates when conducting the association study. However, control for population structure with PCA may not be sufficient to rule out the influence of potential population structure or cryptic relatedness, especially when samples are collected from diverse ancestries. A more powerful strategy is to perform mixed linear modeling.

An implicit assumption of eQTL studies is that each region parsimoniously contains a single eQTL, and most of the available packages have been developed with this in mind. However, most genes are regulated by numerous regulatory elements, including promoter, enhancer and suppressor elements. These can be located several hundred kilo-bases from the transcription start site, yet still contain polymorphisms that contribute to the variance in gene expression.

A common approach to finding multiple causal variants is to first calculate marginal association statistics for each variant, and then perform stepwise conditional analysis including lead associations as covariates in each successive model (Yang et al., 2014). Investigators can then focus on the top ranked independent variants for follow-up studies. Bayesian method have also been shown to be powerful for performing association analysis and fine-mapping, and multiple packages are available, including CAVIAR (Hormozdiari et al., 2014), CAVIARBF (Chen et al., 2015), FINEMAP (Benner et al., 2016), FMQTL (Wen et al., 2015), and DAP (Wen et al., 2016). However, a major caveat that precludes their use on many datasets is that they use only summary statistics (CAVIAR, CAVIARBF, FINEMAP), ignoring any population structure and relatedness, or require external ancestry information to control for population structure in meta-analysis.

Here I present PolyQTL, a statistical approach to association analysis and fine mapping that addresses the limitations of these existing methods. The novelty of PolyQTL lies in that it provides a full powerful Bayesian framework for fine mapping of causal variants that can also be used with cohorts that include samples from related individuals.

3.2 Method

3.2.1 Remove influence of polygenic background

Suppose that the expression of a given gene is under partial control of a polygenetic background, expressed as $y=X\beta+\varepsilon$, where y is a vector of transcript abundance phenotypes with length n, X is a genotype matrix, $\beta$is a vector of random genetic effects distributed

normally N(0, V_g^2/M), M is the number of causal variants, and ε is a random factor with a normal distribution N(0, V_e^2). Covariance between two samples can be described as (1/m $\sum_{k=1}^{m} X_{ik} * X_{jk}$)* V_g^2 + $V_e^2 I$, where i and j index individuals, m is the number of total variants, and *I* is the identity matrix. When there is no population structure or relatedness present, the covariance will be almost identical to (V_g^2 + V_e^2) *I* , and a standard regressionmethod can be applied safely (Joo et al., 2016). However, when there is population structure or relatedness among individuals, the non-diagonal elements in the covariance matrix largely depart from zero, and estimation will be biased. Figure 3.1A illustrates these properties as an n x n matrix of genetic covariance where the bottom left to top right diagonal represents siblings, and the bottom left and top right quadrants imply increased similarity among individuals in two sub-populations.

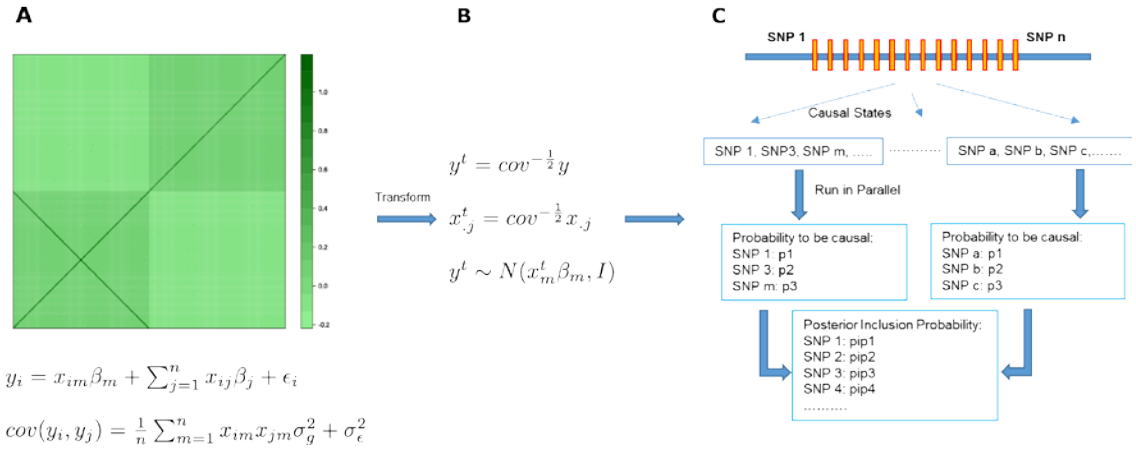

Figure 3.1 Pipeline of PolyQTL. A illustrates the complexity of simulation of related individuals and population structure. Green shading represents the correlation between pairs of samples, with twins represented by the top-left to bottom right diagonal in the bottom left quadrant, and population structure represented by the closer similarity of individuals within the bottom left and top right quadrants. B shows transformation of the phenotype and genotype with the square root of the covariance, and (C) a Bayesian method is used to compute a posterior inclusion probability (PIP) for each variant, ranking candidate variants.

In PolyQTL, I first estimate the covariance parameters V_g^2 and V_e^2 with Restricted Maximum Likelihood (REML), and to remove the influence of population structure, I then transform the phenotype and genotype by dividing by the square root of the phenotype covariance, which results in independent multivariate normal distributions (Figure 3.1B) (Abney et al., 2002). After this transformation, a Bayesian method extending previous methods is used to compute a posterior inclusion probability (PIP) for each variant, leading to a ranking of candidate causal variants (Figure 3.1C). PolyQTL can be used to estimate multiple regulatory variant effects simultaneously, which is computationally burdensome, or run in conditional mode estimating effects at each independent eQTL separately.

3.2.2 Conditional analysis

In eQTL analysis, a cis-region may be defined as 100kb, 200kb or even 1Mb around the transcription start site (TSS), usually containing hundreds or thousands variants to explore. Evaluation of all possible variants in such an interval is constrained by computational burden. To overcome this limitation, and assuming that causal variants are sparsely distributed, I provide the option of conditional analysis, in which newly detected signals are included in the regression model, allowing detection of independent signals at each locus. In each step, the covariance matrix is re-estimated, and used to control for the population structure and relatedness, by implementing the mixed linear regression component of GEMMA (Zhou and Stephens, 2012).

3.2.3 Parallelization

I utilized the C++ OpenMP library to compute causal state posterior probabilities, running the computations in parallel on the Georgia Tech Biocluster when multiple CPUs are available.

3.3 Simulation

3.3.1 Population structure

Two subpopulations were simulated, with ancestral allele frequencies x for causal variants uniformly distributed on [0.1,0.9], and subpopulation allele frequencies were sampled from a beta distribution with parameters $x(1-\text{Fst})/\text{Fst}$ and $(1-x)(1-\text{Fst})/\text{Fst}$, where Fst is the population differentiation index (Yang et al., 2014). To mimic the LD structure in real data, 100 variants in the 400kb cis-regions of randomly chosen genes in the 1,843 1000G non-African individuals (Auton et al., 2015) were used to simulate single eQTL explaining 5% of the phenotype variance. A parameter representing the maximum number of causal variants was set to be 1 both in DAP and my method for each tested interval.

To evaluate the performance in the scenario of multiple causal variants at a locus, I also conducted similar simulations, except that two causal variants were simulated, with effect sizes ranging from 4%~8%. The parameter of the maximum number of causal variants in this case was set to be 2.

3.3.2 Genetic relatedness

A complex family structure may exist in real data. In this simulation, I adopted a simplified family structure design in which I set 400 individuals to be identical twins in one of the two sub-populations. These individuals have identical genotypes, and hence

relatedness equal to 1. Since real-world datasets often include family members with lesser degrees of relatedness, I also simulated situations where 20% of the individuals were set as sibling pairs, namely with 50% allele sharing across the 5,000 simulated causal variants.

3.3.3 Large-effect eQTL and polygenic background

To mimic the LD structure in real data, 100 variants in the 400kb cis-regions of randomly chosen genes in the 1,843 1000G non-African individuals (Auton et al., 2015) were used to simulate two eQTL, each explaining 4%~8% of the phenotype variance, respectively. 600 simulations were performed in each parameter setting, each involving a different randomly selected gene. Some simulations result in both causal variants lying in the same LD block, others separate them into smaller blocks between which $r^2 < 0.3$. In the modeling step, the parameter representing the maximum number of causal variants assumed per interval was set to be 2 both in DAP and PolyQTL.

To evaluate the performance in the scenario of single causal variants, I conducted simulations in which a single causal variants was simulated, with effect size 5%. The parameter of the maximum number of causal variants was set to be 1.

For simulation of the polygenic background, 5,000 variants were chosen to be causal, and the genetic relationship matrix (GRM) was estimated from the genotypes of 4,500 of these causal variants. The genetic contribution was calculated with the GRM and the simulated heritability.

3.3.4 Control of the False Positive Rate

To evaluate how well PolyQTL controls the type 1 error (false positive rate), I performed simulations as before, except that causal variants were modeled as having a null contribution. Population structure (Fst =0.05, and 0.1) was simulated as in previous simulations in PolyQTL, and the results were compared with DAP (which ignores it in the computations). I used a conserved criterion, the Bayesian False Discovery Rate, as described in Wen, 2016. This is calculated as 1- $\sum_{i=1}^{n} PIP_i$, where PIP_i is the posterior inclusion probability (PIP) score for variant *i* in the explored region containing *n* variants, and $\sum_{i=1}^{n} PIP_i$ is a summation of PIP score of variants in one region. It is thus a measurement of the support that there is a causal variant in that region.

3.4 Results

3.4.1 PolyQTL improves statistical power to find the true causal variants.

For all simulations, I compared the performance of PolyQTL and an established method, DAP. I explored the influence of three factors: heritability of gene expression (0.3, 0.6), population structure modeled with Fst (0.1, 0.2) and relatedness (0%, 20% of samples related as identical twins), resulting in a grid of eight different scenarios. For each scenario, I performed 600 simulations. The PIP score of the modeled causal variant was used as a measurement of the statistical power to find true causal variants, and different PIP cutoffs (0.1, 0.3, 0.5) were used to evaluate the fine-mapping resolution. The PIP score from DAP was compared with results from PolyQTL, where improved statistical power is evident where the PIP value of causal variants is greater with one method relative to the other.

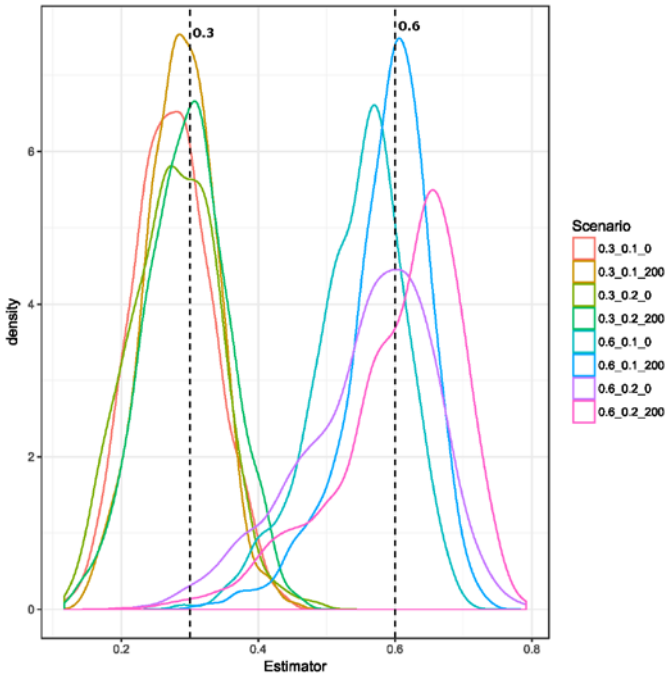

Figure 3.2 The parameter estimation of genetic components

First, I evaluated the performance of PolyQTL in estimating covariance parameters. From Figure 3.2, the estimation of genetic components is consistent with the simulated values. Simulation results reveal that PolyQTL has performance than DAP. Taking the PIP cutoff at 0.1 as an example, in the case of multiple causal variants (n=2), when there is a severe population structure (Fst =0.2), if the phenotype has a relatively low genetic contribution (heritability=0.3), 73.5% of the PIP scores computed with PolyQTL were larger than the cutoff, compared with 66.8% in DAP (Figure 3.3A). Increasing the polygenic contribution (heritability=0.6), PolyQTL had a slightly greater advantage, 75.0% vs 65.5% (Figure 3.3C). Furthermore, in the presence of a high proportion of related individuals in the samples (20% of samples related), control for relatedness brought additional power to find the causal variants, 75.1% vs 63.8% (Figure 3.3D). Similar patterns were observed with more conservative PIP cutoffs (0.3 and 0.5). When there is

only a single causal variant, the advantage of PolyQTL is limited, but performance improvement relative to DAP is in the range of 1%~5%.

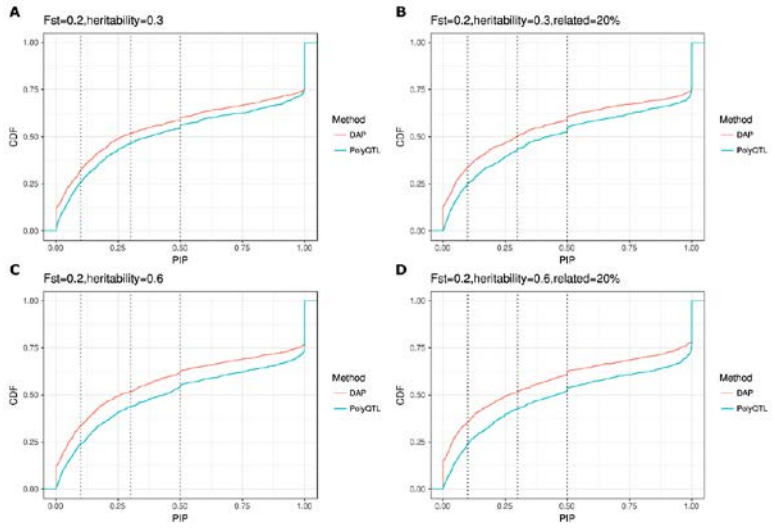

Figure 3.3. Comparison of power to detect causal variants between DAP and PolyQTL when there are 2 causal variants affecting the phenotype and Fst=0.2. The X axis is PIP ranging from 0 to 1, and Y axis is the cumulative distribution quantile (Read for DAP, and Blue for PolyQTL). Three PIP cutoff values (0.1, 0.3, 0.5) were chosen to compare the rate of discovery of causal variants. (A), (B) are low heritability (0.3), and (C), (D) are high heritability (0.6), and 20% of samples were set to be identical twins in (B), (D).

In the presence of subtle population structure (Fst=0.1), I found that PolyQTL had improved performance under conditions of high heritability (0.6), obtaining an improvement of 1.0% to 4.7% (Figure 3.4C). Reduction of the heritability to 0.3 resulted in weaker improvement, less than 1% (Figure 3.4A,B).

Even in this case of the absence of population structure or relatedness in the sample, PolyQTL has a better performance than DAP and with the increase of heritability, the advantage of PolyQTL is more obvious (Figure 3.5). This is because that PolyQTL applied a mixed-linear model to estimate the genetic component, and reduced the estimation noise.

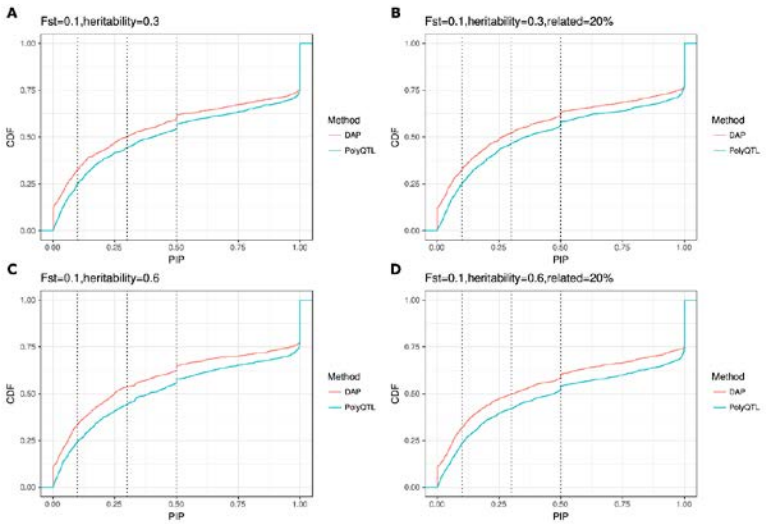

Figure 3.4. Comparison of power to detect causal variants between DAP and PolyQTL when there are 2 causal variants affecting the phenotype and Fst=0.1. The X axis is PIP ranging from 0 to 1, and Y axis is the cumulative distribution quantile (Read for DAP, and Blue for PolyQTL). Three PIP cutoff values (0.1, 0.3, 0.5) were chosen to compare the rate of discovery of causal variants. (A), (B) are low heritability (0.3), and (C), (D) are high heritability (0.6), and 20% of samples were set to be identical twins in (B), (D).

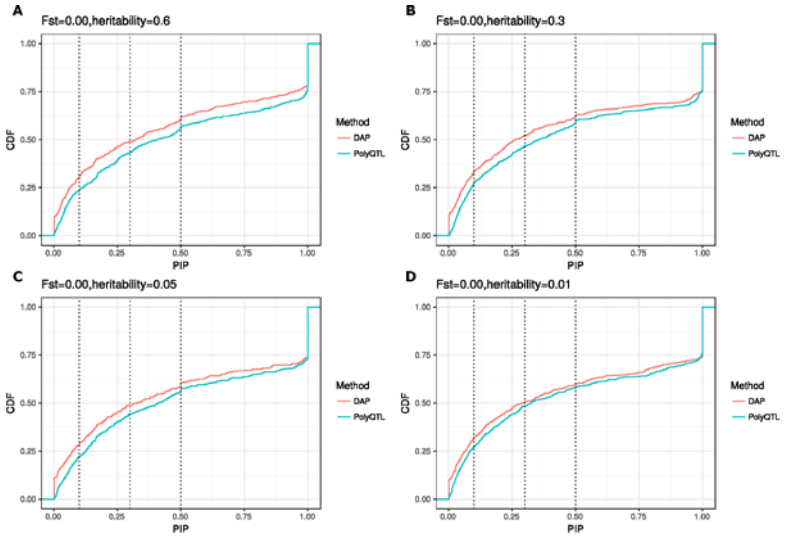

Figure 3.5. Comparison of power to detect causal variants between DAP and PolyQTL when there are 2 causal variants affecting the phenotype and Fst=0. The X axis is PIP ranging from 0 to 1, and Y axis is the cumulative distribution quantile (Red for DAP, and Blue for PolyQTL). Three PIP cutoff values (0.1, 0.3, 0.5) were chosen to compare the rate of discovery of causal variants. (A), (B), (C) to (D) are the results for gradient heritability, ranging from 0.6 to 0.01.

3.4.2 PolyQTL has a good control for false positive rate.

Except for power, another important character of a statistical method is the control for false positive rate. To evaluate the performance of the false positive rate control, I performed 600 simulations, setting the effect size of causal variant to be 0%, and use a measurement, regional PIP, to evaluate the false detection rate.

Table 3.1 Control of type 1 error in PolyQTL.

Heritability	Regional PIP cutoff	Fst			
		0.05		0.1	
		PolyQTL	DAP	PolyQTL	DAP
0.3	0.9	2/600	6/600	4/600	6/600
	0.8	6/600	10/600	10/600	12/600
0.6	0.9	0/600	8/600	2/600	20/600
	0.8	2/600	20/600	2/600	36/600

The table below (Table 3.1) indicates that both PolyQTL and DAP control type 1 error well, with an advantage for PolyQTL which becomes stronger both as heritability increases and the level of population structure increases. For example, if I choose regional PIP 0.9 as the cut-off, on average across the four scenarios only 2 out of 600 null causal variants generate a false positive signal, compared with 10 out of 600 with DAP, a five-fold improvement.

3.5 Conclusion

In this study, I developed a Bayesian multiple eQTL method, named as PolyQTL, which can be used to find multiple eQTL regulation in structured populations. I performed dense simulations to demonstrate that PolyQTL controls population structure and relatedness, improving statistical power to include true causal variants in the list of high probability eQTL SNPs. Meanwhile, PolyQTL also has a reduced false positive rate compared with the method ignoring the potential population structure of relatedness.

CHAPTER 4

Comprehensive Multiple eQTL Detection and its Application in GWAS Interpretation

ABSTRACT: Expression QTL (eQTL) detection has emerged as an important tool for unravelling the relationship between genetic risk factors and disease or clinical phenotypes. Most studies focus on analyses predicated on the assumption that only a single causal variant explains the association signal in each interval. This greatly simplifies the statistical modeling, but is liable to biases in scenarios where multiple linked causal-variants are responsible. Here, my primary goal was to address the prevalence of secondary cis-eQTL signals regulating peripheral blood gene expression locally, utilizing two large human cohort studies, each greater than 2,500 samples with accompanying whole genome genotypes. The CAGE dataset is a compendium of Illumina microarray studies, and the Framingham Heart Study is a two-generation an Affymetrix dataset. I also describe Bayesian co-localization analysis of the extent of sharing of cis-eQTL detected in both studies as well as with the BIOS RNA-seq dataset. Stepwise conditional modeling demonstrates that multiple eQTL signals are present for ~40% of over 3,500 eGenes in both microarray datasets, and that the number of loci with additional signals reduces by approximately two-thirds with each conditioning step. Although fewer than 20% of the peak signals across platforms finemap to the same credible interval, the co-localization analysis finds that as many as 50%~60% of the primary eQTL are actually shared. Subsequently, co-localization of eQTL signals with GWAS hits detected 1,349 genes

whose expression in peripheral blood is associated with 591 human phenotype traits or diseases, including enrichment for genes with regulatory functions such as protein kinase activity and DNA binding, and I found that adding non-primary cis-eQTL to conduct co-localization analysis improves our chance to detect co-localized signals, and contribute to 10%~40% of significant co-localized results. Just one quarter of these co-localization signals replicated, further highlighting the technological and methodological barriers to reconciliation of GWAS and eQTL signals. The results are provided as a web-based resource for visualization of multi-site regulation of gene expression and is association with human complex traits and disease states.

4.1 Background

Since the first GWAS results were published in 2005 (Klein et al. 2015), thousands of genetic regions in human chromosomes have been found to be associated with human phenotypes including disease states (Visscher et al. 2017). Since it is now assumed that the majority of SNP-trait associations identified by GWAS can be attributed to effects on gene expression, precise estimation of the location and effect sizes of regulatory polymorphisms has become important for understanding the relationship between genetic and phenotypic variation (Maurano et al. 2012; Farh et al. 2015). The minimal expectation is that eQTL analysis can identify the gene within a locus that accounts for a GWAS signal, although it has become clear that even this goal is a far from trivial (Chung et al. 2014; Pickrell et al. 2014). Many investigators make the stronger assumption that co-localization of regulatory variants (eSNPs) and GWAS signals to a tight linkage disequilibrium interval implies the

ability to define if not the causal variant, then at least a credible set of SNPs that include the causal site(Trynka et al. 2013; Gaulton et al. 2015; Kichaev and Pasaniuc 2015; Liu et al. 2015).

However, high resolution fine mapping eQTL results aligned with GWAS studies for diverse phenotypes has as yet provided only a few instances with unambiguous evidence that a specific variant affects a human complex trait or diseases through its affect on gene expression. Several recent studies have begun to question the presumed identity of eQTL and GWAS hits: even though there is a highly significant overlap at the level of the locus, it is not so clear that the precise variants are the same. For example, Farh et al (2015) estimated that only ~10% of the GWAS hits take function as eQTL despite the vast majority of those hits mapping to non-coding DNA. Similarly, two recent studies of autoimmune disease have also argued that only approximately one quarter of examined GWAS loci may act as eQTL in the profiled immune cells (Chun et al. 2017; Huang et al. 2017). Furthermore, work based on GTEx gene expression profiling aiming to integrate GWAS and eQTL results found that only a minority of GWAS loci match precisely to eQTL, while the diversity of regulatory effects across tissues can complicate interpretation (Hormozdiari et al. 2016; Gamazon et al. 2018). These results raise the question of why there are so many instances of discordant fine localization: are we simply limited by the low statistical power to detect association signals (Udler et al. 2010), is there mis-estimation of signal strength and location in the case of multiple eQTL per transcript (Zeng et al. 2017), or are regulatory effects so cell-type and context-specific that true co-localization is often missed? In this studies, I will focus on the first two issues by addressing the concordance of signals in two large eQTL datasets where the expectation was that,

despite technical differences between the platforms, shared cis-eQTL signals at the gene level would map to the same credible intervals.

The detection of eQTL is dependent on the accuracy of two technologies designed to estimate transcript abundance (gene expression) and to genotype or impute genetic variants. Genotype calling, whether based on gene chip platforms or whole-genome sequencing, is thought to be highly accurate and robust (1000 Genome Project Consortium, 2015), and methods for imputation of missing genotypes are now generally accepted to be valid for minor allele frequencies of 0.01 or lower. Constraints on gene expression measurement are more problematic, being subject both to the properties of the detection method and of the algorithms use to statistically analyze the data. Microarrays, principally Illumina- and Affymetrix- based for human studies, have been used widely to measure gene expression and have supported the development of expression QTL (eQTL) analyses. By far the largest published study is the 12,000 sample Blood eQTL compendium assembled by Lude Franke and colleagues (Westra et al. 2013), now approaching 30,000. However, the nature of microarray probes provides incomplete coverage of the exons within genes, and there are analytical limitations due to dynamic range of quantitative detection of expression, with the result that estimates of transcript abundance are strongly platform-specific. eQTL artefacts are also known arise due to linkage disequilibrium between regulatory variants and SNPs located with transcript probes. Nevertheless, well-powered studies have detected primary eQTL for over half of all expressed genes in blood, providing ample opportunity to compare the fine-mapping of these signals (Lloyd-jones et al. 2016).

A small number of studies have argued for high replicability of eQTL detected on the same platform. Genotype-Tissue Expression (GTEx) project discovered eQTLs from post-mortem analysis of over 40 tissues, finding extensive sharing of promoter-proximal signals for around half the loci. Zhernakova et al (2017) found that 84% of previous cis-eQTL genes detected with Illumina platform replicated in an RNA-seq data set, the vast majority showing the same direction of allelic effect. Multiple Illumina-based peripheral blood studies carried out on different cohorts by different groups have also reported in excess of 70% shared signals for eQTL detected at 5% false discovery rates (Zeller et al. 2010; Lloyd-jones et al. 2016). However, differences between platforms seem to be much larger than expected; for example, Liang et al found that only between one quarter and one third of eQTL association signals in the MRCE Illumina-based study replicated in a companion MRCE Affymetrix study (Liang et al. 2013). The differences may in part be due to the differential effects of alternative splicing on transcript abundance detected with probes that cover one or a few exons (Illumina) or more of the extent of each gene (Affymetrix), or to the effects of the normalization and other statistical procedures that are used to associate genotypes with transcript abundance estimates. It is also important to recognize that what is described as a shared signal based where a genotype associates with gene expression in two studies may often simply reflect linkage disequilibrium between two independent signals.

Consequently, methods have been developed to evaluate and fine-map co-localization signals, whether across gene expression platforms, or between eQTL and GWAS signals. Most of the current methods seek to distinguish true co-localization from "shared" signal due to linkage disequilibrium. COLOC was one of the first Bayesian

methods which evaluates the relative statistical support of each eQTL-GWAS co-localization hypothesis contingent on LD (Giambartolomei et al. 2014). However, COLOC assumes the default model that a single-causal eQTL exists, which implies a strong prior that variants taking function as eQTL (or associated with a trait), also affect the trait (or expression), potentially leading to false positive co-localization. SMR, or Summary Mendelian Randomization, jointly evaluates the strength of eQTL and GWAS signals using a procedure known as HEIDI to filter heterogeneity of GWAS and eQTL signals in the presence of LD (Zhu et al. 2016). However, SMR is strongly dependent on the accuracy of LD inference from a reference panel, and the HEIDI test has been reported to be conservative. Another Bayesian method, eCAVIAR, calculates a posterior probability of eQTL-GWAS co-localization while allowing for multiple signals in the interval (Hormozdiari et al. 2016). The dependencies of all these methods on sample size has not been well characterized, and it is found only around 50% agreement between them in evaluation of causal variants in a Crohn's disease study (Marigorta et al. 2017). Furthermore, lack of control for population structure or relatedness requires further modification when applied to data sets with large sample size.

In this study, I collected cis-eQTL results from three data sets, and developed a statistical pipeline to achieve the following goals: (a) to evaluating the prevalence of multiple cis-eQTL regulation in human peripheral blood; (b) to estimate the extent of eQTL signal sharing across three expression platform; and (c) to detect co-localization of eQTL signals with GWAS hits contingent on the LD at each locus, revealing the possible biological regulatory mechanisms linking genetic variants to complex human phenotypes.

4.2 Materials and Methods

4.2.1 Datasets

I analyzed three different peripheral blood eQTL data sets. The Consortium for the Architecture of Gene Expression, CAGE dataset consists of Illumina HT12 v3 microarray-based gene expression profiles, as well as whole genome genotype information, from five research studies: the Brisbane Systems Genetics Study (BSGS, N=926) (Powell et al. 2012), Atlanta-based Centre for Health Discovery and Well-Being (CHDWB, N=439) (Wingo and Gibson 2015) and Emory Cardiology Genebank (N=147, Kim et al. 2014), Estonian Genome Centre - University of Tartu (EGCUT) study (N=1065, Schramm et al. 2014), and the Morocco Lifestyle study (N=188, Idaghdour et al. 2010), for a total of 2,765 individuals. IRB approval was obtained for the combination of data into a mega-analysis both by the University of Queensland and for each participating site.

The second dataset from the Framingham Heart Study (FHS) (Huan et al. 2015) contains two-generation data generated on Affymetrix genechips. A total of 5,075 participants with both genotype and gene expression information from the offspring (N = 2,119, 8th examination) and third-generation (N = 2,956, 2nd examination) cohorts were included in this study. Raw genotype and gene expression data were downloaded from dbGAP (phs000007.v25.p9) with IRB approval.

The BIOS RNA-seq summary data was derived from a meta-analysis of results for a total of 2,100 participants from four cohorts (Zhernakova et al. 2017): the Cohort on Diabetes and Atherosclerosis Maastricht (CODAM, 184 individuals included); LifeLines-DEEP (LLD, 626 individuals included); the Leiden Longevity Study (LLS, 654 individuals included); and the Rotterdam Study (RS, 652 individuals included). I downloaded the

summary results of cis-eQTL signals from https://genenetwork.nl/biosqtlbrowser/, so were unable to perform the sequential stepwise regression analyses to detect secondary signals.

4.2.2 Genotype Imputation

Genotype imputation for the CAGE cohort was performed jointly for the five contributing studies to ensure uniformity of assignment of strand identities of SNPs, and is described in detail in Lloyd-Jones et al. (2016) and at https://github.com/CNSGenomics/impute-pipe. Briefly, the pipeline involved pre-imputation quality control, and data-consistency checks, imputation to the 1000G reference panel with Impute2 (Howie et al. 2012), post-imputation quality control (filtering on various data features), and merging of the datasets on common SNPs.

For the Framingham Heart Study data, there were a total of 6,950 individuals before imputation, from which 29 individuals with genotype missing rate ≥5% were removed. Subsequently any SNPs with genotype missing rate ≥5% were also removed along with SNPs with Hardy-Weinberg test P $\leq 10^{-6}$. Prior to imputation, the genotypes were pre-phased using shapeit2 (Delaneau et al. 2013) using the "duohmm" parameter to account for pedigree information. Each chromosome was divided into 5Mb chunks, incorporating the centromere-adjacent region (acen region) into the neighboring chunk, and similarly joining any chunk with <200 SNPs into a neighboring chunk. Imputation was performed with Impute2 (Howie et al. 2012), using qctool to convert gprobs to gen file format, and only SNPs with info value larger than 0.3 were retained for subsequent analyses. The gen file was converted to plink file format, and SNPs with multiple information and InDel variants were filtered out. The remaining SNPs were further reduced to ~6 million SNPs

with >95% call rate across all 5,075 individuals represented by both genotype and gene expression data.

4.2.3 Probe Re-annotation

Since SNP imputation for the CAGE cohort was based on hg19/GRCH37, whereas the Illumina probe annotation was based on hg18/GRCH36, Ire-annotated the probe information by mapping the probe sequences to hg19/GRCH37 with BWA (Li and Durbin 2009), retaining only the uniquely mapped probes. All probe sequences were secondarily mapped to the reference genome with BLAT (Kent 2002), and only probe sequences uniquely mapped with both methods were determined to be high confidence and subsequently used for eQTL detection. Of a total of 45,931 probes mapped to the reference genome, 7,349 probe sequences mapped to multiple regions or remained unmapped, leaving 38,582 probes taken forward for the eQTL analyses. See Table S1 for summary statistics. Since it is well-known (Walter et al. 2007) that SNPs in a probe influence microarray hybridization, I also discarded 3,856 Illumina probes containing SNPs with maf >1% in the 1000 Genomes European sample (Lappalainen et al. 2013). Similarly, SNPs in the Affymetrix probesets were also converted to positions in the hg19 assembly by applying liftOver (UCSC Genome Browser) to the GPL5188 annotation file downloaded from dbGAP, and annotated to the 1000 Genomes. Any SNP with a maf >1% in the 1000 Genome European population and located within a probeset was deemed to be potentially unreliable, and was included as a covariate during the eQTL estimation steps. Among the 280,000 core probesets, 35,000 have such SNPs, and 15,368 transcripts contain at least one SNP in a probe.

4.2.4 Gene Expression Normalization

The gene expression normalization strategy for CAGE required aggressive procedures to account for study-specific biases, as described in detail in Lloyd-Jones LR et al. (2017). It consisted of 5 steps: (1) Variance stabilization using the vsn package (Lin et al. 2008); (2) Quantile normalization forcing the intensity distribution across all probes to have the same shape for all samples; (3) Batch effect correction via linear regression to account for known technical effects, such as RNA extraction date, and physical batch; (4) Batch effect correction (via principal component analysis, removing the first 10 PC to account for unknown confounding procedural, or population-based influences); and (5) Rank normal transformation, namely a final transformation of each probe to a normal distribution with mean 0 and variance 1.

For the FHS data, raw gene expression processed by Affymetrix APT software (version 1.12.0) was downloaded from dbGAP, log2 transformed, and surrogate variable analysis (SVA) (Leek et al. 2012) was used to remove confounding factors, fitting a total of 62 surrogate variables by a linear regression model. Note that the FHS gene expression study (Huan et al. 2015) reported results of a different normalization that included fitting blood cell counts, which I chose to avoid since a similar procedure was not applied to the CAGE data, and because the blood counts were also removed by the SVA fitting.

4.2.5 Multi-site eQTL Detection

For this study, local SNPs were stringently defined as SNPs located within 200 kb upstream or downstream of the gene (defined as the first TSS and last TES listed in the hg19 annotation) containing the probe. Sequential conditional analyses were performed for

each probe, and the genes with significant eSNPs were called eGenes. Since both the CAGE and FHS cohorts contain family-based data (the former for a quarter of the samples, from the BSGS twin study (Powell et al. 2012); the latter for all participants), a mixed linear model was used for eQTL detection in GEMMA (Zhou and Stephens 2012), which fits a genetic relatedness matrix (GRM) as a covariate alongside fixed genotype effects. The multiTrans tool (Joo et al. 2016), which accounts for family structure, was used to specify a study-wise false discovery rate of 5% for genes with multiple independent eSNPs, which was empirically observed to be approximately $P < 10^{-5}$. After first scanning for evidence of at least one local eSNP at this threshold, the residuals after fitting the sentinel SNP were used for a sequential conditional scan for an independent secondary eSNP. This process was iterated until no more signals were observed below $P = 10^{-5}$. SNPs in high LD with each previously detected signal ($r^2 \geq 0.9$) were also filtered out of each sequential analyses. The effect sizes of each discovered SNP were recorded as the sequential conditional estimates. Subsequently, for the multi-site effect size estimates, all discovered independent peak SNPs were fit with the GRM in one mixed model. However, since the GEMMA software does not report the effect sizes of all fixed effects simultaneously, I fit the multi-site models with one SNP specified as the target effect, including the other significant SNPs, as well as the GRM, as covariates. This estimation procedure was repeated for each included SNP, recording the effect size of the target SNP as the multi-site effect, noting that the amount of variance explained by each gene's model is the same for all such models. To control the influence of SNPs located in probes in the FHS data, I incorporated in-probe SNPs with an LD r^2 cutoff 0.75 as covariates during the multi-site modeling step.

4.2.6 Fine-mapping with PolyQTL

Fine-mapping to localize causal variants influencing gene expression was performed using PolyQTL (Zeng and Gibson, 2018), a modification of DAP (Wen and Pique-Regi, 2016) which I developed to account for population structure and ancestry during Bayesian localization in the presence of multiple linked cis-acting variants. I incorporated an option for first performing sequential stepwise regression, using the mixed linear regression component of GEMMA (Zhou and Stephens, 2012) as above to isolate independent QTL. PolyQTL also offers the option to estimate posterior probabilities for all eQTL at a locus simultaneously, but this was not performed here owing to the computational burden.

PolyQTL assumes that there is a single causal variant associated with each independent QTL, and evaluates the posterior probability, given the LD structure at the locus, that each variant in the interval is causal, such that the sum of the posterior probabilities for each independent QTL is between 0 and 1. Genes were modeled as being under partial control of local genotypes as well as the polygenetic background, expressed as $y = X_i\beta_i + G + \varepsilon$, where y is a vector of transcript abundance phenotypes, G represents the influence of the polygenic background, X_i and β_i are the genotype and effect of the explored variant, and ε is a random environmental factor also normally distributed $N(0, V_e^2)$. PolyQTL uses REML to estimate genetic and environmental variances, V_g^2 and V_e^2 given the estimated GRM, K (Yang et al, 2010). To remove the influence of population structure, we transform the phenotype (y) and genotype (X_i) with the square root of the

covariance of the phenotype, ($\widehat{V_g^2}$K + $\widehat{V_e^2}$I), where I is the identity matrix, as this results in independent multivariate normal distributions. We then compute a posterior inclusion probability (PIP) for each variant, leading to a ranking of candidate causal variants (Zeng and Gibson, 2018).

4.2.7 eQTL sharing across expression platforms

Despite the expectation that expression platform influences eQTL detection, I reasoned that cis-eQTL results can complement one another leading to enhanced detection of shared signals by overcoming false negative results from single studies. To this end, I performed joint analysis of the cis-eQTL signals obtained on all three platforms, namely Illumina, Affymetrix, and RNA-seq. I devised a new method based on the eCAVIAR strategy (Hormozdiari et al. 2016), named DPolyQTL, which explores the signal sharing for two phenotypes (either molecular traits or phenotype traits) even where the collected samples are family-based or from diverse ethnicities. DPolyQTL calculates a posterior probability that the causal variants are shared for two phenotype traits, such as expression of a gene measured on two platforms, by multiplying the two posterior probabilities together to generate a colocalization posterior probability (CLPP: Hormozdiari et al. 2016).

Since interpretation of the calculated posterior probability as a shared causal variants is confounded by the complex LD structure in human genome, I conducted permutations to obtain the null distribution of the posterior probability given that a true eQTL is detected in one of the datasets, the discovery dataset, is replicated in the other one, the replication dataset. To do so, the phenotype was permuted in the replication dataset,

and the posterior probability was re-calculated. On this basis the co-localization signal was determined to be true if the CLPP ≥ 0.001 and permutation P-value ≤0.05.

4.2.8 eQTL and GWAS co-localization analysis

Summary results were downloaded for 1,263 phenotype traits or disease from eQTLgen Consortium. For each trait or disease, I defined candidate independent candidate regions as all variants within 100kb of a peak association signal at $P \leq 5\times10^{-8}$. To reduce the computational burden, I excluded all variants in the interval with P ≥0.05.

Co-localization of the eQTL and GWAS signals was then assessed for all genes located within 1Mb of the peak GWAS signal. Similar to the analysis of eQTL sharing between expression platforms, I conducted DPolyQTL on the GWAS and eQTL summary statistics to identify regulatory influences of gene expression on complex phenotypes or disease.

4.3 Results

4.3.1 Multiple eQTL regulation is ubiquitous in human blood

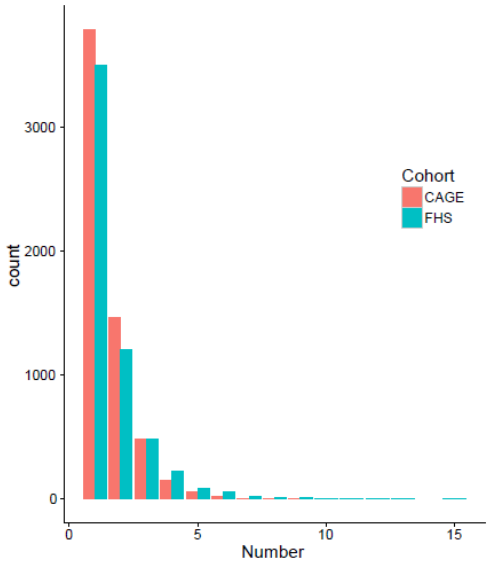

Figure 4.1 Detected independent cis-eQTL in CAGE and FHS cohort.

Our first objective was to estimate the proportion of loci that have multi-site regulation in the two large cohort studies, CAGE and FHS. Since both datasets include siblings, I used GEMMA (Zhou and Stephens 2012) to perform sequential conditional eQTL analysis deploying a genetic relationship matrix based on all measured and imputed genotypes to model family structure and population structure. Applying sequential conditional analysis to CAGE, I detected 5,974 eGenes (37.8% of 15,812 tested genes) with at least one significant eSNP at $P < 10^{-5}$. Of these eGenes, 2,187 (36.6%) contain probes influenced by more than one eSNP and hence appear to be regulated by multiple enhancers. (Note that in the case of genes with multiple probes on the Illumina platform, Ionly required that at least one probe was associated with an eQTL, and for multi-SNP regulation included only the probe with the largest number of significant independent eSNPs). Similarly, in the FHS data, I detected 5,597 (35.3% of 15,853 tested genes), 2,098 (37.5%) of which were regulated by multiple eQTLs. In CAGE, the average variance

explained by detected eSNPs was 6.1%, the same as in FHS, 6.1%, and in both cases these estimates account for more than half of the previously estimated cis heritability (Lloyd-Jones et al, 2017). For those genes with multiple eQTL regulation in CAGE, which have a mean explained variance of 7.2%, the newly detected secondary eSNPs typically explained 20% more variance than the peak SNP alone, namely ~1.2% of the phenotypic variance (6.0% vs 7.2%), also in line with estimates from Lloyd-Jones et al. (2017). For eGenes with multiple eQTLs in FHS, the mean explained variance is 6.3%, and the secondary signals increase the explained variance from 6.5% to 8.3%, an ~28% increase.

Table 4.1. Cross-platform comparison of eSNP detection after adjustment for probe SNPs

	With SNPs-in-Probes			Without SNPs-in-Probes			
No.	CAGE	FHS	Both_any	CAGE	FHS	Both_any	Both_highLD[#]
≤ 1	3175	3874	1330	3787	3499	1186	474
≤ 2	5571	5989	1685	5246	4699	1442	616 (34)
≤ 3	6113	6929	1777	5733	5182	1518	669 (7)
≤ 4	6280	7327	1805	5881	5407	1539	686 (1)
> 0	6383	7713	1812	5974	5597	1565	689 (0)

[#] Cumulative number of eGenes with at least 1 eSNP localized within $r^2 > 0.8$ in both CAGE and FHS, number in brackets indicates cases with 2, 3, or 4 eSNPs all in high LD between datasets.

Figure 4.1 shows frequency histograms for the number of detected eQTL per gene after each sequential step in both studies: the number of loci with additional

independent sites reduces by approximately two thirds with each additional SNP in both CAGE and FHS, up to half a dozen variants, and a few loci have 10 sites. This reduction likely reflects the true prevalence of multi-site effects as well as reduced power to detect SNPs that explain less of the variance than the primary signal. A detailed example of multi-site association is shown for the HBZ locus in CAGE (Figure A4.1), where from left to right, and top to bottom are the results of stepwise conditional analysis yielding 9 independent eQTL signals. The total explained variance is 39.8%, one third more than the 28.4% explained by the highest single-site signal. An example from the FHS is ABHD2 where I detect 5 independent eQTLs explaining 9.3% of the variance, compared with 5.6% for the peak eSNP (although the Affymetrix probeset contains a common variant, rs2283435, that is in linkage equilibrium with each of the five regulatory signals). All of the multiple eQTL results can be downloaded both in tabulated format and as locuszoom plots from Prof. Gibson's lab server at given URL.

I also computed the difference between the estimates fitting all discovered variants jointly and the conditional single-site estimates following eSNP sequential conditional discovery. The average change in estimated beta was 0.04 sdu, plus or minus 0.06 due to a long tail of large deviations.

To compare directly the degree of signal replication from cohorts based on a same platform, I also evaluated the level of replication between the contributing studies in CAGE. For examples, contrasting eQTL detection for chromosome 1 genes between the CHDWB and EGCUT cohorts at FDR<=0.01, I detected 665 significant independent eQTL in EGCUT, and 364 in CHDWB, 315 of which (86.5% of the peak eSNPs) located

in the same credible interval (genotype r^2>=0.8). This result further validates the mega-analysis strategy of combining the different CAGE study cohorts into one large study.

4.3.2 A Bayesian fine-mapping approach increases the power to detect cis-eQTL sharing

4.3.2.1 The r^2 criterion results in poor overlap of cis-eQTL between CAGE and FHS.

Direct comparison of primary results from the CAGE and FHS analyses suggests a disappointingly low level of replication. Primary peaks in CAGE were detected for 53.0% of the eGenes represented in the FHS, and reciprocally 56.5% of the FHS eGenes had primary signals in CAGE, very similar to the proportions reported for eSNPs at $P < 10^{-8}$ across four peripheral blood studies (Zeller et al. 2010) that also had a variety of technical differences. However, the overall overlap between CAGE and FHS for eSNPs within credible intervals defined by LD r^2>0.8 is just 29.1%. Furthermore, only 41.5% of the primary signals in FHS are in LD (r^2>0.8) with the primary eSNP in CAGE, suggesting that different largest-effect regulatory variants are tagged in the two datasets. This overall eSNP replication rate was slightly higher (47.2%) when mapping to 1,314 probesets which map directly to the same exon and have an eQTL signal on both platforms.

The replication rate of secondary, tertiary, and quaternary signals in FHS irrespective of LD was just 19.0%, 11.3%, and 11.0%, indicating successive decay, likely due to reduced power for weaker signals. The reasons for the discrepancies between the studies may have to do with collapsing of probe-level data down to gene-level signals losing information on splice isoforms, the different normalization strategies (which alone can double discovery rates (Qin et al. 2012)), and cross-study biological heterogeneity. Comparison of the percent variance explained by discovered SNPs on the two platforms in

Figure A4.3 shows that effect sizes for many genes are disproportionately tagged by eQTL in the two studies, implying platform effects. Among 139 genes with more than 20% of the variance explained by cis-eSNPs, the replication rates are 64.7% for the primary signal, 34.8% for the secondary, 14.4% for the tertiary, and 8.5% for the quaternary. The subsequent panels confirm that all of these replication rates are proportional to the percent variance explained overall, confirming that statistical power is a major source of low replication.

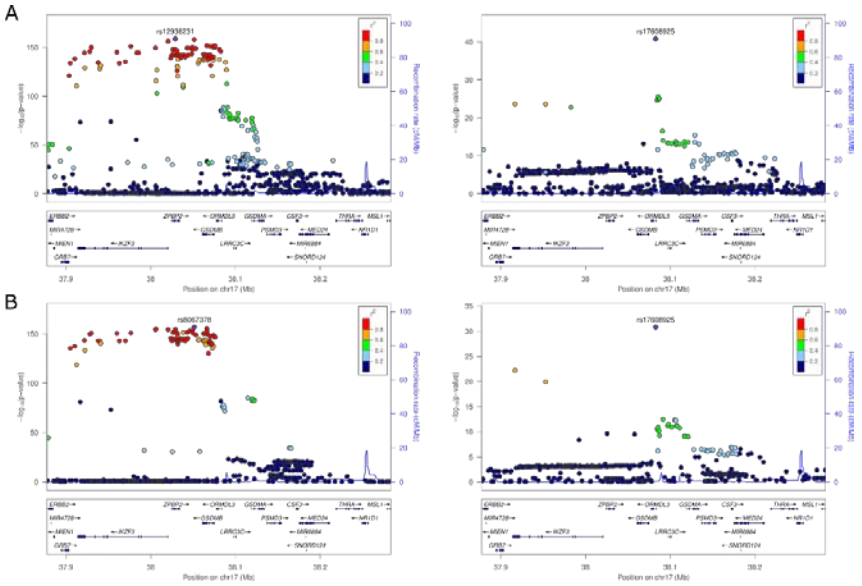

Figure 4.2 An example of shared cis-eQTL signals in CAGE and FHS. I detected two cis-eQTL for ORMDL3 in both CAGE and FHS. For CAGE, two independent rs12936231, and rs17608925 were found to be associated with expression abundance. In FHS, rs17608925 was detected to be independent peak signal, while another independent peak variant, rs8067378 was detected. The genotype r^2 is 1 between rs12936231 and rs8067378.

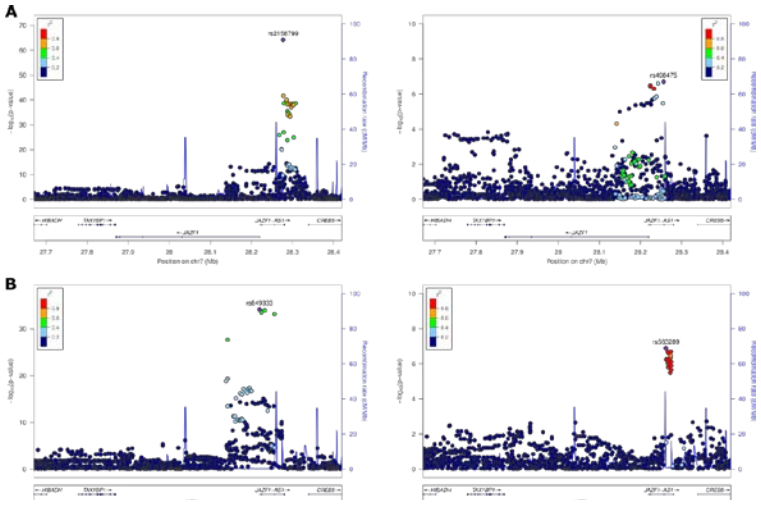

Figure 4.3 An example of rank-changed cis-eQTL signals in CAGE and FHS. I detected two cis-eQTL for ORMDL3 in both CAGE and FHS. For CAGE, two independent rs2158799, and rs498475 were found to be associated with expression abundance (Panel A). In FHS, I detected a primary signal, rs849333, which locates in high LD with the secondary signal rs498475 in CAGE (genotype r^2=0.90), while another independent peak variant, rs563289 was also detected (Panel B).

For the HT12 v3 Illumina probes, 10% of the uniquely mapping probes contain at

least one SNP with MAF>1% in 1000 genome European population. The prevalence of eQTL was twice as great for these probes (59% versus 30% of 23,681 "clean" probes), so Ijust removed the Illumina probes containing SNPs in order to control the false discovery rate. However, since most of the Affymetrix probesets contain at least one SNP, this was not practical for the FHS dataset and instead I employed a conditional analysis strategy incorporating SNPs in probes as covariates. For ~15,000 detected eSNPs, one third of the association signals were abrogated by conditioning on the SNPs in probes, and the number of eGenes correspondingly reduced by 25%. Table 1 contrasts the eQTL results from both platforms before and after controlling for the SNP-in-probe effects. The first three data columns show the cumulative number of eGenes with at least 1, 2, 3, 4, or more detected eSNPs before SNP-in-probe removal, and the next three show the cumulative numbers after. The proportion of overlapping signals is not greatly affected. The last column shows that the number of eGenes where at least one detected signal is likely capturing the same variant is around 44% (689/1565), and that the number where all of the multiply detected signals are within $r^2 > 0.8$ is very small. There are 214 genes with at least two signals in high LD with one another.

4.3.2.2 DPolyQTL increases the proportion of cis-eQTL sharing across different expression platforms

Next, I used DPolyQTL to enhance the power to detect shared cis-eQTL credible intervals in the CAGE, FHS, and BIOS datasets. I extracted each locus by considering variants locating in high LD with the reported peak variants in each eQTL study, and calculated a posterior probability to demonstrate the the likelihood that each variant influences the trait controlling for LD at the locus. Since the available BIOS dataset only

consists of summary results, it was used solely as a discovery dataset. Where genes in CAGE and FHS contained multiple probes or probe-sets, replication is reported where at least one probe in each dataset contains a signal.

Table 4.2 Sharing of cis-eQTL among expression platforms.

Platform	Discovery	Replicate	
		CAGE	FHS
Illumina	CAGE	-	62.60%
Affymetrix	FHS	53.30%	-
RNA-seq	BIOS	44.70%	54.50%

Tables 4.2 reports the cis-eQTL sharing among CAGE, FHS, and BIOS datasets. Shared signals are indicated for 62.6% of the detected cis-eQTL of CAGE in FHS, and for 53.3% of the FHS eGenes in CAGE. For the detected cis-eQTL in BIOS, I found a similar replication rate, namely 53.6%, and 54.7%. Considering that BIOS only reports the eGenes and that for some genes no expression information available in CAGE and FHS, the replication rate considering all genes is 44.7% in CAGE, and 54.5% in FHS. Since DPolyQTL is statistically flexible, it allows multiple eQTL signals to be explored simultaneously. On this basis, 43%~49% of primary eSPNs showed evidence for replication, but the rate was considerably lower, only ~10%, for secondary eSNPs.

Taken together these results indicate greater than 50% cross-platform replication of eGenes across platforms, with evidence that the majority of primary eQTL detected on one

platform are also eQTL on another. However, the primary regulatory variant maps to a different credible interval in more than a third of the cases, and replication of secondary variants is strongly reduced by low statistical power in the presence of multisite regulation.

4.3.3 Biological Annotation of detected multiple eQTLs

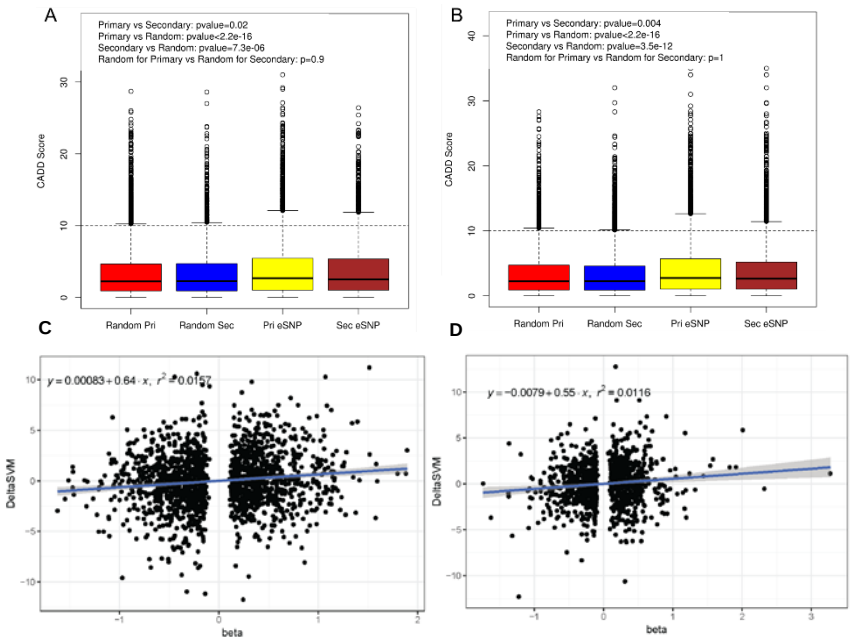

Figure 4.4 Biological annotation for the detected cis-eQTL signals.

Since chromatin marks are often used to enhance fine-mapping, on the basis that peak eSNPs are enriched in the vicinity of ENCODE features such as DNAse hypersensitivity, methylation, and histone modification, I asked whether there is a difference in functional attributes of primary and secondary eQTL. The CADD and deltaSVM scores are two commonly used annotation tools that summarize multiple types of functional evidence. For the CADD score, to determine the significance of enrichment, I created a list of background SNPs with similar allele frequency in the neighboring regions and compared the annotation of eQTLs with that of background SNPs. The distribution of CADD scores was significantly higher for the reported-peak variants, suggesting elevated

likelihood that they are pathogenic and typically selected against (figure 4.2A for CAGE, 2B for FHS). However, although significant, the magnitude of the effect is small relative to the variance in CADD scores and the positive predictive value for each SNP is low. Similarly, potential causal variants defined with the fine-mapping step have slightly elevated probability of locating to regulatory enhancers in human genome defined by the deltaSVM score. Setting any variant with a posterior probability ≥0.8 as a causal variant, I found that there is a significant positive relation between the reported beta value and deltaSVM ($P \leq 10^{-6}$ in both CAGE and FHS, figure 4.2C, 4.2D), though again the overall correlation is weak.

4.3.4 Interpretation of GWAS results

4.3.4.1 eGenes associated with phenotypes are enriched for certain molecular functions

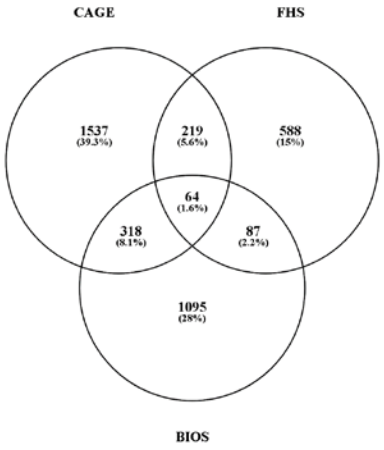

Figure 4.5. Replication of eQTL-GWAS co-localization with different expression platform

In this section, I aimed to identify genes whose expression also associates with phenotypic traits reported in dbGaP. I combined the full summary statistics of 1,263 GWAS results with eQTL signals from CAGE, FHS, BIOS, maximizing statistical power

by performing co-localization analysis based on cis-eQTL detected on all three platforms. This co-localization analysis resulted in 1,349 genes associated with 591 human complex phenotype traits or disease (49. 8% of explored). The highest single platform discovery rate was for the CAGE data on the Illumina platform, and the replication rate across platforms ranged from 24% to 30% (Figure 4.5).

Enrichment analysis with the PANTHER database (Mi et al, 2016) revealed that genes annotated to protein kinase activity or to DNA binding activity were over-represented. PANTHER Pathway analysis further showed that genes involved in Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade (4.2 fold enrichment, 8.5×10^{-4}), VEGF signaling pathway (3.2 fold enrichment, 4.4×10^{-4}), Interleukin signaling pathway (3.1 enrichment, 8.6×10^{-5}), Ras Pathway (2.7 fold enrichment, 2.4×10^{-3}), PDGF signaling pathway (2.3 fold enrichment, 6.0×10^{-4}), Gonadotropin-releasing hormone receptor pathway (2.0 fold enrichment, 7.2×10^{-4}), and Inflammation mediated by chemokine and cytokine signaling pathway (1.92 fold enrichment, 1.1×10^{-3}), were enriched. Furthermore, these 1,349 detected genes were enriched for association with several disease, 327 causing Mendelian diseases (1.4 fold enrichment to background, $p=8.6\times10^{-7}$), providing further evidence that genes defined by highly penetrant mutations also harbor quantitative regulatory variants that influence disease.

4.3.4.2 Co-localization detects known and novel eGenes affecting phenotype and disease

The co-localization results highlight a number of genes that have been reported to affect phenotypic traits through gene expression. For example, I found 5 genes associated

with coronary artery disease, PSRC1, IL6R, LIPA, SWAP70, and VAMP8 (Figure 4A) in the CAGE dataset. Four of these genes have previously been reported to be associated with coronary artery disease. PSRC1 encodes a cysteine protease which has been associated with HDL and LDL levels (Kathiresan et al, 2008), and its expression in mouse liver is significantly associated with plasma LDL cholesterol level (Schadt et al, 2008). LIPA encodes lipase A, which catalyzes the hydrolysis of cholesteryl esters and triglycerides, and in previous studies, it is associated with CAD, where the lead CAD risk allele correlates with increased expression of LIPA mRNA in monocytes (Zeller et al, 2010) and liver (Coronary Artery Disease Genetics Consortium, 2011). SWAP70 encodes a signaling molecule involved in the regulation of filamentous-actin networks in cell migration and adhesion. An intronic SNP has been reported to be a cis-eQTL in naïve and challenged monocytes (Nikpay et al, 2015). Notably, rheumatoid arthritis has been associated with cardiovascular disease prevalence, yet there is little genetic support for this relationship. I found that expression of the IL6R gene is associated with both rheumatoid arthritis and coronary artery disease. I also found that co-localization signals from different expression platforms complement one another, and can capture more biological insights. Thus, with FHS gene expression, there are 11 genes also associated with CAD: ADAMTS7, CARF, CDKN2A, GGCX, HECTD4, IL6R, LIPA, PCSK9, PSRC1, USP39, VAMP8, including four of the CAGE genes (IL6R, LIPA, VAMP8, and PSRC1). Of the remaining genes, manual review of the literature finds that 5/7 have been reported to be associated with CAD.

Our co-localization analysis also identified genes relevant to multiple linked traits or diseases. In previous studies, IKZF3 was reported to affect the autoimmune diseases

Crohn's disease, ulcerative colitis and rheumatoid arthritis. These findings are replicated in my data, and enhanced further by associations with the additional autoimmune diseases, type 1 diabetes, and primary biliary cirrhosis as well as with asthma (Figure 4.4B). Expression of IKZF3 is also associated with neutrophil cell and white blood cell counts. Most of these co-localization signals are replicated in FHS or BIOS data.

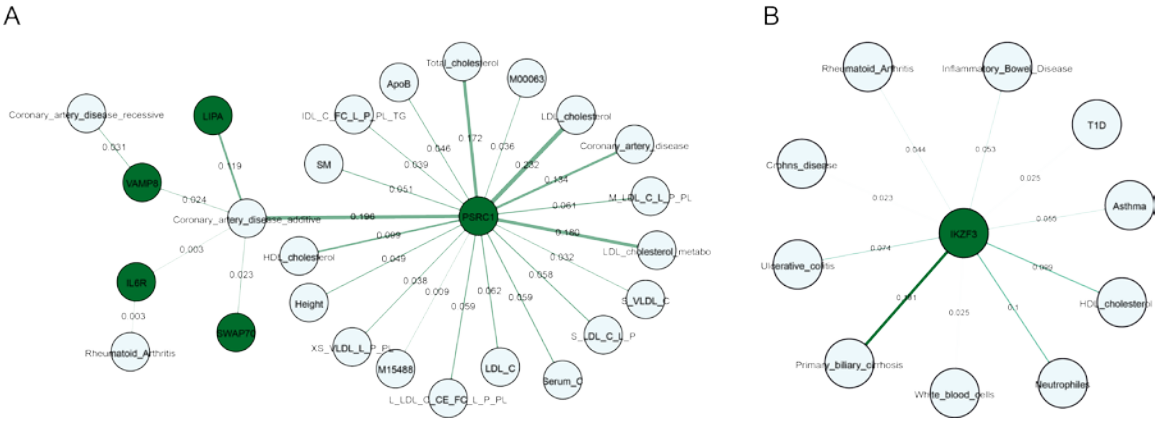

Figure 4.6 Two examples of eQTL-GWAS co-localization. One example is that IKZF3 is previous reported to be associated with diverse auto-immune diseases. In my analysis, I not only replicated previous findings, and also found the association with other auto-immune diseases and some immune-associated phenotype traits. Another is for the detected genes whose expressions were demonstrated to affect CAD.

4.3.4.3 Gene expression regulated by non-primary cis-eQTL mediates genetic effect to phenotype and disease

In previous co-localization studies, it is reported that only limited proportion of GWAS variants take function as eQTL (Farh et al. 2013; Chun et al. 2017). However, those studies mainly focus on primary cis-eQTL. I expect that co-localization analysis based on additional eQTL should increase our chance to find the co-localized eQTL-GWAS signal. For each expression-phenotype pair with at least significant co-localized signal, I first find variants showing sign of colocation in the credible interval of primary cis-eQTL, and if

variants are found, I assert that the co-localization results from primary cis-eQTL. If none variant is found, then switch to secondary cis-eQTL, tertiary signals, so on and so forth until no cis-eQTL remains.

For 2,138 co-localized signals in CAGE, I detected 82.0% of the co-localized signals are from primary eQTL, and 11.0% from secondary signals, 3.9% from tertiary signals. In FHS, there are 958 co-localized eQTL-GWAS, 69.5% are from primary signals, and 15.1% from secondary signals, and 11.9% from tertiary signals. For instance, like shown in Figure 4.3, I detected two independent cis-eQTL, rs2158799 and rs498475 for JAZF1 in CAGE, and the secondary cis-eQTL, rs498475 locates in high LD (r^2=0.92) with a type 2 diabetes GWAS hit, rs849135, and co-localization analysis does find a significant signals. The co-localization results indicate a great contribution of non-primary cis-eQTL.

4.3.5 eQTLHub: A R application based on shiny package to visualize the multiple cis-eQTL results, and the co-localization signals of eQTL-GWAS.

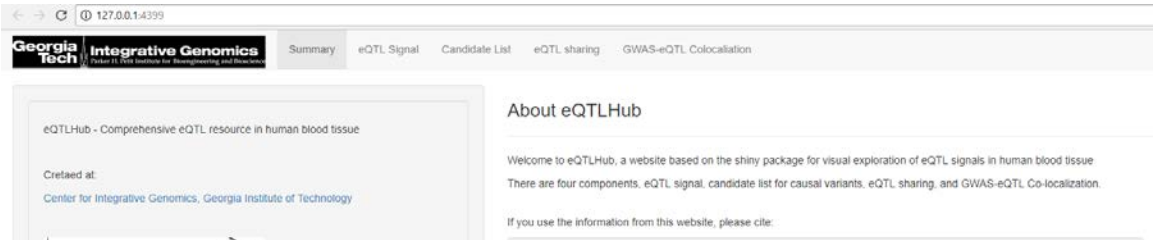

Figure 4.7 Interface of eQTLHub providing access to multiple eQTL results and eQTL-GWAS co-localization.

To present the multiple eQTL results, I developed eQTLHub, a website based on the shiny package for visual exploration of eQTL signals in human blood tissue. There are four components, eQTL signal, candidate list for causal variants, eQTL sharing, and GWAS-eQTL Co-localization.

4.4 Discussion

In summary, I find extensive evidence for secondary and tertiary cis-eQTL associations explaining gene expression variation in peripheral blood. At least a third of all highly expressed genes display such effects, consistent with recent evidence from very large-scale GWAS that at least one quarter of loci harbor multiple associations within a Mb interval. However, the fine mapping of eQTL across platforms is considerably lower than expected and accordingly replication of co-localization with visible phenotypes and disease risk is also modest, despite the large sample size of my two cohorts. Since secondary and tertiary effect sizes are generally smaller than primary ones, statistical power remains a major detriment to the joint fine mapping of regulatory variants to GWAS credible intervals.

Resolution of GWAS associations to single causal variants is a major current objective of human genetics. Four general strategies are being deployed: very large GWAS and eQTL studies, including cross-population analyses, intended to narrow peaks; sophisticated co-localization approaches; filtering on functional attributes associated with SNPs; and high-throughput experimental validation. The first objective is to define credible intervals that are highly likely to contain the causal variant or variants within a linkage disequilibrium block. However, several recent studies have reported that as few as one third of disease associations map to the same credible interval as the lead eQTL, even in cases such an autoimmune Crohn's disease where the eQTL mapping is carried out in the presumably relevant peripheral blood tissue consisting of immune cells (Huang et al. 2017; Chun et al. 2017).

Two classes of explanation may account for this discrepancy between expectation and observation: biological and technical. The obvious biological explanation is that the causal variant detected by GWAS for some phenotypic trait does not directly regulate gene expression. It may for example influence chromatin structure, preparing the locus for induction under conditions not sampled in the transcriptomic study (Alasoo et al. 2018), and indeed there is some evidence for greater overlap of methylation QTL than expression QTL with Crohn's disease associations (Huang et al. 2017). A corollary would be that the influence on gene expression would only be seen if the appropriate tissue (or the most important cell type within a mixture of cell types, such as peripheral blood) is sampled, or under more appropriate conditions of stimulation either ex vivo or in vivo (such as inflamed tissue-resident immune cells). The prevalence of response-eQTL provides good evidence in support of this claim (Fairfax et al. 2014).

Technical explanations relate to statistical methodology and power, as well as platform effects. It is remarkable in my study that both the Illumina and Affymetrix datasets yielded very similar proportions of eGenes, as well as distributions of secondary and tertiary signals. Yet the overlap between these signals was only approximately one half for the primary eSNPs, and less than 20% for conditional associations. Implementation of DPolyQTL provided evidence that statistical power is a major source of failure to replicate, both by enhancing the detection of shared primary signals between the datasets and showing that detection rates drop as effect sizes of secondary, tertiary and quaternary associations reduce. Nevertheless, it is also clear that platform effects result in major differences in blood cis-eQTL detection. These are only partially ameliorated by

focusing on probes that capture the same exon within a transcript, implying that detection of alternate splicing and isoform usage is just one aspect of the platform effect.

Irrespective of the causes of differential localization of primary eSNPs, an important practical implication of my findings is cautioning against the common use of summary eQTL statistics as evidence that a GWAS hit acts as an eQTL. Given the extensive linkage disequilibrium typically observed over long stretches of regulatory DNA, it is not uncommon for the GWAS variant to be included in a list of eQTL highly significant summary statistics listed on browsers such as the Blood eQTL browser. Visual inspection of the profile of association across the locus will often be sufficient to illustrate that the eQTL and GWAS peaks are not actually the same. Formal tests of the hypothesis of equivalence are provided by software tools such as COLOC, but these are designed for supervised analysis locus-by-locus and may be biased by the assumption that a single causal variant is responsible for each eQTL effect. The HEIDI test in SMR attempts to adjust the inference that an eQTL mediates the phenotypic association for local LD, providing genome-wide estimation of cases of heterogeneity of effects. Alternatively, the Bayesian eCaviar approach, implemented here in DPolyQTL to adjust for population structure and familial relatedness, more directly adjusts for LD in the derivation of posterior probabilities of joint association. I recommend using a combination of these approaches to explore the likelihood that eQTL explain GWAS effects, and to this end have developed a web browser which for the first time allows users to explore the profile of primary and secondary signals in peripheral blood.

Contrary to the expectation that mega-analysis of large eQTL studies would improve the resolution of eQTL signals, I instead find levels of complexity that complicate

the ability to reduce genetic associations to single causal variants. Most clearly, it is apparent that multiple regulatory variants affect the expression of the majority of transcripts expressed in peripheral blood. Similarly, meta-analysis of GWAS including hundreds of thousands of individuals increasingly find secondary associations at individual loci (Wood et al. 2014; GIANT Consortium, 2018). I have previously shown by simulation that the presence of multiple variants in LD blocks typical of human genes biases both the localization of eSNPs and the estimation of their effect sizes, with as many as 20% of effects potentially located outside detected credible intervals (Zeng et al. 2017). While functional data collected by the ENCODE project and measures of evolutionary conservation are often used to filter or adjust eQTL estimation, my analyses only confirm a modest enrichment of such marks at eQTL peaks. Elsewhere it is shown that this is in large part due to the high correlation of functional scores within credible intervals (Liu et al. 2018). Consequently, functional assays will continue to provide the gold standard for demonstration that specific SNPs associate with trait function through gene regulation.

CHAPTER 5

Trans-eQTL detection in the FHS cohort and cis-trans eQTL mediation analysis

ABSTRACT: Although cis-eQTL have been demonstrated to be powerful in interpretation of GWAS results, it can only provide local regulation information, while trans-eQTL can be used to find downstream genes and perform gene network analysis. However, trans-eQTL detection has been blocked by the fact that trans-eQTL have smaller effect size, and also the huge burden of multiple test correction. In this analysis, I applied a statistical strategy based on mixed linear model to conduct trans-eQT detection. Results reveal that the method has a good control for false positive, and the detected trans-eQTL have a consistent estimation with other results based on large sample size studies. Comparisons of eQTL signals among expression platform found that cis-eQTL detection is more easily affected by platform-specific factors than trans-eQTL. Co-localization analysis reveals that most trans-eQTL take function as cis-factor on the local genes to affect downstream gene expression.

5.1 Background

The advance of high-resolution genotyping technology has led to a wave of genome-wide association studies (GWAS) of hundreds of phenotypes relevant to human health and disease. Yet, the vast majority of detected hits from GWAS significantly

associated with clinical traits and disease states locate in non-coding regions, and are assumed to function as regulatory factors instead of changing protein function. Thus, to improve understanding of biological mechanisms, exploring the relationship between genetic variants and RNA expression abundance is a critical step toward ultimate improvements in diagnosis, prevention, and treatment of disease. This endeavor begins with analysis of variation in messenger RNA (mRNA) expression levels associated with genotypic variation to identify expression quantitative trait loci (eQTLs) across the human genome.

Statistical methods in eQTL detection are similar to those used in GWAS, and the statistical strategies developed for GWAS can be directly applied for eQTL studies. However, eQTL detection involves thousands of genes and hundreds of thousands variants, or even millions of variants after imputation. For cis-eQTL, analysis is simplified because only several hundred or thousand variants are used and effect sizes tend to be large, while for trans-eQTL, there is not only a greater computational burden, but control for multiple test correction reduces statistical power. For these reasons, currently only cis-eQTL are widely explored, and trans-eQTL detection has lagged behind. A consortium of researchers is needed to collect a large sample size to gain statistical power for trans-eQTL analysis.

In contrast to cis-eQTLs, analysis of trans-eQTLs is vastly more computationally challenging and reported trans-eQTLs have proven to be less replicable across studies. Therefore, many eQTL studies focus only on cis-eQTLs or a subset of variants associated with phenotypic traits or important biological functions. However, according to results from studies estimating gene expression heritability (Lloyd Jones et al, 2017), at least half of expression heritability can be accounted for by trans-eQTL factors. When SNPs at a

trans-eQTL locus affect the expression of multiple genes, the region is usually referred to as a trans-eQTL hotspot. Cis-eQTLs typically reside close to transcription start sites (TSSs), suggesting that they directly impact gene expression. The mechanisms by which trans-eQTLs alter transcription of their linked trans-eGenes are largely unknown and likely often reflect indirect or cryptic regulation. For example, it has been proposed that expression of trans-eGenes could be mediated by transcription factors encoded by genes located close to the corresponding trans-eQTLs. This phenomenon suggests that cis-eQTLs might influence the expression of master regulators for a large number of trans-eGenes, in the manner of biological networks.

In this chapter, I report results of an Affymetrix microarray-based genome-wide eQTL study, considering both cis and trans elements, in whole blood samples from over 5000 participants in the Framingham Heart Study (FHS), a multi-generational community-based prospective study. In this analysis, I aimed to ascertain to what extent SNPs affect genes in cis and in trans and to determine whether eQTL mapping in peripheral blood could identify downstream pathways that might be drivers of disease processes.

5.2 Material and methods

5.2.1 Study cohort

The Framingham Heart Study (FHS) is a cohort study initiated in 1948, with the aim of identifying risk factors for heart disease. Starting in 1971, the offspring and offspring spouses (N = 5,124) of the original FHS cohort participants were recruited and they have been examined approximately every 4 years since. From 2002 to 2005, the adult children (third generation cohort, N = 4,095) of the offspring cohort participants were

recruited and are also being examined in an ongoing manner. For this study, a total of 5,075 participants who provided both genotype and gene expression information from the offspring (N = 2,119) and third-generation (N = 2,956) cohorts were included. Whole blood samples were collected at the eighth examination of the offspring cohort and the second examination of the third generation cohort. Fasting peripheral whole blood samples (2.5 ml) were stored in PAXgeneTM tubes (PreAnalytiX, Hombrechtikon, Switzerland) and the Affymetrix Human ExonArray ST 1.0 (Affymetrix, Inc., Santa Clara, CA) was utilized to measure mRNA expression levels. Genotyping was performed with the Affymetrix 500K mapping array and the Affymetrix 50K gene-focused MIP array. Genotype imputation was conducted using impute2 against 1000 Genomes Phase 3 reference.

5.2.2 Trans-eQTL Detection Pipeline for FHS

Imputation results were converted to bgen format with genotype dosage as independent variables. A genetic relatedness matrix was constructed using all of the imputed genotypes using GEMMA (Zhou et al., 2012). For the gene expression data, the first 20 non-genetic PCs of gene expression were regressed out, and residuals were used as adjusted phenotypes for the association studies.

Prior to trans-eQTL detection, cis-eQTL were first detected by stepwise sequential conditional analysis. Variants located within a distance of less than 1Mb from either the 5' or 3' ends of the explored gene coding region were deemed to be cis, and only SNPs with a minor allele frequency (MAF) of >0.01 and a Hardy-Weinberg equilibrium P value of >0.001 were included in the analyses. In each iteration of the conditional analysis, the peak signal with a P value$<10^{-6}$ also computed using the GEMMA package, was determined to

be an independent eSNP. The residuals from discovery of each SNP were then taken forward as the dependent variable in a new scan for additional independent SNP(s).

After cis-eQTL detection, residuals removing all cis-eQTL effects at each gene were used as the phenotype, and trans-eQTL detection was performed on 10,562 variants previously shown to be associated with phenotypic traits. The signal was evaluated with a mixed linear model in GEMMA, controlling for population structure and relatedness. To obtain the null distribution by permutation, the covariance component was firstly estimated by REML, and the square root of the covariance matrix was used to transform the phenotype and genotype matrices (Abney et al., 2002), after which, the transformed phenotype is exchangeable, and can be safely used to conduct permutation analysis. The false discovery rate was controlled relative to 10 phenotype permutations, retaining the co-expression structure by permuting the sample IDs which were used in common for all expression phenotypes.

5.2.3 Effective Population Size Estimation

As one of the cohorts in the eQTLGen consortium, my trans-eQTL results were integrated into the final meta-analysis results. The Framingham Heart Study was the only dataset in the study which did not consist of unrelated individuals but was instead a family-based cohort. Although the analysis strategy took family relationship into account, I needed to determine the effective sample size to use the proper weight for this dataset in the weighted Z-score meta-analysis. For a specific variant, the effective population size can be calculated with the formula:

$N_{eff} = \frac{var(y)-2p_i(1-p_i)\beta_i^2}{2p(1-p)*var(\hat{\beta})}$, in which, p is the minor allele frequency of the explored variant, and $var(\hat{\beta})$ is the variance of the estimator. The effective population size of one gene can be estimated with the mean or median N_{eff} of genome-wide variants by:

Formula 1: $N_{eff} = \frac{\sum_{i=1}^{m}[var(y)-2p_i(1-p_i)\beta_i^2]/[2p_i(1-p_i)S_i^2]}{m}$ or

Formula 2: $N_{eff} = Median(var(y) - 2p_i(1-p_i)\beta_i^2]/[2p_i(1-p_i)S_i^2)$, $i \in \{1,2,...,m\}$ is the SNP index.

I selected 20 random genes and used eQTL effects for genome-wide SNPs to estimate the effective sample size (N_{eff}).

5.2.4 Trans-cis eQTL Co-localization Analysis

I used cis-eQTL results from 31,684 blood samples and a subset of variants that are associated from the trans-eQTL data for the 4,339 FHS samples to evaluate the co-localization between cis-eQTL and trans-eQTL signals. A total of 4,397 independent trans-eQTLs were each treated as a single eQTL. As to the cis-eQTL signals, cis-eQTL genes were only included in the coloc analysis if they had more than 200 variants shared with tested trans-eQTL. Next, I used eQTLGen meta-analysis Z-scores, allele frequencies and sample sizes to estimate beta and var(beta), using equations outlined in the Supplementary Text of (Zhu et al., 2016). In the formula, approximate sample sizes of 31,000 for cis-eQTLs and 4,300 for trans-eQTLs, together with minor allele frequencies from eQTLGen (without Framingham Heart Study cohort) were used. To test for co-localization, the coloc v3.1 package (Giambartolomei et al., 2014) was used with default priors (1×10^{-4} for both,

cis- and trans-eQTL, and 1×10^{-5} for sharing of eQTL signals). As cis-eQTLs were regressed out from expression matrices prior to trans-eQTL mapping, I expected that co-localization signals would be unlikely to reflect spurious correlation between cis- and trans-eQTL genes caused by unknown confounders.

I downloaded curated Gene Ontology gene sets (2018. year version) (Ashburner et al., 2000; The Gene Ontology Consortium, 2017) from the Enrichr web site (Chen et al, 2013; Kuleshov et al, 2016). Those gene sets were used to conduct hypergeometric over-representation analyses as implemented into the R package ClusterProfiler (Guangchuang et al., 2012), while using all the cis-eQTL genes showing co-localization with any trans-eQTL (coloc PH4>0.8) as a test set and all the cis-eQTL genes included to the co-localization analysis as the background.

5.3 Results

5.3.1 Most of trans-eQTL variants affect neighboring genes.

Co-localization was tested between the subset of 10,683 cis-eQTL genes and 2,950 trans-eQTL genes. In summary, 181,821 co-localization tests were conducted between a cis-eQTL and a GWAS locus with a trans-eQTL, of which 2,776 resulted in H4>0.8 (Coloc hypothesis 4, posterior probability of one causal SNP explaining both the eQTL and GWAS signals), suggesting that the cis-eQTL gene mediates the trans-effect in 1.5% of all such comparisons.

418 of the cis-eQTL genes (4%) have at least one downstream gene co-localizing with H4>0.8, suggesting cis-trans co-localization for an average of 6.6 such targets per cis-

eQTL (compared with 0.26 per all tested cis-eQTLs). Some of the cis-eGenes have as many as 100% of the potential targets significant. 1,528 of the trans-eQTL genes (52%) have at least one cis-eQTL gene with a colocalizing eQTL effect (H4>0.8), strongly suggesting possible cis-trans co-localization, an average of 1.8 cis-mediators per trans-eQTL (compared with 0.9 per tested trans-eQTLs). No more than 33% of the potential mediators for any given trans-eQTL were significantly co-localizing (H4>0.8).

GO enrichment analysis found that trans-eGenes are enriched for transcription factors, but clearly not all trans-eGenes encode transcription factors.

5.3.2 Conditional analysis based on GEMMA controls the false positive rate

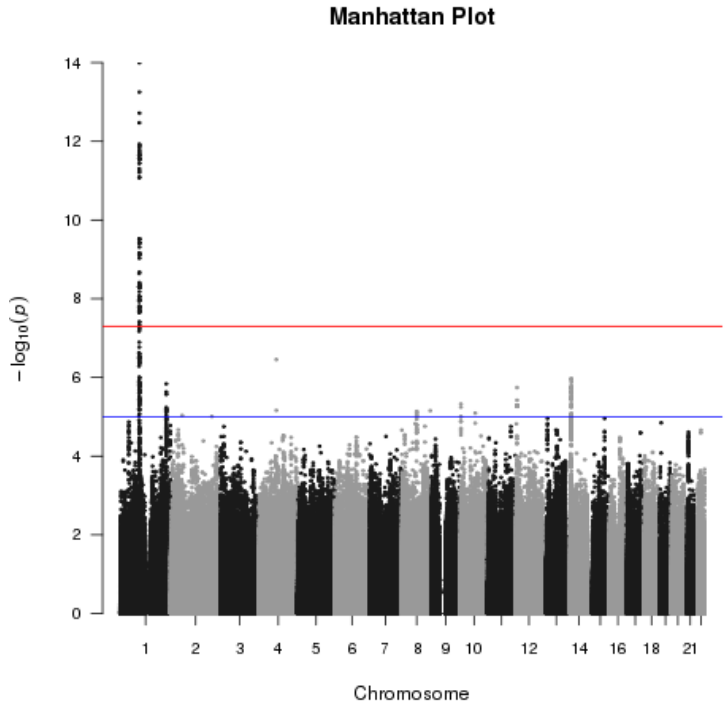

Figure 5.1 Manhattan plot of trans-eQTL for DNTTIP2 on Chromosome 1.

Most of the samples in the FHS are genetically related with other samples, so to control for the relatedness and the potential population structure, I applied GEMMA's mixed linear model. GEMMA is a widely used method in mixed linear modelling. Compared with other MLM-based methods, GEMMA is an efficient exact method, estimating variance components for each variant resulting in a better control for false positive and also a large statistical power.

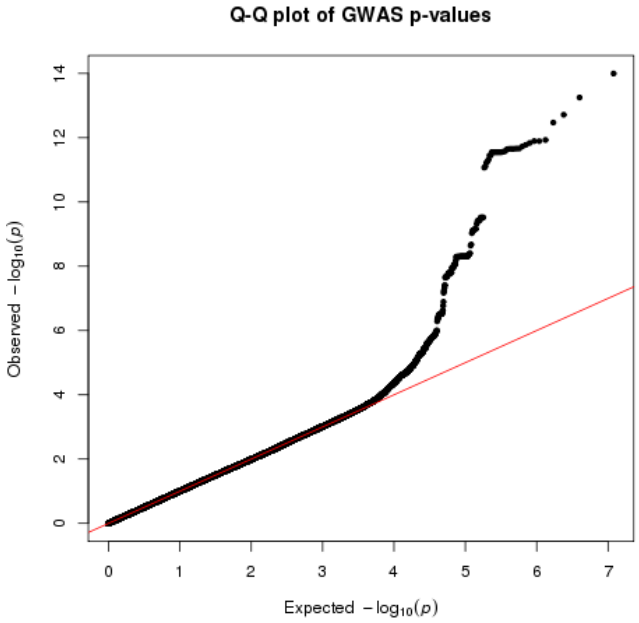

Figure 5.2 QQ-plot for the DNTTIP2 gene.

To evaluate the control of the false positive rate, I randomly chose 20 genes, and conducted eQTL detection with genome-wide variants. Analysis of genomic inflation factors for the 20 chosen genes shows that all these genes have a lambda around 1 (Table 5.1), indicating good control of the false positive rate. Figure 5.1 and Figure 5.2 show the Manhattan plot for one trans-eQTL signal, and corresponding Q-Q plot, for quality evaluation of a randomly chosen gene, DNTTIP2. Despite the excess of small P-values greater than NLP4 at target trans-eQTL, there is no genome-wide inflation.

Table 5.1 Genomic inflation factor for 20 random genes in FHS.

Gene	Heritability	Lambda
DNTTIP2	0.032	1.001
HELLS	0.322	0.947
PTPN11	0.049	1.001
PARP4	0.205	0.970
ZFYVE26	0.057	0.994
LPCAT2	0.213	0.997
SNX20	0.062	1.009
VPS4A	1E-05	0.996
PNMT	0.007	0.997
OR7G2	1E-05	0.968
CBLN4	1E-05	0.993
DHX35	0.079	0.988
CELSR1	0.036	1.000
LAMP3	0.055	1.007
NSUN3	0.420	1.000
CASP3	0.180	0.990
THAP6	0.034	1.003
POM121L12	1E-05	0.967
CDKL5	0.105	0.997
TAF7L	0.023	1.005

5.3.3 Effective sample size estimation.

As the effective sample size was not dramatically different from the actual sample size (mean N_{eff} over all 20 genes was 4,837; median N_{eff}=4,865 as compared to actual N = 5,075; Table 5.2), relatedness is not likely to influence the results of integrated meta-analysis considerably, so I opted to use the actual sample size as a weight in the meta-analysis.

Table 5.2 Effective sample size of random 20 genes in FHS.

Gene	$N1_{eff}$	$N2_{eff}$
DNTTIP2	4907	4976
HELLS	4383	4829
PTPN11	4850	4970
PARP4	4506	4445
ZFYVE26	4835	5033
LPCAT2	4500	4600
SNX20	4828	5024
VPS4A	5026	4900
PNMT	4999	4977
OR7G2	5026	4984
CBLN4	5031	4971
DHX35	4769	5073
CELSR1	4901	4744
LAMP3	4837	4810
NSUN3	4336	4519
CASP3	4559	4648
THAP6	4900	4813
POM121L12	5027	4828
CDKL5	4702	4593
TAF7L	4938	5005

5.3.4 Comparison with trans-eQTL estimates in the eQTLgen Consortium

With the strategy developed to detect trans-eQTL in structured populations, I conducted trans-eQTL in the FHS data set. I applied the criteria that any variant was determined to be significant, if there is no absolute Z-score from permutations larger that in real data, which was demonstrated to be enough to have a FDR control below 0.05 (Westra et al., 2013). After eQTL detection, I compared my results with other data in the eQTLgen Consortium.

Before the detection of trans-eQTL, I first conducted cis-eQTL detection, and a total of 7,274 associations were found to be significant. Cross-validation with other platform data indicated a platform-bias. For FHS cis-eQTL, the best concordance was found in NTR-NESDA, which is also an Affymetrix platform: among the explored 6,388 (87.8%) FHS associations, 50.1% replicated in NTR-NESDA, and 97.2% of the shared signals had same direction of effect (Figure 5.3). By contrast, in an Illumina data set, 4,510 (62.0%) of the FHS associations could be tested, and 74.6% of the tested associations were replicated with only 89.7% showing consistent direction of effect (Figure 5.4). Similarly, replication in an RNA-seq dataset was also less than for the same platform. These results serve as a further reminder for researchers to be aware of platform-specific biases in eQTL analysis.

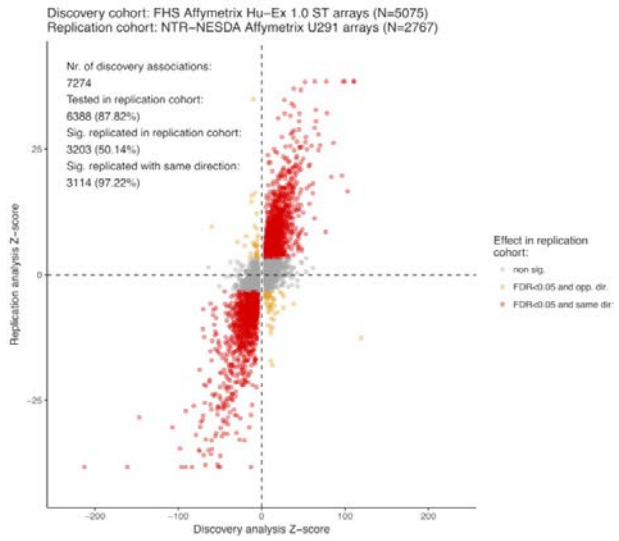

Figure 5.3 Cis-eQTL comparison between FHS and Affymetrix data set.

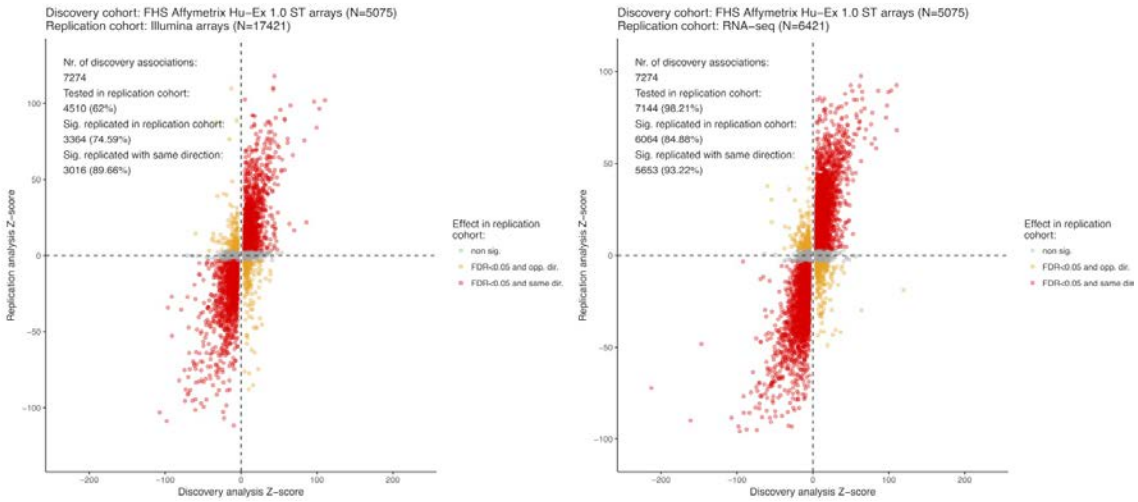

Figure 5.4 Cis-eQTL comparison between FHS and Illumina (left) and RNA-seq (right) data sets.

By contrast, trans-eQT results are free of plat-form bias, and shown high replicate rate. In summary, I detected a total of 4,422 significant associations in FHS dataset. To validate the accuracy of my trans-eQTL detection, I performed comparisons with trans-eQTL results based on other data set in eQTLgen Consortium (Figures 5.5, 5.6). Cross-comparison demonstrates that trans-eQTL results in FHS have similar concordance levels.

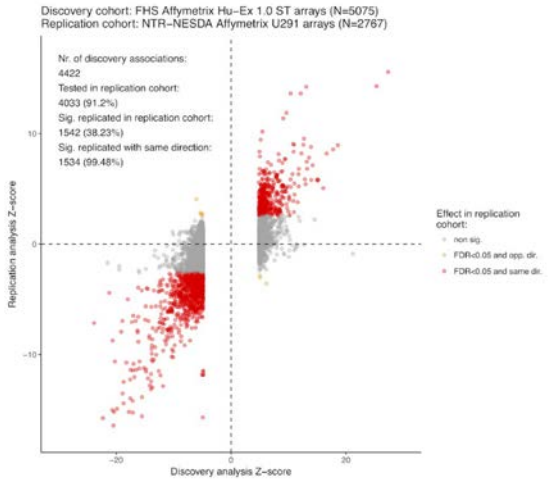

Figure 5.5 Trans-eQTL comparison between FHS and Affymetrix data set.

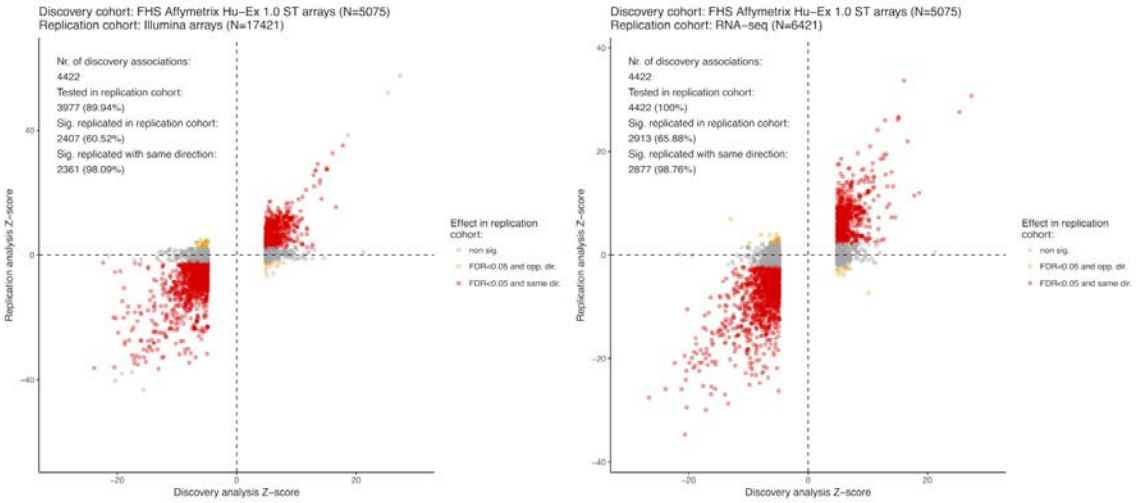

Figure 5.6 Trans-eQTL comparison between FHS and two Illumina (left) and RNA-seq (right) data sets

Comparing with meta-analysis results of 1,7000 individuals on the Illumina platform, I contrasted 3,977 (89.9%) associations, and 60.5% of the tested associations from FHS replicated, with 98.1% of these showing the same direction of effect size. Similarly, I measured the replication rate in RNA-seq trans-eQTL results, and all of the significant trans-eQTL in FHS could be tested, of which 65.9% replicated, 98.8% showing the same direction (Figure 5.6). I thus did not detect a higher concordance with NTR-NESDA based on Affymetrix U291 array, since even though (because of the smaller sample size (2,767)), only a third could be evaluated, 99.5% had the same direction of effect (Figure 5.5).

5.4 Conclusion

In this chapter, I mainly described the trans-eQTL method in a structured population. In trans-eQTL detection, the major hindering factor is that millions of variants are tested, so multiple test correction reduces statistical power. To overcome the limitation and improve the statistical power to find trans-eQTL signals, I applied a mixed linear model

method, also utilizing a well-designed permutation method. My association strategy has good control of the false positive rate, and the detected signals have good consistency with results based other datasets. Unexpectedly, the comparisons of detected eQTL signals in different platforms revealed that the detection of cis-eQTL is more vulnerable to platform bias than trans-eQTL. This phenomenon may result from the fact that the design strategy difference: in Illumina platform, the expression abundance is measured by the hybridization signals of 50 bp probes, while in Affymetrix, expression is measured by a set of probes targeting most of the exon region, and in RNA-seq, the expression was measured by the mapped sequenced reads.

CHAPTER 6

CONCLUSION AND DISCUSSION

In this dissertation, I have mainly focused on solving two problems in current human genetics: 1. Whether or not the expression of genes tends to be regulated by multiple eQTL; 2. Why GWAS hits are not more commonly found to co-localize with eQTL. My conclusion is that there are two reasons for the low co-localization level, one is the limited statistical power, and the other is regulation by multiple-causal variants.

To address these two problems, I collected two datasets with large sample sizes to improve the statistical power to detect true, biological eQTL signals, and also developed a statistical method to perform multiple eQTL detection with control for relatedness and population structure. In Chapter 2, to demonstrate the idea of constraints brought by multiple eQTL regulation in current QTL analysis, I performed massive simulations based on real data to demonstrate that currently used methods have low power to detect co-localization signals in the presence of multiple eQTL regulation. In Chapter 3, I described a method named PolyQTL, which I developed to solve the potential problem of relatedness and population structure, which may be common when combining samples from diverse populations. In Chapter 4, I applied my method to real data, explored the eQTL signals overlapping between three expression platforms, and also used it to evaluate GWAS-eQTL co-localization. I used it to find causal genes and causal variants thought to function as

eQTL which affect human complex traits or diseases. In Chapter 5, I described my work as a member of the eQTLgen Consortium.

Power is defined as the likelihood that a study will detect an effect when there is an effect, and can be calculated as $1-\frac{1}{\sqrt{2\pi}}\int_{\Phi^{-1}\left(\frac{\alpha}{2}\right)-\lambda}^{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)-\lambda} e^{-1/2x^2}\,dx$, in which, α is the significance level, λ is the non-centrality parameter, calculated as beta/(sqrt(variance)/n) with beta the allelic effect size, n is the sample size, and the variance is the phenotype variance. From this formula, for a given significance level, we can see that the larger the value of λ, the larger power we have. In order to attain a large λ, we require either large beta, small variance or large sample size n. Because of the limited availability of samples in current eQTL applications (most published studies contain less than 1000 samples), low power hampers the possibility to detect subtle effect eQTL. In addition, the phenomenon of winner's curse further complicates the situation, since non-causal variants with an overestimated effect size are often chosen. Although in this study, my colleagues and I collected an unusually large set of samples to detect cis-eQTL, only half of the genes were found to be regulated by eQTL, and the replication rate is still low for secondary cis-eQTL. It is thus necessary to continue to gather more samples. Currently, the sample size used in my studies is too small to consistently support detection of significant secondary signals in human cis-eQTL studies. Vosa et al (in preparation) show by meta-analysis that with more than 30,000 samples, we can detect cis-eQTL for almost all explored genes in human blood, demonstrating the prevalence of genetic regulation of gene expression. If I could extend the analysis to a larger cohort, I would have more power to detect the secondary signals, and reveal their role in traits and disease.

When gene expression is regulated by multiple eQTL, estimation of the effect sizes of explored variants will be influenced by the combination of multiple causal variants, depending on the LD structure at the locus. My results do reveal that gene expression in human blood tissue are largely regulated by multiple eQTL, and that cis-eQTL are to some extent shared when measurement of transcript abundance is performed on different expression platforms. Co-localization based on multiple-eQTL does find that for as many as 15% of genes, it is secondary cis-eQTL, instead of primary signals, that co-localize with GWAS hits. These secondary signals may be the ones that function as expression regulatory factors to affect human complex traits and diseases, raising the question of why the primary signals are not GWAS hits.

This result is puzzling, and leads me to consider potential biological mechanisms. There are two main possibilities. One is that if there are two causal variants that affect phenotypic traits through expression, due to low power in GWAS detection, the variant with the smaller effect size may finally be chosen and reported, while in eQTL detection, both of the two causal variants are found with conditional statistical analysis. The second possibility is that only certain specific cell types are associated with phenotypic traits or diseases, and the secondary cis-eQTL only functions in these cell types. Since I have measured gene expression abundance in whole blood consisting of a mixture of over a dozen common cell types, the influence of secondary eQTL causal variant is diluted. Although some studies based on cell type specific expression have revealed that only a limited proportion of GWAS hits co-localize with cell eQTL signals, I anticipate that if I could collect more cell-specific samples (for example by single cell RNA-seq of peripheral

blood), and perform cell-specific eQTL, I could resolve more co-localized eQTL-GWAS variants.

In my studies, biological annotation information has not been used. Apart from the statistical association signals between genotype and phenotype, it is also possible to obtain information from biological and functional genomic contexts. For example, as shown in my eQTL-GWAS co-localization analysis, transcription factors are enriched to affect phenotype through gene expression. By integrating these kinds of biological information into the statistical model, I should be able to further enhance the power to find causal variants and genes. Some developed methods (Yang et al., 2017) implicitly make the assumption that variants associated with one trait, should also affect other traits. For instance, COLOC chooses the default prior that a variant affects either expression or phenotype to be 10^{-4}, while the prior that the variant affects both traits is 10^{-5}, which indicates an enrichment of 1000 fold ($10^{-5}/(10^{-4}*10^{-4})$). Although this kind of assumption is now still controversial, some studies reveal that integrating biological annotation into the model may increases power (Yang et al., 2017). Results from the ENCODE and ROADMAP Consortia (2015) have provided a rich atlas of functional information, so extension of my PolyQTL method to jointly integrate functional and association data should improve the accuracy of fine mapping.

REFERENCES

Akey, J. M., S. Biswas, J. T. Leek and J. D. Storey, 2007 On the design and analysis of gene expression studies in human populations. ***Nat Genet*** 39: 807-808.

Andiappan, A. K., R. Melchiotti, T. Y. Poh, M. Nah, K. J. Puan et al., 2015 Genome-wide analysis of the genetic regulation of gene expression in human neutrophils. ***Nat Commn*** 6:7971.

Ardlie, K. G., D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young et al., 2015 The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. ***Science*** 348: 648-660.

Battle, A., S. Mostafavi, X. Zhu, J. B. Potash, M. M. Weissman et al., 2014 Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. ***Genome Res*** 24: 14-24.

Benner, C., C. C. Spencer, A. S. Havulinna, V. Salomaa, S. Ripatti et al., 2016 FINEMAP: efficient variable selection using summary data from genome-wide association studies. ***Bioinformatics*** 32: 1493-1501.

Brem, R. B., G. Yvert, R. Clinton and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. ***Science*** 296: 752-755.

Bryois, J., A. Buil, D. M. Evans, J. P. Kemp, S. B. Montgomery et al., 2014 Cis and trans effects of human genomic variants on gene expression. ***PLoS Genet*** 10: e1004461.

Burkhardt, R., H. Kirsten, F. Beutner, L. M. Holdt, A. Gross et al., 2015 Integration of genome-wide SNP data and gene-expression profiles reveals six novel loci and regulatory mechanisms for amino acids and acylcarnitines in whole blood. ***PLoS Genet*** 11: e1005510.

Chen, W., B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva et al., 2015 Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. ***Genetics*** 200: 719-736.

Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K.-Y. Jen et al., 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. ***Nat Genet*** 33: 422-425.

Cheung, V. G., R. S. Spielman, K. G. Ewens, T. M. Weber, M. Morley et al., 2005 Mapping determinants of human gene expression by regional and genome-wide association. ***Nature*** 437: 1365-1369.

International Human Genome Sequencing Consortium, 2005 Initial sequence of the chimpanzee genome and comparison with the human genome. ***Nature*** 437: 69-87.

Choy, E., R. Yelensky, S. Bonakdar, R. M. Plenge, R. Saxena et al., 2008 Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. ***PLoS Genet*** 4: e1000287.

Chun, S., A. Casparino, N. A. Patsopoulos, D. C. Croteau-Chonka, B. A. Raby et al., 2017 Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. ***Nat Genet*** 49: 600-605.

Chung, D., C. Yang, C. Li, J. Gelernter and H. Zhao, 2014 GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. ***PLoS Genet*** 10: e1004787.

Civelek, M., Y. Wu, C. Pan, C. K. Raulerson, A. Ko et al., 2017 Genetic regulation of adipose gene expression and cardio-metabolic traits. ***Am J Hum Genet*** 100: 428-443.

Claussnitzer, M., S. N. Dankel, K.-H. Kim, G. Quon, W. Meuleman et al., 2015 FTO obesity variant circuitry and adipocyte browning in humans. ***N Engl J Med*** 2015: 895-907.

Colantuoni, C., B. K. Lipska, T. Ye, T. M. Hyde, R. Tao et al., 2011 Temporal dynamics and genetic control of transcription in the human prefrontal cortex. ***Nature*** 478: 519-523.

Corradin, O., A. Saiakhova, B. Akhtar-Zaidi, L. Myeroff, J. Willis et al., 2014 Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. ***Genome Res*** 24: 1-13.

Coronary Artery Disease Genetics Consortium, 2011 A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. ***Nat Genet*** 43: 339-344.

Crawford, N. P., X. Qian, A. Ziogas, A. G. Papageorge, B. J. Boersma et al., 2007 Rrp1b, a new candidate susceptibility gene for breast cancer progression and metastasis. ***PLoS Genet*** 3: e214.

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel et al., 2009a Common regulatory variation impacts gene expression in a cell type–dependent manner. ***Science*** 325: 1246-1250.

Dimas, A. S., S. Deutsch, B. E. Stranger, S. B. Montgomery, C. Borel et al., 2009b Common regulatory variation impacts gene expression in a cell type-dependent manner. ***Science*** 325: 1246-1250.

Ding, J., J. E. Gudjonsson, L. Liang, P. E. Stuart, Y. Li et al., 2010 Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. ***Am J Hum Genet*** 87: 779-789.

Dixon, A. L., L. Liang, M. F. Moffatt, W. Chen, S. Heath et al., 2007 A genome-wide association study of global gene expression. ***Nat Genet*** 39: 1202-1207.

Duan, S., R. S. Huang, W. Zhang, W. K. Bleibel, C. A. Roe et al., 2008 Genetic architecture of transcript-level variation in humans. ***Am J Hum Genet*** 82: 1101-1113.

Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink et al., 2008 Genetics of gene expression and its effect on disease. ***Nature*** 452: 423-428.

Encode Project Consortium, 2012 An integrated encyclopedia of DNA elements in the human genome. ***Nature*** 489: 57-74.

Fadista, J., P. Vikman, E. O. Laakso, I. G. Mollet, J. L. Esguerra et al., 2014 Global genomic and transcriptomic analysis of human pancreatic islets reveals novel genes influencing glucose metabolism. ***Proc Natl Acad Sci USA*** 111: 13924-13929.

Fairfax, B. P., P. Humburg, S. Makino, V. Naranbhai, D. Wong et al., 2014 Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. ***Science*** 343: 1246949.

Fairfax, B. P., S. Makino, J. Radhakrishnan, K. Plant, S. Leslie et al., 2012 Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. ***Nat Genet*** 44: 502-510.

Farh, K. K.-H., A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley et al., 2015 Genetic and epigenetic fine mapping of causal autoimmune disease variants. ***Nature*** 518: 337-343.

Fehrmann, R. S., R. C. Jansen, J. H. Veldink, H.-J. Westra, D. Arends et al., 2011 Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. ***PLoS Genet*** 7: e1002197.

Gaulton, K. J., T. Ferreira, Y. Lee, A. Raimondo, R. Mägi et al., 2015 Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. ***Nat Genet*** 47: 1415-1425.

Gibbs, J. R., M. P. van der Brug, D. G. Hernandez, B. J. Traynor, M. A. Nalls et al., 2010 Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. ***PLoS Genet*** 6: e1000952.

Göring, H. H., J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth et al., 2007 Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. ***Nat Genet*** 39: 1208-1216.

Grundberg, E., V. Adoue, T. Kwan, B. Ge, Q. L. Duan et al., 2011 Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. ***PLoS Genet*** 7: e1001279.

Grundberg, E., K. S. Small, Å. K. Hedman, A. C. Nica, A. Buil et al., 2012 Mapping cis- and trans-regulatory effects across multiple tissues in twins. ***Nat Genet*** 44: 1084-1089.

GTEx Consortium, 2015 The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. ***Science*** 348: 648-660.

Gusev, A., G. Bhatia, N. Zaitlen, B. J. Vilhjalmsson, D. Diogo et al., 2013 Quantifying missing heritability at known GWAS loci. ***PLoS Genet*** 9: e1003993.

Gusev, A., A. Ko, H. Shi, G. Bhatia, W. Chung et al., 2016 Integrative approaches for large-scale transcriptome-wide association studies. ***Nat Genet*** 48: 245–252

Gusev, A., S. H. Lee, G. Trynka, H. Finucane, B. J. Vilhjalmsson et al., 2014 Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. ***Am J Hum Genet*** 95: 535-552.

Hao, K., Y. Bossé, D. C. Nickle, P. D. Paré, D. S. Postma et al., 2012 Lung eQTLs to help reveal the molecular underpinnings of asthma. ***PLoS Genet*** 8: e1003029.

Hill, W., and A. Robertson, 1968 Linkage disequilibrium in finite populations. ***Theor Appl Genet*** 38: 226-231.

Hofman, A., S. D. Murad, C. M. van Duijn, O. H. Franco, A. Goedegebure et al., 2013 The Rotterdam Study: 2014 objectives and design update. ***Eur J Epidemiol*** 28: 889-926.

Hore, V., A. Viñuela, A. Buil, J. Knight, M. I. McCarthy et al., 2016 Tensor decomposition for multiple-tissue gene expression experiments. ***Nat Genet*** 48: 1094-1100.

Hormozdiari, F., E. Kostem, E. Y. Kang, B. Pasaniuc and E. Eskin, 2014 Identifying causal variants at loci with multiple signals of association. ***Genetics*** 198: 497-508.

Hormozdiari, F., M. van de Bunt, A. V. Segre, X. Li, J. W. J. Joo et al., 2016 Colocalization of GWAS and eQTL Signals Detects Target Genes. ***Am J Hum Genet*** 99: 1245-1260.

Horvath, S., B. Zhang, M. Carlson, K. Lu, S. Zhu et al., 2006 Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. ***Proc Natl Acad Sci USA*** 103: 17402-17407.

Huan, T., C. Liu, R. Joehanes, X. Zhang, B. H. Chen et al., 2015 A systematic heritability analysis of the human whole blood transcriptome. ***Hum Genet*** 134: 343-358.

Hussin, J. G., A. Hodgkinson, Y. Idaghdour, J.-C. Grenier, J.-P. Goulet et al., 2015 Recombination affects accumulation of damaging and disease-associated mutations in human populations. ***Nat Genet*** 47: 400-404.

Idaghdour, Y., W. Czika, K. V. Shianna, S. H. Lee, P. M. Visscher et al., 2010a Geographical genomics of human leukocyte gene expression variation in southern Morocco. ***Nat Genet*** 42: 62-67.

Idaghdour, Y., W. Czika, K. V. Shianna, S. H. Lee, P. M. Visscher et al., 2010b Geographical genomics of human leukocyte gene expression variation in southern Morocco. ***Nat Genet*** 42: 62-67.

Innocenti, F., G. M. Cooper, I. B. Stanaway, E. R. Gamazon, J. D. Smith et al., 2011 Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. ***PLoS Genet*** 7: e1002078.

Inouye, M., K. Silander, E. Hamalainen, V. Salomaa, K. Harald et al., 2010 An immune response network associated with blood lipid levels. ***PLoS Genet*** 6: e1001113.

Jacob, F., and J. Monod, 1961 Genetic regulatory mechanisms in the synthesis of proteins. ***J Mol Biol*** 3: 318-356.

Kasela, S., K. Kisand, L. Tserel, E. Kaleviste, A. Remm et al., 2017 Pathogenic implications for autoimmune mechanisms derived by comparative eQTL analysis of CD4+ versus CD8+ T cells. ***PLoS Genet*** 13: e1006643.

Kathiresan, S., O. Melander, C. Guiducci, A. Surti, N.P. Burtt, et al., 2008 Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. ***Nat Genet*** 40: 189-197.

Kichaev, G., and B. Pasaniuc, 2015 Leveraging functional-annotation data in trans-ethnic fine-mapping studies. ***Am J Hum Genet*** 97: 260-271.

Kichaev, G., W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin et al., 2014 Integrating functional data to prioritize causal variants in statistical fine-mapping studies. ***PLoS Genet*** 10: e1004722.

Kim, J., N. Ghasemzadeh, D. J. Eapen, N. C. Chung, J. D. Storey et al., 2014 Gene expression profiles associated with acute myocardial infarction and risk of cardiovascular death. ***Genome Med*** 6: 40.

Klein, R. J., C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler et al., 2005 Complement factor H polymorphism in age-related macular degeneration. ***Science*** 308: 385-389.

Koopmann, T. T., M. E. Adriaens, P. D. Moerland, R. F. Marsman, M. L. Westerveld et al., 2014 Genome-wide identification of expression quantitative trait loci (eQTLs) in human heart. ***PLoS One*** 9: e97380.

Lappalainen, T., M. Sammeth, M. R. Friedländer, P. AC't Hoen, J. Monlong et al., 2013 Transcriptome and genome sequencing uncovers functional variation in humans. ***Nature*** 501: 506-511.

Lee, M. N., C. Ye, A.-C. Villani, T. Raj, W. Li et al., 2014 Common genetic variants modulate pathogen-sensing responses in human dendritic cells. ***Science*** 343: 1246980.

Li, Q., J.-H. Seo, B. Stranger, A. McKenna, I. Pe'er et al., 2013 Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. ***Cell*** 152: 633-641.

Li, Y., O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu et al., 2006 Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. ***PLoS Genet*** 2: e222.

Liang, L., N. Morar, A. L. Dixon, G. M. Lathrop, G. R. Abecasis et al., 2013 A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. ***Genome Res*** 23: 716-726.

Lindholm, M. E., M. Huss, B. W. Solnestam, S. Kjellqvist, J. Lundeberg et al., 2014 The human skeletal muscle transcriptome: sex differences, alternative splicing, and tissue homogeneity assessed with RNA sequencing. ***The FASEB J*** 28: 4571-4581.

Liu, J. Z., S. van Sommeren, H. Huang, S. C. Ng, R. Alberts et al., 2015 Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. ***Nat Genet*** 47: 979-986.

Lloyd-Jones, L. R., A. Holloway, A. McRae, J. Yang, K. Small et al., 2017 The Genetic Architecture of Gene Expression in Peripheral Blood. ***Am J Hum Genet*** 100: 228-237.

Loh, P. R., G. Kichaev, S. Gazal, A. P. Schoech and A. L. Price, 2018 Mixed-model association for biobank-scale datasets. ***Nat Genet*** 50: 906-908.

Mallick, S., H. Li, M. Lipson, I. Mathieson, M. Gymrek et al., 2016 The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. ***Nature*** 538: 201-206.

Mancuso, N., H. Shi, P. Goddard, G. Kichaev, A. Gusev et al., 2017 Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. ***Am J Hum Genet*** 100: 473-487.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff et al., 2009 Finding the missing heritability of complex diseases. ***Nature*** 461: 747-753.

Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen et al., 2012 Systematic localization of common disease-associated variation in regulatory DNA. ***Science*** 337: 1190-1195.

Mehta, D., K. Heim, C. Herder, M. Carstensen, G. Eckstein et al., 2013 Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. ***Euro J Hum Genet*** 21: 48-54.

Mellars, P., 2006 Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. ***Proc Natl Acad Sci USA*** 103: 9381-9386.

Metzger, B. P., D. C. Yuan, J. D. Gruber, F. Duveau and P. J. Wittkopp, 2015 Selection on noise constrains variation in a eukaryotic promoter. ***Nature*** 521: 344-347.

Mi, H., X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P.D. Thomas, 2017 PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. ***Nucleic Acids Res*** 45: D183-D189.

Moffatt, M. F., M. Kabesch, L. Liang, A. L. Dixon, D. Strachan et al., 2007 Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. ***Nature*** 448: 470-473.

Monks, S., A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak et al., 2004 Genetic inheritance of gene expression in human cell lines. ***Am J Hum Genet*** 75: 1094-1105.

Montgomery, S. B., M. Sammeth, M. Gutierrez-Arcelus, R. P. Lach, C. Ingle et al., 2010 Transcriptome genetics using second generation sequencing in a Caucasian population. ***Nature*** 464: 773-777.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens et al., 2004 Genetic analysis of genome-wide variation in human gene expression. ***Nature*** 430: 743-747.

Musunuru, K., A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt et al., 2010 From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. ***Nature*** 466: 714-719.

Myers, A. J., J. R. Gibbs, J. A. Webster, K. Rohrer, A. Zhao et al., 2007 A survey of genetic human cortical gene expression. ***Nat Genet*** 39: 1494-1499.

Narahara, M., K. Higasa, S. Nakamura, Y. Tabara, T. Kawaguchi et al., 2014 Large-scale East-Asian eQTL mapping reveals novel candidate genes for LD mapping and the genomic landscape of transcriptional effects of sequence variants. ***PloS One*** 9: e100924.

Nédélec, Y., J. Sanz, G. Baharian, Z. A. Szpiech, A. Pacis et al., 2016 Genetic ancestry and natural selection drive population differences in immune responses to pathogens. ***Cell*** 167: 657-669.

Nica, A. C., S. B. Montgomery, A. S. Dimas, B. E. Stranger, C. Beazley et al., 2010 Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. ***PLoS Genet*** 6: e1000895.

Nicolae, D. L., E. Gamazon, W. Zhang, S. Duan, M. E. Dolan et al., 2010 Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. ***PLoS Genet*** 6: e1000888.

Nikpay, M., A. Goel, H.H. Won, L.M. Hall, C. Willenborg, et al., 2015 A comprehensive 1,000 Genomes-based genome-wide association meta-analysis of coronary artery disease. ***Nat Genet*** 47: 1121-1130.

Ongen, H., C. L. Andersen, J. B. Bramsen, B. Oster, M. H. Rasmussen et al., 2014 Putative cis-regulatory drivers in colorectal cancer. ***Nature*** 512: 87-90.

Peters, J. E., P. A. Lyons, J. C. Lee, A. C. Richard, M. D. Fortune et al., 2016 Insight into genotype-phenotype associations through eQTL mapping in multiple cell types in health and immune-mediated disease. ***PLoS Genet*** 12: e1005908.

Peterson, C. B., A. J. Jasinska, F. Gao, I. Zelaya, T. M. Teshiba et al., 2016 Characterization of Expression Quantitative Trait Loci in Pedigrees from Colombia and Costa Rica Ascertained for Bipolar Disorder. ***PLoS Genet*** 12: e1006046.

Pickrell, J. K., 2014 Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. ***Am J Hum Genet*** 94: 559-573.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt et al., 2010 Understanding mechanisms underlying human gene expression variation with RNA sequencing. ***Nature*** 464: 768-772.

Pierce, B. L., L. Tong, L. S. Chen, R. Rahaman, M. Argos et al., 2014 Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. ***PLoS Genet*** 10: e1004818.

Polderman, T. J., B. Benyamin, C. A. de Leeuw, P. F. Sullivan, A. van Bochoven et al., 2015 Meta-analysis of the heritability of human traits based on fifty years of twin studies. ***Nat Genet*** 47: 702-709.

Powell, J. E., A. K. Henders, A. F. McRae, A. Caracella, S. Smith et al., 2012 The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. ***PLoS One*** 7: e35430.

Preininger, M., D. Arafat, J. Kim, A. P. Nath, Y. Idaghdour et al., 2013 Blood-informative transcripts define nine common axes of peripheral blood gene expression. ***PLoS Genet*** 9: e1003362.

Price, A. L., N. Patterson, D. C. Hancks, S. Myers, D. Reich et al., 2008 Effects of cis and trans genetic ancestry on gene expression in African Americans. ***PLoS Genet*** 4: e1000294.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. ***Am J Hum Genet*** 69: 1-14.

Quach, H., M. Rotival, J. Pothlichet, Y.-H. E. Loh, M. Dannemann et al., 2016 Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. ***Cell*** 167: 643-656.

Raj, T., K. Rothamel, S. Mostafavi, C. Ye, M. N. Lee et al., 2014a Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. ***Science*** 344: 519-523.

Raj, T., K. Rothamel, S. Mostafavi, C. Ye, M. N. Lee et al., 2014b Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. ***Science*** 344: 519-523.

Ramasamy, A., D. Trabzuni, S. Guelfi, V. Varghese, C. Smith et al., 2014 Genetic variability in the regulation of gene expression in ten regions of the human brain. ***Nat Neurosci*** 17: 1418-1428.

Roadmap Epigenomics Consortium, 2015 Integrative Analysis of 111 Human Reference Epigenomes. ***Nature*** 518: 317-330

Schadt, E. E., C. Molony, E. Chudin, K. Hao, X. Yang et al., 2008 Mapping the genetic architecture of gene expression in human liver. ***PLoS Biol*** 6: e107.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che et al., 2003 Genetics of gene expression surveyed in maize, mouse and man. ***Nature*** 422: 297-302.

Scott, L. J., M. R. Erdos, J. R. Huyghe, R. P. Welch, A. T. Beck et al., 2016 The genetic regulatory signature of type 2 diabetes in human skeletal muscle. ***Nat Commun*** 7:11764.

Simon, L. M., L. C. Edelstein, S. Nagalla, A. B. Woodley, E. S. Chen et al., 2014 Human platelet microRNA-mRNA networks associated with age and gender revealed by integrated plateletomics. ***Blood*** 123: e37-e45.

Soldner, F., Y. Stelzer, C. S. Shivalila, B. J. Abraham, J. C. Latourelle et al., 2016 Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. ***Nature*** 533: 95-99.

Sotoodehnia, N., A. Isaacs, P. I. De Bakker, M. Dörr, C. Newton-Cheh et al., 2010 Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. ***Nat Genet*** 42: 1068-1076.

Spielman, R. S., L. A. Bastone, J. T. Burdick, M. Morley, W. J. Ewens et al., 2007 Common genetic variants account for differences in gene expression among ethnic groups. ***Nat Genet*** 39: 226-231.

Stranger, B. E., M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch et al., 2005 Genome-wide associations of gene expression variation in humans. ***PLoS Genet*** 1: e78.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley et al., 2007a Relative impact of nucleotide and copy number variation on gene expression phenotypes. ***Science*** 315: 848-853.

Stranger, B. E., A. C. Nica, M. S. Forrest, A. Dimas, C. P. Bird et al., 2007b Population genomics of human gene expression. ***Nat Genet*** 39: 1217.

Tanaka, T., J. Shen, G. R. Abecasis, A. Kisialiou, J. M. Ordovas et al., 2009 Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. ***PLoS Genet*** 5: e1000338.

Teslovich, T. M., K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou et al., 2010 Biological, clinical and population relevance of 95 loci for blood lipids. ***Nature*** 466: 707-713.

Thibodeau, S. N., A. French, S. McDonnell, J. Cheville, S. Middha et al., 2015 Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set. ***Nat Commun*** 6:8653.

Tigchelaar, E. F., A. Zhernakova, J. A. Dekens, G. Hermes, A. Baranska et al., 2015 Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. ***Brit Med J*** 5: e006772.

Tönjes, A., E. Zeggini, P. Kovacs, Y. Böttcher, D. Schleinitz et al., 2010 Association of FTO variants with BMI and fat mass in the self-contained population of Sorbs in Germany. ***Euro J Hum Genet*** 18: 104-110.

Torgerson, D. G., A. R. Boyko, R. D. Hernandez, A. Indap, X. Hu et al., 2009 Evolutionary processes acting on candidate cis-regulatory regions in humans inferred from patterns of polymorphism and divergence. ***PLoS Genet*** 5: e1000592.

Trynka, G., C. Sandor, B. Han, H. Xu, B. E. Stranger et al., 2013 Chromatin marks identify critical cell types for fine mapping complex trait variants. ***Nat Genet*** 45: 124-130.

Turpeinen, H., I. Seppälä, L.-P. Lyytikäinen, E. Raitoharju, N. Hutri-Kähönen et al., 2015 A genome-wide expression quantitative trait loci analysis of proprotein convertase subtilisin/kexin enzymes identifies a novel regulatory gene variant for FURIN expression and blood pressure. ***Hum Genet*** 134: 627-636.

Udler, M. S., J. Tyrer and D. F. Easton, 2010 Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. ***Genet Epidemiol*** 34: 463-468.

Marigorta, U. M., L. A. Denson, J. S. Hyams, K. Mondal, J. Prince *et al.*, 2017 Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. ***Nat Genet*** 49**:** 1517-1521.

van de Bunt, M., J. E. M. Fox, X. Dai, A. Barrett, C. Grey et al., 2015 Transcript expression data from human islets links regulatory signals from genome-wide association studies for type 2 diabetes and glycemic traits to their downstream effectors. ***PLoS Genet*** 11: e1005694.

Walsh, A. M., J. W. Whitaker, C. C. Huang, Y. Cherkas, S. L. Lamberth et al., 2016 Integrative genomic deconvolution of rheumatoid arthritis GWAS loci into gene and cell type associations. ***Genome Biol*** 17: 79.

Webster, J. A., J. R. Gibbs, J. Clarke, M. Ray, W. Zhang et al., 2009 Genetic control of human brain transcript expression in Alzheimer disease. ***Am J Hum Genet*** 84: 445-458.

Wen, X., Y. Lee, F. Luca and R. Pique-Regi, 2016 Efficient integrative multi-SNP association analysis via Deterministic Approximation of Posteriors. ***Am J Hum Genet*** 98: 1114-1129.

Wen, X., F. Luca and R. Pique-Regi, 2015 Cross-population joint analysis of eQTLs: fine mapping and functional annotation. ***PLoS Genet*** 11: e1005176.

Westra, H.-J., M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann et al., 2013 Systematic identification of trans eQTLs as putative drivers of known disease associations. ***Nat Genet*** 45: 1238-1243.

Westra, H. J., D. Arends, T. Esko, M. J. Peters, C. Schurmann et al., 2015 Cell Specific eQTL Analysis without Sorting Cells. ***PLoS Genet*** 11: e1005223.

Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers et al., 2014 Defining the role of common variation in the genomic and biological architecture of adult human height. ***Nat Genet*** 46: 1173-1186.

Wright, F. A., P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun et al., 2014 Heritability and genomics of gene expression in peripheral blood. ***Nat Genet*** 46: 430-437.

Yan, H., W. Yuan, V. E. Velculescu, B. Vogelstein and K. W. Kinzler, 2002 Allelic variation in human gene expression. ***Science*** 297: 1143-1143.

Yang, E., G. Wang, J. Yang, B. Zhou, Y. Tian et al., 2016 Epistasis and destabilizing mutations shape gene expression variability in humans via distinct modes of action. ***Hum Mol Genet***: 25(22):4911-4919.

Yang, J., T. Ferreira, A. P. Morris, S. E. Medland, P. A. Madden et al., 2012 Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. ***Nat Genet*** 44: 369-375.

Yang, J., L. G. Fritsche, X. Zhou, G. Abecasis and C. International Age-Related Macular Degeneration Genomics, 2017 A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. ***Am J Hum Genet*** 101: 404-416.

Ye, C. J., T. Feng, H.-K. Kwon, T. Raj, M. T. Wilson et al., 2014 Intersection of population variation and autoimmunity genetics in human T cell activation. ***Science*** 345: 1254665.

Zeller, T., P. Wild, S. Szymczak, M. Rotival, A. Schillert et al., 2010 Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. ***PloS One*** 5: e10693.

Zeng, B. and G. Gibson, 2018. PolyQTL: Bayesian multiple eQTL detection with control for population structure and sample relatedness. ***Bioinformatics*** In press PMID: 30165584

Zeng, B., L.R. Lloyd-Jones, A. Holloway, U.M. Marigorta, A. Metspalu, et al., 2017 Constraints on eQTL fine mapping in the presence of multisite local regulation of gene expression. ***G3*** 7: 2533-2544.

Zou, F., H. S. Chai, C. S. Younkin, M. Allen, J. Crook et al., 2012 Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. ***PLoS Genet*** 8: e1002707.