ABSTRACT

Title of Document: ANALYSIS OF CONSENSUS GENOME-WIDE EXPRESSION-QTLS AND THEIR RELATIONSHIPS TO HUMAN COMPLEX TRAIT DISEASES.

Chen-Hsin Yu, Doctor of Philosophy, 2014

Directed By: Professor John Moult, Institute for Bioscience and Biotechnology Research. Department of Cell Biology and Molecular Genetics, University of Maryland

Genome-wide association studies of human complex disease have identified a large number of disease associated genetic loci. However, most of these risk loci do not provide direct information on the biological basis of a disease or on the underlying mechanisms. Recent genome-wide expression quantitative trait loci (eQTLs) association studies have provided information on genetic factors, especially SNPs, associated with gene expression variation. These eQTLs might contribute to phenotype diversity and disease susceptibility, but interpretation is handicapped by low reproducibility of the expression results. Our first major goal was to establish a list of consensus eQTLs by integrating publicly available data for specific human

populations and cell types. We used linkage disequilibrium data from Hapmap and the 1000 Genomes Project to integrate the results of eQTL studies. Overall, we find over 4000 genes that are involved in high confidence eQTL relationships. We also assessed the possible underlying mechanisms of tissue dependent eQTLs by mapping these to known genome sites of functional elements. Results of comparison of eQTLs across studies on the same cell type versus those on different cell types suggest that tissue specific eQTLs are less common than pan-tissue eQTLs. Our second major goal was to use these results to elucidate the role eQTLs play in human common diseases. For this purpose, we matched the high confidence eQTLs to a set of 335 disease risk loci identified from the Wellcome Trust Case Control Consortium (WTCCC1) genome-wide association study and follow-up studies for seven human common diseases. Our results show that the data are consistent with approximately 50% of these disease loci arising from an underlying expression change mechanism. In many cases, the results provide a proposed expression mechanism for genes previously suggested as disease relevant, in others, new disease relevant genes are identified. A web-based database, ExSNP, was designed to provide comprehensive access to the eQTL data and results from our analysis, including original eQTLs, high-confidence eQTLs, cell type dependent eQTLs, population dependent eQTLs, disease associated eQTLs, and functionally annotated eQTLs. The website also incorporates a genome browser that allows visualization of the relative positions of eQTL SNPs to their associated genes and other neighboring genes, as well as the relationship to functional elements and disease associations.

ANALYSIS OF CONSENSUS GENOME-WIDE EXPRESSION-QTLS AND THEIR RELATIONSHIPS TO HUMAN COMPLEX TRAIT DISEASES

By

Chen-Hsin Yu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Professor John Moult, Chair
Professor Stephen M. Roth
Associate Professor Sridhar Hannenhalli
Associate Professor Stephen M. Mount
Associate Professor Mihai Pop

Dedication

To my family

Acknowledgements

Pursuing a Ph.D. has been the most wonderful and unforgettable journey in my life. From the first day I entered IBBR, which we used to call “CARB”, I always felt excited about exploring new technology and novel scientific ideas. I would like to thank all the members of staff at IBBR for creating an incredible research environment. I could never have reached the heights or explored the depths without the help, support, guidance and efforts of a lot of people.

First and foremost, I cannot express enough thanks to my advisor, Dr. John Moult, whose infectious enthusiasm and intellectual guidance have instilling in me the qualities required in being a good scientist. I always enjoy our scientific discussions, and these have been a constant source of inspiration, excitement, advice, and guidance throughout my studies. I also deeply thank him for the unprecedented freedom he provided me to explore my curiosity in solving some daunting problems in biology.

I am sincerely grateful for the help and suggestions from all my committee members, Dr. Jonathan Dinman, Dr. Sridhar Hannenhalli, Dr. Stephen Mount, Dr. Mihai Pop, and Dr. Stephen Roth, throughout this project and through my entire program of study. I especially appreciate Dr. Mount’s generous support and discussions on my research. I also want to thank Dr. Pop for his knowledgeable teaching in bioinformatics.

My completion of this project could not have been accomplished without the support of all my lab members. I would like to thank Dr. Lipika Ray Pal. She always offered the most generous help and continuous encouragement when I was in trouble. I would

also like to thank Maya Zuhl for her knowledge of computational skills and web maintenance. Many thanks to Dr. Nuttinee Teerakulkittipong, Dr. Zhen Shi, Mr. Yizhou Yin, Dr. Chen Cao, for sharing the experience of research and for the mutual support in the lab.

Finally, I would like to thank my family for their constant love and immeasurable sacrifice during my life. It is to them I dedicate this thesis. First, I would like to thank my parents. Without their priceless support and countless encouragements, I would never come this far. Lastly, and most importantly, I would like to thank my wife, Ching-Ching Lin, for accompanying me on this journey and being proud of me every step of the way.

Table of Contents

Dedication ii

Acknowledgements iii

Table of Contents v

List of Tables viii

List of Figures ix

List of Abbreviations xi

Chapter 1: Introduction 1

1.1: Genome-wide expression quantitative trait loci (eQTL) 1

1.2: Mapping complex disease traits with global gene expression 4

1.3: Integration of eQTL studies 8

Chapter 2: Meta-analysis of Genome-wide Expression Quantitative Trait Loci (eQTLs) for Various Human Populations and Cell Types 12

2.1: Introduction 12

2.2: Results 14

Summary of Genome-wide eQTL. 14

Linkage Disequilibrium relationships between eQTLs 22

Pair-wise comparisons show low agreement between eQTL datasets 26

High-confidence eQTL identification 32

Tissue dependence of eQTL relationships 40

2.3: Methods 49

Data sources 49

Data preparation 51

Linkage disequilibrium 51
Functional elements 53
2.4: Discussion 54
Chapter 3: The role of Human expression quantitative Traits in complex trait disease 57
3.1: Introduction 57
3.2: Results 62
High-confident eQTLs 62
Identification of Disease loci with a possible expression related mechanism 63
Examples of disease associated eQTL relationships 68
ADAM15 for Crohn's disease 68
TSPAN3 and PSTPIP1 for Type 2 Diabetes 69
GALNT4 for Hypertension 70
3.3: Methods 73
High-confidence eQTL Data 73
Genome-wide associations studies of human common diseases 73
LD Relationships 73
CentiMorgan distance calculation 74
Comparison of disease and eQTL markers 74
3.4: Discussion: 82
Chapter 4: Web-based database for query and visualization of human genome-wide expression quantitative trait loci 88
4.1: Introduction 88

4.2: Construction and content 91
Database construction 91
eQTL data sources and processing 92
Identification of consensus eQTLs 93
Population-dependent and cell type-dependent eQTLs 94
Functional interpretation of eQTLs 95
Disease associated eQTLs 95
Web implementation 96
4.3: Use of the resource 97
eQTL related Queries 97
eQTL browsing 97
An example of a tissue dependent high-confidence eQTL relationship in liver 100
An example of a high confidence eQTL relationship associated with human disease 101
4.4: Conclusion 103
Chapter 5: Conclusions and perspectives 104
5.1: High Confidence eQTL sets 104
5.2: Disease associated eQTLs 104
5.3: Web-based resource 105
5.4: Future perspectives 106
Appendix 108
Bibliography 132

List of Tables

Table 1.1. Summary of candidate causal genes for human common diseases that were discovered using eQTL data 7

Table 2.1. eQTL data for 16 selected genome-wide eQTL association studies..... 18

Table 2.2. eQTL associations and unique eQTL relationships for each dataset..... 24

Table 2.3. Pair-wise comparisons of eQTL datasets 30

Table 2.4. Classification of eQTL studies by cell type and population..... 33

Table 2.5. Summary of eQTL relationships and high-confidence eQTL relationships found in each integrated set 37

Table 2.6. Number of exSNPs that fall on each type of functional element, and number of associated exGenes, for each integrated eQTL set 46

Table 3.1 Summary of high-confident eQTLs from 16 integrated sets..... 62

Table 3.2. Number of disease risk loci with possible underlying expression mechanisms in seven common diseases. 65

Table 3.3. Number of genes in each category for each disease. 66

Table 3.4**.** The 2x2 table of numbers of genes in that are disease candidates and/or involved in in high-confidence eQTL relationships. 68

List of Figures

Figure 2.1. Fraction of cis- and trans- eQTL associations in each dataset. 20

Figure 2.2. Distribution of distances between each cis-exSNP and the associated exGene. 21

Figure 2.3. Hierarchical clustering of the fraction of common exGenes between pairs of eQTL datasets. 31

Figure 2.4. Hierarchical clustering of the fraction of common exGenes with LD related exSNPs between pairs of eQTL datasets. 32

Figure 2.5. Identification of high confidence unique eQTL relationships. 36

Figure 2.6. Number of high-confidence exGenes with 1, 2, 3... eQTL relationships at various LD thresholds (r^2) in the AllCell_AllPop integrated set. 38

Figure 2.7. Number of HC-exGenes with support from 1, 2, 3, ... studies at various LD threshold (r^2) in the AllCell_AllPop integrated set. 39

Figure 2.8. Approximate quality of each dataset, as reflected in the % of high-confidence exGenes relative to the LCL_CEU integrated set, at an $r^2 > 0.3$ LD threshold. 40

Figure 2.9. Comparisons of fractions of common exGenes between pairs of eQTL datasets of the same cell type and pairs with different cell types for the MuTHER study. 42

Figure 2.10. Comparisons of fractions of common exGenes between pairs of eQTL datasets of the same cell type and pairs with different cell types for the 3C study. 43

Figure 2.11. Comparisons of fractions of common exGenes between datasets in the same population versus datasets from different populations. 44
Figure 2.12. Distribution of tissue-specific exGenes associated transcription factor binding sites. 48
Figure 2.13. Histogram of differences in LD values for pairs of SNPs derived from Hapmap and 1000 Genomes data. 52
Figure 3. 1. Percentage of disease loci with possible expression mechanisms as a function of the cM distance between the closest disease and expression marker SNPs. 72
Figure 3.2. Model for identifying those disease associated loci with a probable underlying expression mechanism. 75
Figure 3.3. Manhattan plots for a locus associated with Type 1 diabetes in the WTCCC1 data. 79
Figure 4.1. Workflow for the construction of the ExSNP database. 91
Figure 4.2. Sample screenshot from the ExSNP browser. 99
Figure 4.3. Visualization of a Liver dependent high-confidence eQTL relationships for APOC4. 101
Figure 4.4. Visualization of eQTL relationships for the chromosome region 17q12 that is associated with the risk of Asthma and some autoimmune diseases. 103

List of Abbreviations

eQTL	expression quantitative trait locus
EBV	Epstein–Barr virus
GWAS	genome-wide association study
LCL	lymphoblastoid cell line
LD	linkage disequilibrium
SNP	single nucleotide polymorphism
WTCCC	the Wellcome Trust Case Control Consortium

Chapter 1: Introduction

Genome-wide association studies (GWAS) have been successful in identifying genetic variants associated with numerous human traits and common diseases (Welter et al., 2014). There is, however, a substantial gap between these findings and a full understanding of how the loci contribute to complex trait diseases. Despite the large number of single nucleotide polymorphisms (SNPs) discovered to be reproducibly associated with traits, only rarely does a SNP affect protein function in an obvious manner (Manolio et al., 2009). A large number of SNPs from GWAS lie in intergenic or intronic regions, where the downstream mechanism by which the phenotype is influenced is unclear. In this thesis, we relate complex trait diseases to global gene expression, with the goal of identifying those disease loci where the underlying mechanism may involve change in expression of a gene. In the first part of the work, we systematically investigated the association of human genetic variants with gene expression. In the second part, we linked these expression relationships to GWAS results for disease traits, in this way providing putative expression mechanisms for a subset of disease related loci.

1.1: Genome-wide expression quantitative trait loci (eQTL)

Gene expression variation (i.e. transcript abundance) among individuals can be considered as a highly heritable quantitative trait in human populations (Cheung et al., 2005; Morley et al., 2004; Storey et al., 2007). Transcript levels of genes may be modified by genetic variants through various mechanisms in transcription and post-

transcriptional processes. For instance, SNPs located on cis-regulatory regions, such as transcription factor binding sites, microRNA binding sites, splicing sites, or sites that regulate the decay rate of mRNA, may alter mRNA expression levels by affecting the binding affinity or activity of corresponding functional elements. An approach, genetical genomics, first introduced by Jansen and Nap, aims to identify genetic variants that modulate gene expression by merging the analyses of genetic variations and expression levels (Jansen & Nap, 2001). Such expression quantitative trait loci (eQTL) mapping utilizes statistical techniques to identify correlations between quantitative measurements of mRNA expression and genetic polymorphisms segregating in a population (Farrall, 2004; Gilad, Rifkin, & Pritchard, 2008). There are two main strategies for QTL mapping: association tests and linkage analysis in a cross population (Alberts et al., 2005). Technological developments and cost decreases in microarrays now allow the simultaneous measurement of the expression levels of thousands of genes in a large number of individuals from various species, as well as genotyping the status of up to a million variants in each individual, usually SNPs. Early studies have mapped genetic variants to gene expression in a number of model organisms, establishing the power of the approach. Organisms have included maize (Salvi et al., 2007; E.E. Schadt et al., 2003), Arabidopsis (DeCook, Lall, Nettleton, & Howell, 2006), yeast (Brem, Yvert, Clinton, & Kruglyak, 2002; Yvert et al., 2003), Caenorhabditis elegans (Y. Li et al., 2006), fly (Wayne & McIntyre, 2002), mice (Bystrykh et al., 2005; Chesler et al., 2005; Doss, Schadt, Drake, & Lusis, 2005), rats (Hubner et al., 2005; Petretto, Mangion, Pravanec, Hubner, & Aitman, 2006), and humans (Cheung et al., 2005; Deutsch et al., 2005; Monks et al., 2004;

Morley et al., 2004; Stranger et al., 2005). Next generation sequencing of RNA (RNA-Seq) is beginning to supplant microarray technology, using high-throughput sequencing platforms to obtain relatively unbiased measurements of expression across the entire length of a transcript (Zhong Wang, Gerstein, & Snyder, 2009). This technology has several advantages, including access to rare transcripts, more accurate quantification of abundance transcripts, novel gene structure, and alternative splicing (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008; Pan, Shai, Lee, Frey, & Blencowe, 2008; Sultan et al., 2008; E. T. Wang et al., 2008). Most recently, complete genome sequencing has also begun to supplant the use of microarrays for genotyping (Lappalainen et al., 2013).

Many genome-wide eQTLs mapping studies have now been performed in a variety of human tissues and populations (Grundberg et al., 2012; Lappalainen et al., 2013). So far, hundreds of thousands of cis- and trans- regulatory eQTLs have been discovered. Cis-regulatory eQTL, where the presence of a genetic variant is associated with the level of transcripts from a gene located within a few hundred kilobases, have been predominantly reported. Trans-regulatory eQTL associations, where the genetic variant is distant from the transcript locus, are much harder to reliably identify due to multiple testing problems: analysis of trans effects involves of the order of 10^4 more statistical tests than for cis effects.

Initially, most human eQTL mapping studies measured transcript abundance in easily accessible blood cells (peripheral blood lymphocytes and Epstein-Barr virus (EBV) transformed immortalized lyphobalstoid cell lines (LCLs)) (Morley et al., 2004).

Since gene expression profiles vary in different tissues, it is only to be expected that some eQTLs are tissue specific, and it has been reported that 33-69% of eQTLs, depending on the analysis method and the tissue type, are not discovered in other tissues (Ding et al., 2010; Michaelson, Alberts, Schughart, & Beyer, 2010; Zeller et al., 2010). Recently, a number of studies have been performed on other human tissues or cell types (e.g., brain (Gibbs et al., 2010; Myers et al., 2007), liver (Greenawalt et al., 2011; Innocenti et al., 2011; Eric E Schadt et al., 2008), adipose (Emilsson et al., 2008; Greenawalt et al., 2011; Nica et al., 2011), fibroblasts (Dimas et al., 2009), and skin (Ding et al., 2010; Nica et al., 2011). Dimas et al. (Dimas et al., 2009) identified eQTLs in three cell types: primary fibroblasts, LCLs and T–cells and estimated that a large proportion of regulatory variants are cell type-specific.

1.2: Mapping complex disease traits with global gene expression

Genome-wide association studies (GWAS) are providing a powerful approach to identifying common disease loci, and the GWAS catalog currently contains ~17,600 loci where variants have been found to be associated with the phenotypes of ~1100 human complex traits (http://www.genome.gov/gwastudies). In each of these loci, there must be some mechanisms whereby genetic variants affect the function of one or more gene products.

There are a variety of possible mechanisms, including missense, where a resulting amino acid substitution in some way affects the level of function of a protein; effects on splicing; and effects on expression. Since eQTL studies have shown that SNPs affecting expression are widespread, it is likely that some of these are involved in the underlying mechanisms in disease loci.

Two studies have demonstrated that SNPs associated with human traits are in general enriched for eQTLs (Cookson, Liang, Abecasis, Moffatt, & Lathrop, 2009; Nicolae et al., 2010). Another study also showed that chemotherapeutic drug susceptibility associated SNPs is enriched in eQTLs (Gamazon, Huang, Cox, & Dolan, 2010). A number of studies have incorporated information from eQTL association results and shown these provide a promising approach for improving functional interpretation of disease GWAS findings (Chu et al., 2011; Ertekin-Taner, 2011; Heid et al., 2010; Hrdlickova, Westra, Franke, & Wijmenga, 2011; Hsu et al., 2010; Lango Allen et al., 2010; Moffatt et al., 2007; Richards et al., 2012; Speliotes et al., 2010; Wu et al., 2012) and for prioritizing genes in an association region for functional experiments using animal models (Teslovich et al., 2010). Most of these studies used the most accessible eQTL data from LCL, and the general suitability of these for the study of traits/diseases not relevant to LCL has still to be established. Some studies have used eQTL data derived from a tissue appropriate to the disease of interest to link to disease-associated SNPs (Ding et al., 2010; Fransen et al., 2010; Innocenti et al., 2011; Kang, Morgan, & Chen, 2012; Liu, 2011; Richards et al., 2012; Eric E Schadt

et al., 2008; Zhong et al., 2010). Ding et al. (Ding et al., 2010) reported an eQTL study of human skin that aimed to elucidate the role of regulation of gene expression in psoriasis. Innocenti et al. (Innocenti et al., 2011) and Schadt et al. (Eric E Schadt et al., 2008) mapped eQTLs in human liver tissue and demonstrated the role of some candidate genetic variants that affect gene expression and so play a role in human common diseases, for example NOD2 expression in leprosy, C2orf43 in prostate cancer, SORT1 expression in Coronary artery disease, CELSR2 expression in LDL cholesterol levels, and RPS26 expression in Type 1 diabetes. These studies demonstrate that eQTL mapping can facilitate efforts to understand the relationship between expression differences caused by genetic variations and human common diseases. Table 1.1 shows the studies that have used eQTL data to discover candidate causal genes for some human common diseases. Genes identified in this way might be useful to prioritize candidate genes and pathways associated with the risk of complex diseases and traits, such as basal cell carcinoma in a skin cancer GWAS (M. Zhang et al., 2013) and type 2 diabetes (Zhong et al., 2010). These studies suggest that genome-wide eQTL results provide an important reference source for investigating the expression effects of disease-associated SNPs and for prioritizing disease causal gene.

Table 1.1. Summary of candidate causal genes for human common diseases that were discovered using eQTL data

Complex trait	Genes	Reference
Acute lymphobalstic leukemia	GSTM2; GAPDH; NCOR1	(French et al., 2008)
Alzheimer disease	Miltiple genes	(Webster et al., 2009)
Asthma	GSDMA; ORDML3; HCG26; MEF2C; HLA-DQB1; GPSM3; PBX2; NUP35; POM121L2	(L. Li et al., 2013; Moffatt et al., 2007)
Celiac disease	ILI8RAP; CCR3; IL12A; RGS1; SH2B3; TAGAP	(Heap et al., 2009; Hunt et al., 2008)
Coronary artery disease	SORT1	(Eric E Schadt et al., 2008)
Crohn's disease	UBE2L3; BCL3	(Fransen et al., 2010)
Drug metabolism	ADME	(Schröder et al., 2011)
Glucocorticoids	NQO1; AIRE; SGK1	(Maranville et al., 2011)
Graves' disease	RNASET2; FGFR1OP; GDCG4p14	(Chu et al., 2011)
LDL cholesterol levels	CELSR2	(Eric E Schadt et al., 2008)
leprosy	NOD2	(Innocenti et al., 2011)
pancreatic cancer	BACH1	(Wu et al., 2012)
prostate cancer	C2orf43	(Innocenti et al., 2011)
Psoriasis	FUT2, TMEM77,RPS26, LOC348751, C17orf45, ERAP2, TNRC6B, ENDOD1	(Ding et al., 2010)
Type 1 diabetes	RPS26	(Eric E Schadt et al., 2008)
Type 2 diabetes	ME1	(Zhong et al., 2010)
Waist-hip ratio	TBX15, AA553656, GRB14, PIGC, ZNRF3, STAB1	(Heid et al., 2010)

1.3: Integration of eQTL studies

Despite the large number of identified human eQTL associations, few have been convincingly reproducible in multiple studies (Dixon et al., 2007; Göring et al., 2007; Myers et al., 2007; Stranger et al., 2007; Veyrieras et al., 2008). For example, Choy et al. showed that they were unable to detect evidence that eQTLs are convincingly associated with drug response by using eQTL data from LCL, probably as a consequence of biological noise and in vitro confounding artifacts (Choy et al., 2008). These observations show that putative eQTLs derived from a single study should be considered with care. Indeed, in our comparison of the results from 16 human genome-wide eQTL mapping studies, we observed that a large fraction of reported eQTLs are not reproduced in other studies. This is the case even for studies using the same cell types and with the same individuals, such as LCL data for the HapMap CEU population (Altshuler et al., 2010).

A number of suggestions have been made to explain the apparent underlying low reproducibility of eQTLs. First, various expression microarray chips with different selected probes were used in different eQTL studies. Secondly, mRNA sequence polymorphisms in probe regions are known to influence hybridization on microarrays considerably (Gilad, Rifkin, Bertone, Gerstein, & White, 2005). Thus, mRNA fully matching the probes on the microarrays may hybridize better than mRNA that happens to carry a genetic variant in particular individuals, causing a difference in signal unrelated to expression level. The relatively new technology for gene

expression profiling, RNA-seq (Zhong Wang et al., 2009), already used in some eQTL studies, may provide a more reliable way for large-scale measurement of differences in gene expression.

Thirdly, the complexity of microarray-based gene expression analysis has been shown to be one of the critical reasons for the difficulty reproducing expression studies (Ioannidis et al., 2009). There are many steps of data processing and analysis for which parameters and procedures are some times inadequately described, making replication difficult.

Fourthly, some hidden confounding factors, including population structure (e.g., race, family-relatedness) and microarray array artifacts (batch effects), could also lead to spurious or missed associations (Listgarten, Kadie, Schadt, & Heckerman, 2010). In addition to these general issues for expression analysis, there are many confounders in specific cases. For example, for the studies that have used the most accessible cell type, LCL, there are many non-genetic factors that may be introduced in the path from the human donor to the study of an LCL *in vitro*. When new immortalized LCLs are obtained by infecting B-cells with the EB virus, a different sub-population of B-cell will be selected, the amount of an individual response to the EB virus will vary, and so will the history of passage in cell culture as well as culture conditions.

The apparent unreliability of eQTL studies hinders their use in analysis of complex trait disease. One way to address this issue is to compare the data across studies, and

in this way identify sets of eQTLs that have been observed multiple times, independently. To this end, we have integrated a number of publicly available eQTL studies, and have extracted several merged datasets that incorporate the consensus SNP-gene association pairs across various tissues and population combinations. In this study, linkage disequilibrium (LD) relationships were used to assist in determining these consensus associations. Linkage disequilibrium is a measure of the degree to which the presence of alleles at two loci are correlated. LD varies across the human genome, but typically extends up to at least 200 Kilobases. Thus, measurement of the status of one SNP in an individual provides information about the probability of the presence many SNPs nearby. Without this property, current microarrays, which only measure that status of about one million SNPs, would not effective for either eQTL or disease GWAS studies. However, there is a down side. LD also makes it quite difficult to distinguish which of the many alleles in a region that is associated with a phenotype is the real causal allele – many may exhibit approximately equal association strength with the phenotype. In this study, we used LD information to identify sets of SNPs that are highly correlated with marker eQTL SNPs and so define a consensus region representing one underlying QTL relationship.

1.4: Overview

This dissertation is organized as follows. In Chapter 2, we start with an overview of current available human genome-wide eQTL studies and show that consistency across these eQTL datasets is low. We then introduce a method to identify the high-confidence eQTL associations by comparison across eQTL datasets. We also assess

the population-specific and tissue-specific eQTLs and the possible mechanisms underlying expression regulation for these eQTL associations. Chapter 3 first describes a method to map GWAS disease risk loci with the high-confidence eQTL associations. We then describe the application of this method for seven human diseases and diseases relevant to four tissue types to identity a set of disease risk loci with potential underlying expression mechanisms. Chapter 4 first summarizes a list of web-based eQTL databases and software. We then introduce an interactive and user-friendly integrated web database that we have designed for querying and visualizing all available human eQTL data and high-confident eQTL associations. This comprehensive database will also allow better utilization of these data to elucidate the role of eQTLs in a set of complex trait diseases. Chapter 5 summarizes the conclusions from this project and discusses current challenges and future perspectives in genome-wide human eQTL studies, especially in relationship to complex trait disease.

Chapter 2: Meta-analysis of Genome-wide Expression Quantitative Trait Loci (eQTLs) for Various Human Populations and Cell Types

2.1: Introduction

Genome-wide eQTL association studies, combining whole-genome scale SNP genotyping arrays and whole-transcriptome expression arrays, have provided a powerful means of linking genetic variants to gene expression. Advances in high-throughput technology in microarrays make it feasible to efficiently and quantitatively measure mRNA levels of thousands of genes in parallel (Schena et al., 1996). Genotyping microarrays allow the status of a representative set of up to a million SNPs to be determined in a set of individuals. Statistical analysis of these two complementary types of data then permits associations between the presence of a SNP and the level of each transcript, so identifying expression quantitative trait loci (eQTLs). Application of these technologies for expression quantitative trait loci (eQTL) mapping studies for cells from various organisms, such as maize (Salvi et al., 2007; E.E. Schadt et al., 2003), Arabidopsis (DeCook et al., 2006), yeast (Brem et al., 2002; Yvert et al., 2003), Caenorhabditis elegans (Y. Li et al., 2006), fly (Wayne & McIntyre, 2002), mice (Bystrykh et al., 2005; Chesler et al., 2005; Eric E Schadt et al., 2005), rats (Hubner et al., 2005; Petretto et al., 2006), and humans (Cheung et al., 2005; Deutsch et al., 2005; Monks et al., 2004; Morley et al., 2004; Stranger et al., 2005) have revealed a large number of SNP/gene associations. Newer technologies, particularly RNA-seq for the measurement of transcript levels and whole genome

sequencing for genotyping are beginning to replace microarray use (Lappalainen et al., 2013).

A number of genome-wide eQTL association studies have been conducted in Epstein-Barr virus-transformed lymphoblastoid cell lines (LCLs), utilizing genotype data of various human populations to discover the genetic variants contributing to differences in gene expression within and among human ethnic groups (Price et al., 2008; Spielman et al., 2007; Storey et al., 2007; W. Zhang, Duan, Kistner, & Bleibel, 2008). In addition, a number of studies have reported eQTL associations identified in other cell types, including primary fibroblasts (Dimas et al., 2009), primary monocytes (Fairfax et al., 2012; Rotival et al., 2011; Zeller et al., 2010), as well as cells from brain (Gibbs et al., 2010; Myers et al., 2007), liver (Innocenti et al., 2011; Eric E Schadt et al., 2008), adipose (Emilsson et al., 2008; Nica et al., 2011), and skin (Ding et al., 2010; Nica et al., 2011) tissues.

Several studies have demonstrated that eQTLs are involved in higher-level cellular phenotypes, such as development, differentiation, and maintenance, as well as whole-body traits including disease susceptibility (Cookson et al., 2009; Hamza et al., 2011; Loo et al., 2012; Moffatt et al., 2007; Nica et al., 2010; Nicolae et al., 2010) and personalized drug response (Choy et al., 2008; Schröder et al., 2011).

While producing large quantities of data, these studies suffer from considerable noise arising from the use of variety of statistical models and experimental limitations. A

natural next step is to derive more complete and reliable eQTL associations by combining results from multiple studies. Our objective in this work was to develop a method to integrate the results from available eQTL studies, and to establish a database that provides an efficient way to prioritize eQTLs for further analysis, particularly in the context of the role of expression variation in complex trait disease.

To this end, we have integrated eQTL data from 16 publicly available genome-wide eQTL studies covering various human tissues and populations, and identified consensus SNP-gene associations across studies. We have also compared eQTLs across different tissues and populations so as to estimate the proportions of tissue-dependent and population-dependent relationships. In order to help identify mechanisms underlying these eQTL associations, we have mapped eQTLs to annotated functional elements, discovering enrichments of tissue-specific transcription factor binding sites.

2.2: Results

Summary of Genome-wide eQTL.

Table 2.1 summarizes 16 publicly available genome-wide eQTL studies collected in this study, and categorized them into 29 datasets by tissue and population. The majority of studies as far have been performed on Lymphoblastoid cell lines (LCLs). In addition to LCLs, we also include data from various tissues, mostly from Caucasian populations. The 3C study (Dimas et al., 2009) covers three cell types (3CL:LCLs, 3CF: primary fibroblasts, and 3CT: primary T-cells). Two studies (Gibbs

et al., 2010; Myers et al., 2007), BR and BR2, assessed the transcriptomes of cells from different brain regions. Two studies, MO (Zeller et al., 2010) and IM (Fairfax et al., 2012), were performed on various circulating immune cells, specifically primary monocytes (MO; IM_MO) and B-cells (IM_B). Two studies, LV (Eric E Schadt et al., 2008) and LV2 (Innocenti et al., 2011), investigated eQTLs in liver cells. A study of eQTLs in skin, SKN (Ding et al., 2010), is also included.

Most of the studies are on Caucasian populations. Hapmap (The International HapMap 3 Consortium, 2010) populations have often been used, and in addition to Caucasian (HA_CEU; HA2_CEU; HRC), we include Chinese (HA_CHB), Japanese (HA_JPT), and Yoruba (HA_YRI; HA2_YRI; HRY) data derived from Hapmap populations (Duan et al., 2008; Montgomery et al., 2010; Pickrell et al., 2010; Stranger et al., 2007) .

Most studies used a combination of genotyping microarrays and transcription microarrays. Three studies, HRC (Montgomery et al., 2010), HRY (Pickrell et al., 2010), and E-GEUV (Lappalainen et al., 2013), all on LCLs, used RNA sequencing technology rather than the older microarray technology to determine expression levels. One study, E-GEUV, used the 1000 genomes project populations (EUR and YRI), and so was able to include the genotypes of all SNPs down to about a frequency of 1%, instead of the limited number represented on a genotyping microarray.

In this study, we define 'exSNPs' as the SNPs that correlate with change of expression of one or more genes. The corresponding genes are referred to as 'exGenes', and an 'eQTL association' represents the relationships between one exSNP and its associated exGene. After processing the raw data from the 16 studies, there are totally 796,908 unique eQTL associations covering 15,170 unique exGenes and 548,344 unique exSNPs. The number of eQTL associations varies widely across studies (522 ~ 390,813). Variation in population sample size is probably the biggest factor in this spread (sample sizes range from 30 to 1490). The expression level of most exGenes is associated with the presence of multiple exSNPs, primarily as a result of linkage disequilibrium, and in most cases only a single variant is likely actually causative of a change in expression.

As in common practice, we consider 'cis-eQTL' associations to be those where the exSNPs are located within 1Mb of either the 5' or 3' end of the associated exGene. eQTL associations between an exGene and an exSNP located more than 1 Mb distant away from the gene region are referred to as 'trans-eQTL' associations. Figure 2.1 shows the proportion of cis and trans-eQTLs in each dataset. Most datasets have a much higher fraction (> 60%) of cis eQTLs. The predominance of cis-eQTLs is largely a consequence of the increased statistical power obtained by limiting the genome window in which associations are examined, thereby greatly reducing the size of multi-testing correction needed. Figure 2.2 shows the distribution of distances between exSNP-exGene pairs. The density falls off rapidly with distance, and 85% of cis-regulatory exSNPs are within 200Kb of the corresponding exGene. cis-eQTLs are

approximately symmetrically distributed both upstream and downstream of the corresponding exGene, as well as within the gene. About 25% of cis-regulatory exSNPs fall within a gene region, and were assigned a distance of zero. Although linkage disequlibrium broadens this distribution, it is still apparent that the majority of SNPs involved in cis-eQTL relationships are located in the vicinity of the affected gene, including the 5' and 3' UTRs, and neighboring up-stream and down-stream regions. Because of linkage disequilibrium, it is difficult to determine the exact location of the underlying causal variants that directly affect gene expression.

Table 2.1. eQTL data for 16 selected genome-wide eQTL association studies

Study ID	Samples (size)	Cell type	Genotyping	Phenotyping	eQTL associations	exSNPs	exGenes
HA	HapMap CEU (30) HapMap CHB (45) HapMap JPT (45) HapMap YRI (30)	LCL	HapMap Project	Illumina Human WG-6	3858 4066 5254 3524	3686 3780 5061 3283	239 253 274 306
BR	Caucasians (193)	Brain Cortex	Affymetrix 500K	Illumina HumanRefseq-8	624	545	209
AS	Childhood Asthma (206)	LCL	1. Illumina Human-1 2. Illumina HumanHap 300	Affymatrix HG-U133	21116	12121	2632
LV	Caucasian liver donors (427)	Liver cell	1. Illumina 650Y 2. Affymetrix 500K	Custom ink-jet microarray	4362	2527	3824
HA2	30 HapMap CEU (30) 30 HapMap YRI (30)	LCL	HapMap Project	Affymetrix GeneChip Human Exon 1.0	4453 5027	3699 4086	722 1659
3C	Caucasians (75)	LCL Fibroblast T-cell	Illumina 550K	Illumina Human WG-6	554 522 546	544 508 540	436 424 429
MO	German (1490)	Monocyte	Affymetrix 6.0	Illumina Human HT-12	37694	29948	2752
HRC	HapMap CEU (60)	LCL	HapMap Project	RNA-Seq	8908	3896	930
HRY	HapMap YRI (69)	LCL	HapMap Project	RNA-Seq	799	779	786
BR2	Caucasians (150)	Cerebellum Frontal cortex Temporal cortex Pons	Illumina Infinium HumanHap 550	Illumina HumanRef-8	5243 5512 5335 3411	4399 5198 4059 3284	317 329 385 275
SKN	Healthy skin individuals (57)	Skin	Perlegen Sciences array	Affymatrix HG-U133	5410	4782	222
LV2	Liver donors (266)	Liver cell	1. Illumina 610	1. Agilent-014850	1170	1161	1170

			2. Illumina HumanHap 550	2. Illumina HumanRef-8			
IM	British (288)	Monocyte B-cell	Illumina Human OmniExpress-12	Illumina HumanHT-12	33740 22453	28956 20333	6063 5449
MuTHER	Caucasian female twins (~160)	LCL Skin Adipose	Illumina 1. HumanHap 300 2. HumanHap 610Q 3. 1M-Duo 4. 1.2MDuo 1M	Illumina Human HT-12	211977 103537 138885	149684 82933 109689	3945 2495 3136
MRC	Childhood Asthma (MRCA: 405) & Atopic Dermatitis (MRCE: 950)	LCL	1. Illumina Human-1 2. Illumina HumanHap 300	1. Affymetrix HG-U133 2. Illumina Human WG-6	176848	109763	1251
E-GEUV	1000 Genome - EUR (373) 1000 Genome - YRI (89)	LCL	1000 Genome Project	Illumina HiSeq 2000	390813 19314	281446 16932	3048 472

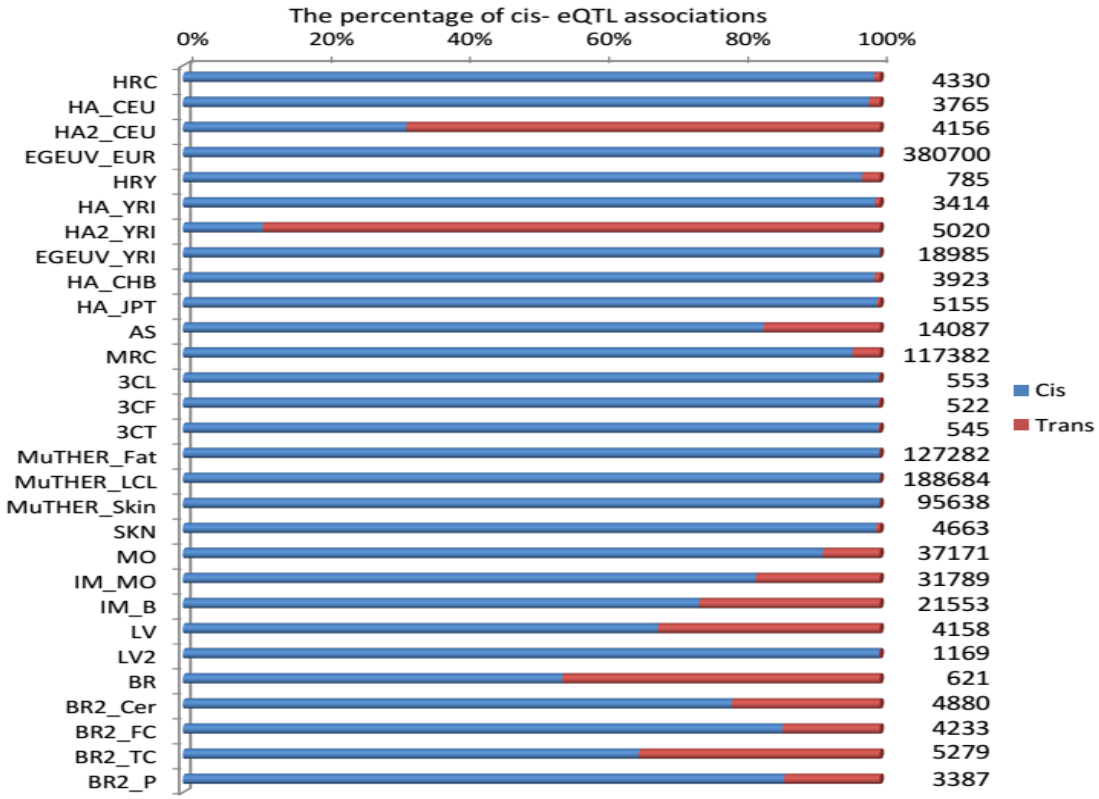


Figure 2.1. Fraction of cis- and trans- eQTL associations in each dataset.

The number at the end of each row is the total number of eQTL associations. The fraction of associations classified as trans-eQTL in the HA2 study is much higher than others as a result of that study considering all associations out to 4Mb, as opposed to the more usual 1Mb.

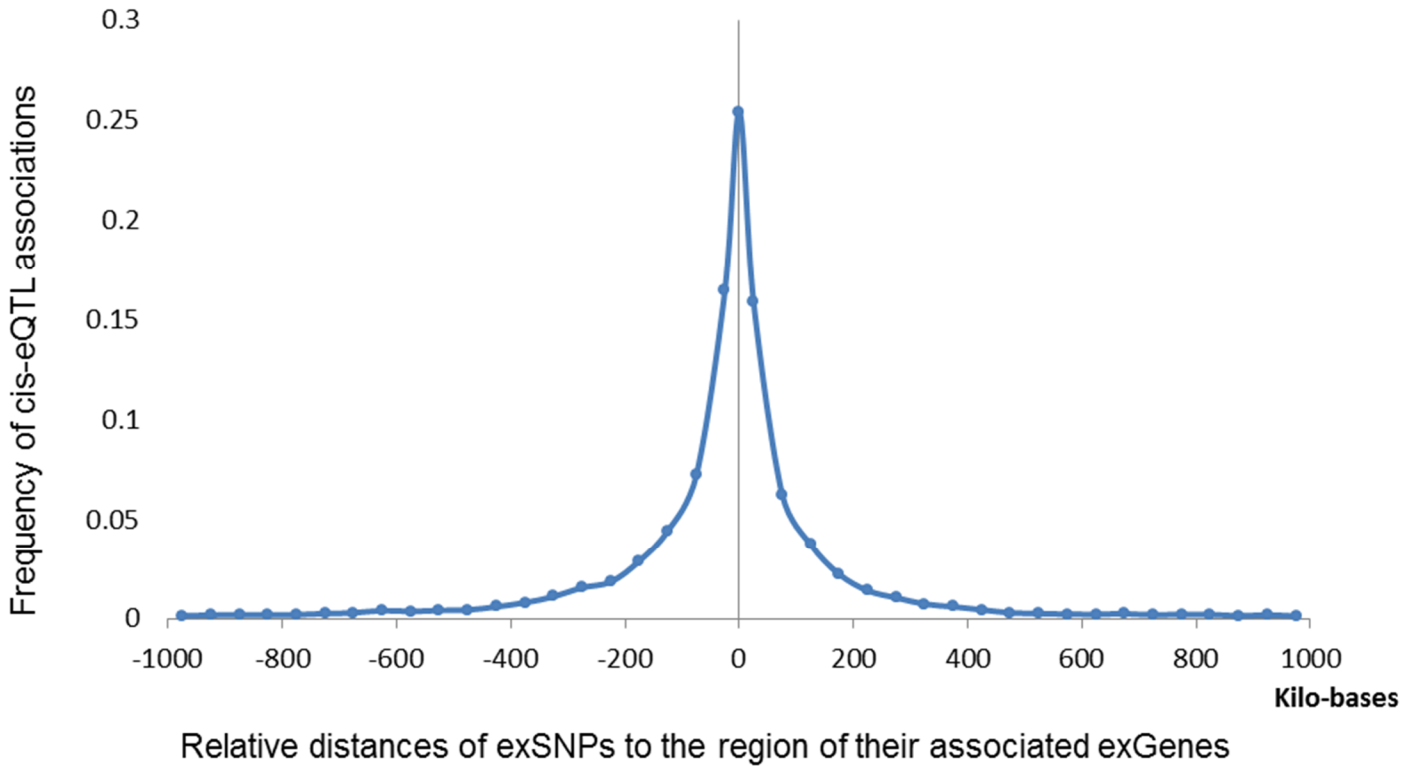


Figure 2.2. Distribution of distances between each cis-exSNP and the associated exGene.

Most distances are relatively short – less than 200Kb.

Linkage Disequilibrium relationships between eQTLs

We presume that the underlying mechanistic origin of a cis-eQTL relationships is that a particular SNP or other variant falls on a functional element, such as a transcription factor binding site, a microRNA binding site, or splice site where the change leads to non-sense mediated decay. Then an association study will reveal a statistical relationship between the presence of that causal SNP and the level of expression of the gene. Typically, because of incomplete recombination within human genomes, any such variant is in linkage disequilibrium with a number of others nearby. That is, the presence of these other SNPs is correlated with the presence of the mechanism SNP. As a consequence, these nearby SNPs will also exhibit a correlation with the expression level of the same gene. In principle, it might be possible to identify which of the set of such SNPs is causal from the strength of the correlation between its presence and the level of gene expression. In practice, LD is often close to 1 for a number of neighboring SNPs, and the data are usually noisy, so it is not possible to make such a determination. Further, low sampling of SNPs using typical microarray genotyping technology (about 1 million out of the approximately 40 million SNPs) are usually included, only some of the nearby SNPs are assayed, and it is unlikely the causal SNP will itself be assayed.

In spite of these limitations, it is usually possible to group exSNPs into LD blocks, and so approximately identify the number of unique causal relationships – each block will usually represent one relationship. To this end, for each dataset, we first determined the LD relationship (r^2) between all the SNPs exhibiting an eQTL

relationship with the same gene, using LD data for the corresponding populations from the Hapmap (Altshuler et al., 2010) and 1000 Genomes projects (Abecasis et al., 2010). For the two studies where the ethnicity of samples is not reported, IM and MO, we selected only SNP pairs with strong LD relationships appeared in all populations.

Table 2.2 shows the number of total eQTL associations and the corresponding number of unique exSNPs and unique exGenes in each dataset. It also shows the number of unique eQTL relationships, each of which represent a set of LD related exSNPs associated with the same exGene, at three LD thresholds, $r^2 \geq 0.8$, 0.5, and 0.3. Here each eQTL relationship likely represents one mechanistic relationship between the presence of a causal variant and the expression level of the gene. The proportion of exGenes with single eQTL relationship ranges from 54% - 100% with an LD threshold of 0.8 to 72% - 100% at a threshold of 0.3. Some studies, EGEUV, MRC, MO studies have many genes with multiple eQTL relationships may be because the LD information between exSNPs associated with the same exGene is missing. Within each single study, we have confirmed that exSNPs in the same LD block are overwhelmingly associated with a change of expression of the same exGene (see Appendix).

Table 2.2. eQTL associations and unique eQTL relationships for each dataset

Dataset	Unique eQTL associations	Unique exGenes	Unique exSNPs	Unique eQTL relationships ($r^2 \geq 0.8$)	Unique eQTL relationships ($r^2 \geq 0.5$)	Unique eQTL relationships ($r^2 \geq 0.3$)
HRC	4362	930	3896	1453	1116	1038
HA_CEU	3787	239	3686	451	286	252
HA2_CEU	4163	722	3699	1273	1166	1141
EGEUV_EUR	390696	3048	281446	135826	103879	88142
HRY	794	786	779	794	792	790
HA_YRI	3419	306	3283	619	372	336
HA2_YRI	5027	1659	4086	3007	2835	2813
EGEUV_YRI	19314	472	16932	9349	6887	5709
HA_CHB	3930	253	3780	453	293	265
HA_JPT	5165	274	5061	481	317	290
AS	14348	2632	12121	6596	4178	3328
MRC	119958	1251	109763	17019	10894	8959
3CL	554	436	544	531	494	469
3CF	522	424	508	501	462	443
3CT	546	429	540	525	475	462
MuTHER_Fat	128181	3136	109689	19704	9056	5367
MuTHER_LCL	189983	3945	149684	28861	12913	7379
MuTHER_Skin	96412	2495	82933	14236	6471	3883
SKN	4916	222	4782	384	243	227
MO	37580	2752	29948	29690	23130	17598
IM_MO	31914	6063	28956	27794	23695	19929
IM_B	21674	5449	20333	19244	16665	14361
LV	4171	3824	2527	4145	4126	4117

LV2	1170	1170	1161	1170	1170	1170
BR	624	209	545	358	323	315
BR2_Cer	5241	317	4399	572	374	344
BR2_FC	5429	329	5198	625	381	347
BR2_TC	5280	385	4059	681	441	409
BR2_P	3389	275	3284	475	312	285
Total	-	15170	578094	-	-	-

Unique eQTL associations are the numbers of unique exGene-exSNP pairs. Unique exGenes are the numbers of unique genes that have at least one eQTL association. Unique exSNPs are the numbers of SNPs that are involved in at least one eQTL association. Unique eQTL relationships are the numbers of unique blocks of LD-linked exSNPs, where all exSNPs in a block are associated with a change of expression of the corresponding exGene.

Pair-wise comparisons show low agreement between eQTL datasets

To investigate how often the same eQTL relationships are found in different studies, we compared the eQTL associations between each pair of datasets and identified the common exGenes and the exSNPs that are associated with these. Table 2.3 summarizes the level of agreement among the 16 different eQTL data sets. The diagonal shows the number of genes with at least one eQTL relationship discovered in each study (the exGenes), and the numbers in the upper triangle show the percentage of exGenes which are common between each pair of studies. In general, the agreement of most (92%) of pairwise comparisons between datasets is low (4%-49%).

Some differences between eQTL studies presumably arise from different biology as a function of cell type and population. However, the fraction of common exGenes for studies on the same population and cell type are also often low. For example, the fractions of common exGenes among studies performing in LCLs for Caucasian populations (HRC, HA_CEU, HA2_CEU, and EGEUV_EUR) are usually not high (8% - 27%), with the exception of the comparison between EGEUV_EUR and HA_CEU (57%). Similarly, for the African population studies, HRY, HA_YRI, HA2_YRI, and EGEUV_YRI, the common exGene fractions between pairs are also usually low (14% - 33%), with the exception of the comparison between EGEUV_YRI and HRY (42%). In this latter case, the relatively high agreement may be because both studies used RNA-Seq technology. Studies on other cell types (skin,

liver, and two types of brain tissue) with the same populations tend to show an intermediate level of agreement (45%-59%).

Studies on the same cell line but different populations have agreements of 7%-35%, in the same range as those with the same population. There are two exceptions. First are the studies between EGEUV_CEU and HRY (56%), both RNA-Seq studies. Second are the studies using the 1000 genomes Caucasian and African populations, EGEUV_CEU and EGEUV_YRI, which share the highest fraction (77%) of exGenes, again likely reflecting the high quality of the recent RNA-Seq studies (Lappalainen et al., 2013), and also because essentially all SNPs with the frequency greater than 1% are included, removing the difficulties of comparing results from microarrays with different SNPs subsets. An anomaly is a relatively high level of agreement between EGEUV_EUR and HA_YRI (48%). The latter is an older study using a transcription array.

Three sets of studies (MuTHER, 3C, and IM) measure expression in different cell types from the same population. The MuTHER study (Grundberg et al., 2012) used adipose, LCL, skin in a Caucasian population. Here levels of agreements are high (54%-60%). It is unclear to what extent this is a consequence of protocol and technology as opposed to expression being independent of sample type. The 3C (Dimas et al., 2009) study used LCLs, fibroblast, and T-cell lines in a Caucasian population. Here agreement is lower (29%-31%). The third study, IM (Fairfax et al.,

2012) measured expression in monocyte and B-cell lines in a British population, producing an intermediate level of agreement (47%).

The numbers in the lower triangle show the percentage of common exGenes which are associated with the same or LD related exSNPs between each pair of studies. This fraction should indicate if the unique eQTL relationships found in one study are the same as found in others. In general, HA_CEU, HA_CHB, HA_JPT and EGEUV_EUR have higher fractions of common exGenes with LD related exSNPs compared to other datasets, especially in LCL (50%-95%). Datasets from the IM study, IM_MO and IM_B, shows much lower fractions (0%-35%).

There are several possible reasons for low consistency between studies, besides that due to different cell types and populations. First, a variety of genotyping arrays, with different tag SNPs and different probes have been used (see Table 2.1). Secondly, early studies relied on RNA microarrays to estimate transcript levels. Only three studies used more recent RNA-Seq technology. Thirdly, the analysis procedures and statistical models used in each study vary (for example, linear regression models, Spearman rank correlation). In addition, there are other possible confounders arising in the experimental procedures, for example the history of a cell culture and culture conditions, and differences in experimental protocols. Despite these issues, there is evidence that a substantial proportion of the cis-eQTL findings are reproducible (Greenawalt et al., 2011; Innocenti et al., 2011). Innocenti et al. estimated 49% -

67% cis-eQTL reproducibility between several datasets conducted in liver, which is consistent with the comparison between LV and LV2 (59%) in our datasets.

We also performed hierarchical clustering to further compare these datasets (Figure 2.3 & 2.4). Figure 2.3 shows the hierarchical cluster of all datasets based on the fraction of common exGenes. Two factors dominate the tree topology, cell type and specific study. Most of the datasets that used LCLs are grouped in one major branch, and studies that used monocytes or liver are also grouped. Datasets from the same study are usually grouped together, for example the 3C study, the BR2 study, and the MuTHER study (excepting MuTHER_LCL which is in the LCL group). Figure 2.4 shows the hierarchical cluster based on the fraction of common exGenes which are associated with LD related exSNPs between pairs of studies. The tree structure here is similar to that based just on common exGenes, again with the major factors determining tree topology being the cell type and the study.

Table 2.3. Pair-wise comparisons of eQTL datasets

Overlapped gene_ratio	HRC	HA CEU	HA2 CEU	EGEUV EUR	HRY	HA YRI	HA2 YRI	EGEUV YRI	HA CHB	HA JPT	AS	MRC	3CL	3CF	3CT	MuTHER Fat	MuTHER LCL	MuTHER Skin	SKN	MO	IM MO	IM B	LV	LV2	BR	BR2 Cer	BR2 FC	BR2 TC	BR2 P
HRC	928	18	8	27	10	11	17	11	13	15	25	15	14	10	10	23	33	19	14	23	38	33	25	12	8	10	10	6	9
HA_CEU	53	239	12	57	23	35	14	15	37	38	62	53	27	20	21	47	65	45	14	54	47	37	49	38	6	11	11	12	9
HA2_CEU	25	52	722	24	8	7	24	7	12	10	20	14	8	5	7	22	32	18	9	20	38	33	24	9	6	6	5	7	7
EGEUV_EUR	37	78	22	3048	56	48	21	77	58	54	32	59	50	37	40	29	49	29	42	34	37	34	30	39	24	33	35	26	26
HRY	35	54	31	66	780	33	16	42	24	24	37	24	22	12	14	34	49	28	17	31	39	34	30	19	6	10	11	8	7
HA_YRI	35	76	27	69	73	304	14	24	32	34	47	40	15	9	11	38	54	36	10	40	44	36	44	24	7	8	7	7	7
HA2_YRI	5	21	12	10	23	32	1659	16	15	15	19	13	13	12	12	20	29	17	13	20	35	31	21	13	12	14	12	9	13
EGEUV_YRI	55	68	52	78	74	80	36	472	17	14	39	24	14	8	8	28	45	25	12	33	37	31	31	21	5	7	4	4	5
HA_CHB	36	89	39	77	52	78	16	62	253	47	49	46	19	15	15	42	62	40	13	47	46	36	44	30	8	11	10	11	9
HA_JPT	28	87	37	76	64	81	10	54	95	274	52	47	18	14	15	46	64	43	11	49	48	43	43	26	7	8	8	8	8
AS	27	75	20	66	35	52	6	48	77	69	2629	49	45	36	38	28	39	25	64	27	40	36	30	34	27	31	30	26	26
MRC	31	86	19	89	68	80	8	77	86	83	81	1251	29	23	22	42	68	35	32	45	51	46	39	23	24	22	22	19	20
3CL	45	74	44	75	24	48	12	46	67	65	68	77	435	29	31	44	60	40	18	51	42	42	41	25	6	12	11	10	12
3CF	24	67	26	54	17	48	8	31	58	42	56	68	76	424	29	50	52	43	19	51	46	39	41	25	7	13	12	10	13
3CT	27	63	24	53	21	36	9	35	57	57	57	63	80	72	428	38	49	35	15	50	43	37	40	21	9	11	12	10	11
MuTHER_Fat	21	70	10	60	42	52	5	47	74	66	52	72	68	70	65	3134	54	60	48	41	47	40	33	43	35	47	49	42	42
MuTHER_LCL	23	77	12	79	59	64	5	57	79	77	66	90	80	70	66	77	3944	58	51	48	47	42	30	44	32	50	53	43	48
MuTHER_Skin	20	73	12	64	46	58	6	45	72	69	55	76	72	74	67	85	80	2495	45	38	46	41	34	36	28	40	44	36	37
SKN	50	88	29	78	51	82	28	58	86	88	87	83	73	81	61	85	80	90	222	50	43	42	59	44	7	12	11	11	8
MO	23	74	13	62	38	50	4	40	67	64	55	77	67	63	63	77	76	79	79	2748	59	40	35	42	33	39	36	32	35
IM_MO	5	11	1	22	9	13	1	11	21	21	11	34	13	16	12	33	33	35	16	53	6063	47	36	45	47	39	41	38	39
IM_B	4	15	4	23	8	11	0	10	26	22	15	28	18	8	14	21	32	22	12	10	1	5446	32	36	37	39	37	34	38
LV	14	52	7	35	10	25	2	16	44	41	34	52	37	33	30	46	43	46	62	38	7	5	3818	59	37	39	38	35	36
LV2	22	72	21	61	16	35	8	32	60	64	58	73	48	50	44	72	74	75	76	38	9	7	57	1169	20	23	22	22	22
BR	12	85	8	33	8	43	4	27	69	67	40	50	58	53	53	42	46	51	93	41	4	4	26	19	209	6	6	8	6
BR2_Cer	33	73	26	45	24	67	9	29	79	73	53	56	67	78	56	60	57	74	78	56	10	5	31	41	83	315	28	28	25
BR2_FC	22	96	22	46	26	67	7	29	88	78	52	57	69	80	63	67	58	77	76	62	10	7	44	53	85	86	329	36	28
BR2_TC	29	76	12	46	19	60	6	29	81	77	51	64	67	78	58	61	59	69	75	57	7	5	35	48	69	82	91	385	29
BR2_P	32	82	16	44	22	47	6	54	77	52	56	56	50	69	60	61	53	66	78	54	5	5	38	58	69	88	94	91	274

The diagonal gives the number of unique exGenes found in each study. The numbers in the upper triangle are the percentage of common exGenes between each pair of studies. The numbers in the lower triangle are the percentage of common exGenes associated with the same or LD-linked exSNPs between each pair of studies.

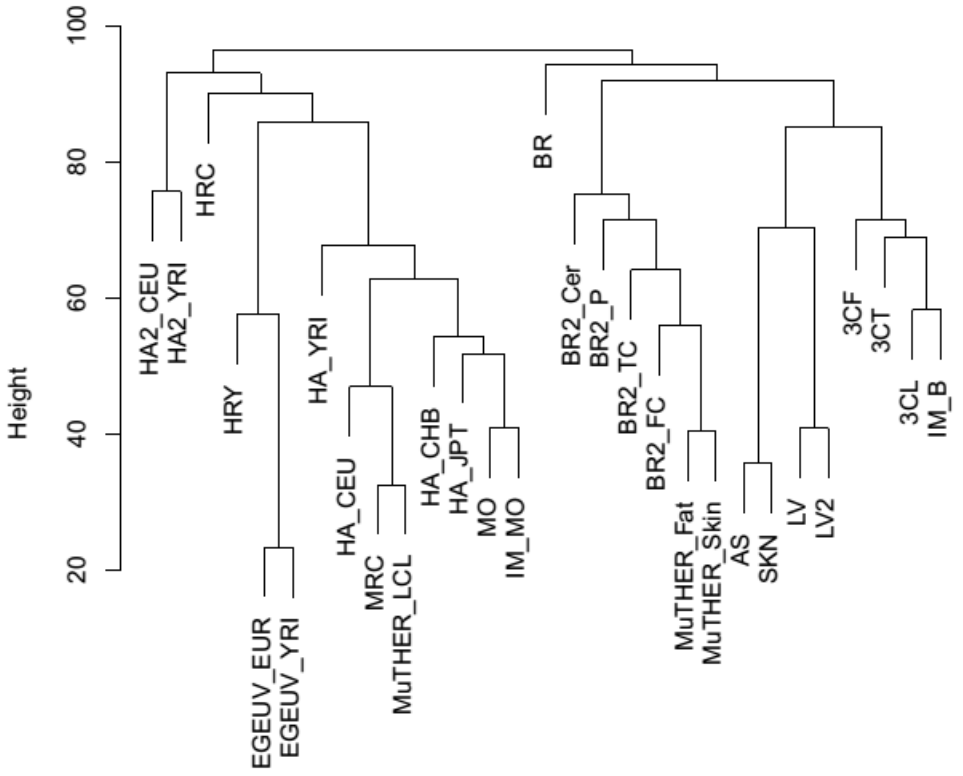


Figure 2.3. Hierarchical clustering of the fraction of common exGenes between pairs of eQTL datasets.

Distance scale is based on the % of common exgenes between pairs of datasets.

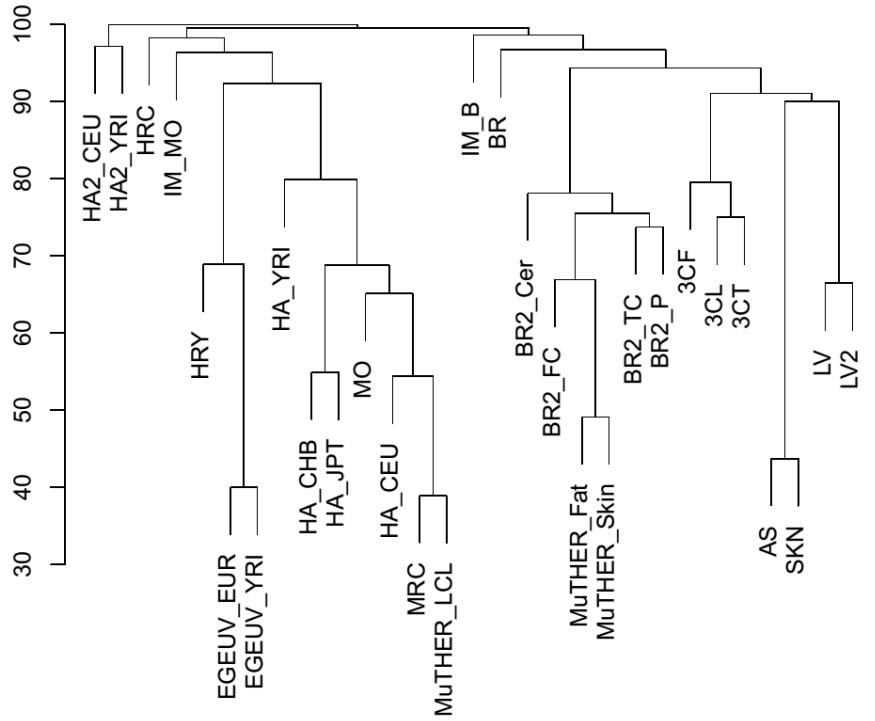


Figure 2.4. Hierarchical clustering of the fraction of common exGenes with LD related exSNPs between pairs of eQTL datasets.

Distance scale is based on the % of common exgenes with LD related exSNPs between pairs of datasets.

High-confidence eQTL identification

Given the high level of the variability between the studies apparently due to non-biological causes, such as different microarrays and inherent noise in the data, it is desirable to identify the more reliable eQTL relationships. For this purpose, we compiled eQTL relationships that have been observed in at least two studies, for set of studies within the same population, for studies on the same cell types, and between any pair of studies, independent of population and cell type.

For this purpose, the studies were grouped into 13 subsets (Table 2.4). The largest integrated set, AllCell_AllPop, includes all 29 datasets. The LCL_CEU, LCL_YRI, and LCL_ASN, and LCL sets integrate datasets performed in LCLs for Caucasian, African, Asian, and all populations, respectively (the Chinese CHB and Japanese JPT populations were combined into one Asian set (ASN)). Datasets for each of the other eight cell types were merged, independent of population (most are in fact Caucasian).

Table 2.4. Classification of eQTL studies by cell type and population

Class	Tissue	Population	Studies	Datasets
AllCell_AllPop	All	All	16	All
LCL_CEU	LCL	CEU	8	HA_CEU, HA2_CEU, HRC, AS, MRC, 3CL, EGEUV_EUR, MuTHER_LCL
LCL_YRI	LCL	YRI	4	HA_YRI, HA2_YRI, HRY, EGEUV_YRI
LCL_ASN	LCL	ASN	1	HA_CHB, HA_JPT
LCL	LCL	All	9	HA_CEU, HA_YRI, HA_CHB, HA_JPT, HA2_CEU, HA2_YRI, HRC, HRY, AS, MRC, 3CL, EGEUV_EUR, EGEUV_YRI, MuTHER_LCL
Bcell	B-cell	All	1	IM_B
Monocyte	Monocyte	All	2	MO, IM_MO
Tcell	T-cell	All	1	3CT
Brain	Brain	All	2	BR, BR2_Cer, BR2_FC, BR2_TC, BR2_P
Liver	Liver	All	2	LV, LV2
Skin	Skin	All	2	SKN, MuTHER_Skin
Fibroblast	Fibroblast	All	1	3CF
Fat	Adipose	All	1	MuTHER_Fat

ASN indicates pooled CHB+JPT populations.

To identify the common eQTL associations in the integrated set, the same algorithm was used as for finding agreements between pairs of studies. A high-confidence

eQTL relationship is defined as one for which supporting eQTLs are found in more than one study within an integrated set. Figure 2.5 illustrates how supporting eQTLs are identified. The number of studies in which supporting eQTLs are found is used as basis for an approximate confidence score. The idea here is that the more studies with data supporting the same eQTL relationship, the higher its reliability. We identified high-confidence unique eQTL relationships within the eight integrated sets that contain more than one study. Table 2.5 shows the number of unique eQTL relationships and high-confidence unique eQTL relationships in each integrated set at various LD levels. For the biggest integrated set, AllCell_AllPop, at the lowest LD threshold ($r^2 \geq 0.3$), the 133,658 unique eQTL relationships in this set resulted in 6,754 high-confidence unique eQTL relationships involving a total of 4,709 exGenes (HC-exGgenes). There are more high-confidence unique eQTL relationships in the LCL integrated set than others, as a consequence of the larger number of contributing studies. The comparison between one pair of studies, BR and BR2, shows the lowest agreement, and so there are the fewest high-confidence unique eQTL relationships in the Brain integrated set, consisting of only 16 HC-exGenes.

In general, most exGenes (77%) contain only one high-confidence unique eQTL relationship in each integrated set at the lowest LD level ($r^2 \geq 0.3$) (Figure 2.6). As the LD level (r^2 threshold) increases, the LD haplotype block is broken out to multiple blocks. Therefore, the number of high confidence exGenes that contain more than one high-confidence eQTL relationship will also increase. We assumed that it is likely that in fact most exGenes are only involved in only one eQTL relationship - so

performed most of the analysis in terms of the number of exGenes, rather than relationships, to avoid double counting. Figure 2.7 shows the distribution of the number of studies in which each high-confidence exGene is identified, at various LD levels, for the AllCell_AllPop integrated set. Most (17-18%) HC-exGenes appear in four studies and 63-69% of HC-exGenes appear in fewer than five studies.

As an estimate of the relative of quality of the eQTL datasets, we calculated the fraction of exGenes in each dataset that are part of high-confidence unique eQTL relationships (HC-exGenes) in the LCL_CEU integrated set (Figure 2.8). This quality measure varies widely. The lowest fraction of HC-exGenes is for the HA2 dataset (6.5%). The MRC dataset has the highest fraction of HC-exGenes (84%).

We compared the number of high confidence eQTL relationships identified with that expected by chance. For this purpose, we first generated 1000 random sets of pseudo eQTL relationship data for each of the 11 selected eQTL datasets. For each these 1000 full sets of pseudo data, we then used the algorithm described above to identify implied high-confidence eQTL relationships and also calculated the number of pseudo high-confidence eQTL genes so generated. In all cases the number of pseudo eQTL relationships is much much lower than that for the real data (typically more than a factor or 10), and the probability of the real data occurring by chance is too low to calculate. Thus, according to this model, the high-confidence eQTL genes in each integrated set represent significant agreement between datasets, way beyond what would be expected by chance. Fuller details are given in the Appendix.

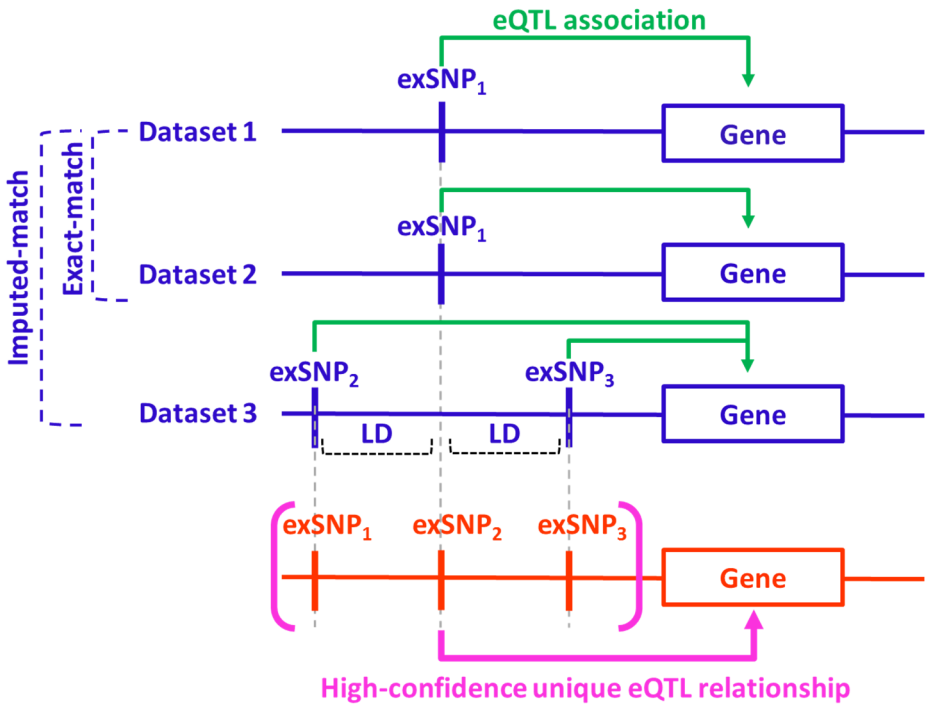


Figure 2.5. Identification of high confidence unique eQTL relationships.

A high confidence eQTL relationship is defined as one found in two or more datasets. This figure illustrates two ways, either exact-match or imputed-match, to determine consensus associations. Exact-match: In Dataset 1, the presence of exSNP1 is associated with altered expression of the gene. Dataset 2 contains the exact same SNP – gene association, sufficient to classify the association as high confidence. Imputed-match: Dataset 3 has an association between two other SNPs, exSNP2 and exSNP3 and the expression level of the same gene. These SNPs are both in LD with exSNP1, so are considered to represent the same underlying relationship. As seen in datasets 1 and 2.

Table 2.5. Summary of eQTL relationships and high-confidence eQTL relationships found in each integrated set

'LD(r^2)' is the linkage disequilibrium threshold used for relating exSNPs. See text for dataset definitions.

	LD(r^2)	AllCell_AllPop	LCL_CEU	LCL_YRI	LCL	Brain	Liver	Skin	Monocyte
Unique exSNPs		548344	431758	23536	441971	9169	3588	84578	55943
Unique exGenes		15170	7869	2725	8918	1171	4295	2616	7186
Total unique eQTL relationships	0.8	240785	190902	17110	192093	1939	4914	14419	53656
HC unique eQTL relationships	0.8	18615	9585	257	9237	21	393	153	3562
HC unique exGenes	0.8	4252	2079	203	2245	16	393	91	857
Total unique eQTL relationships	0.5	169031	140831	14688	140481	1468	4803	6597	43065
HC unique eQTL relationships	0.5	9506	4249	229	4210	16	485	93	3285
HC unique exGenes	0.5	4482	2161	222	2345	16	485	93	899
Total unique eQTL relationships	0.3	133658	114892	12892	113989	1390	4750	3999	34118
HC unique eQTL relationships	0.3	6754	3032	238	3174	16	530	94	2741
HC unique exGenes	0.3	4709	2231	237	2438	16	530	94	943

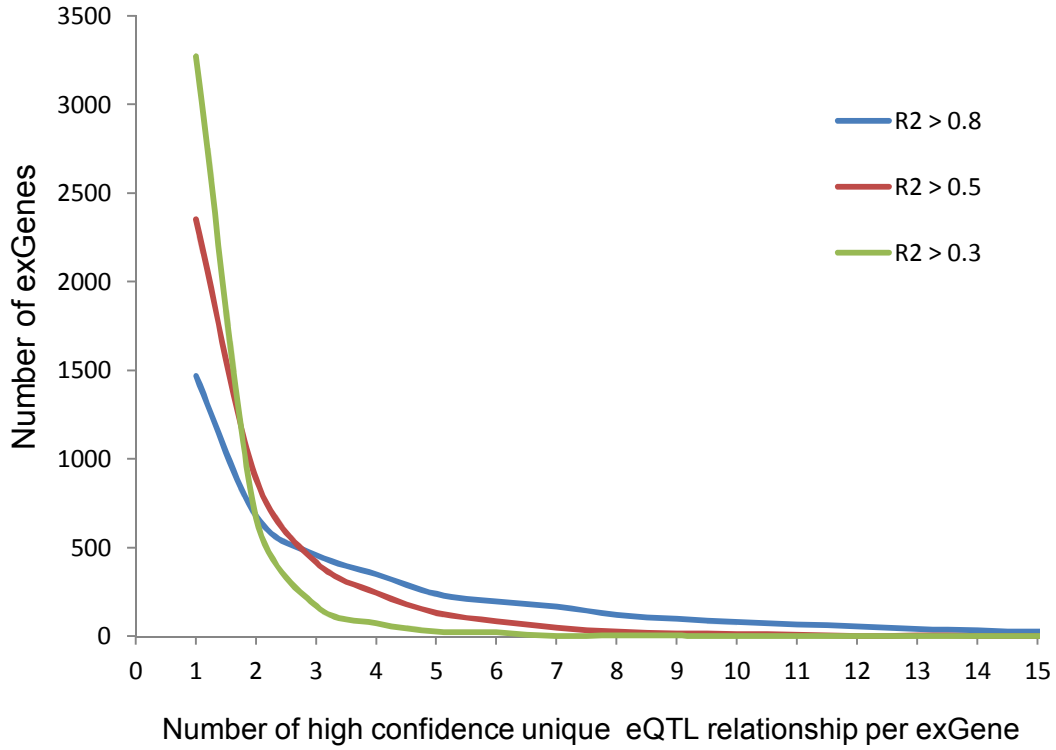


Figure 2.6. Number of high-confidence exGenes with 1, 2, 3... eQTL relationships at various LD thresholds (r^2) in the AllCell_AllPop integrated set.

At all thresholds, most exGenes appear to be involved in a single relationship. The proportion of exGenes with only one high-confidence unique eQTL relationship is 34.5%, 55.4%, and 77% for $r^2 \geq 0.8$, 0.5, and 0.3, respectively.

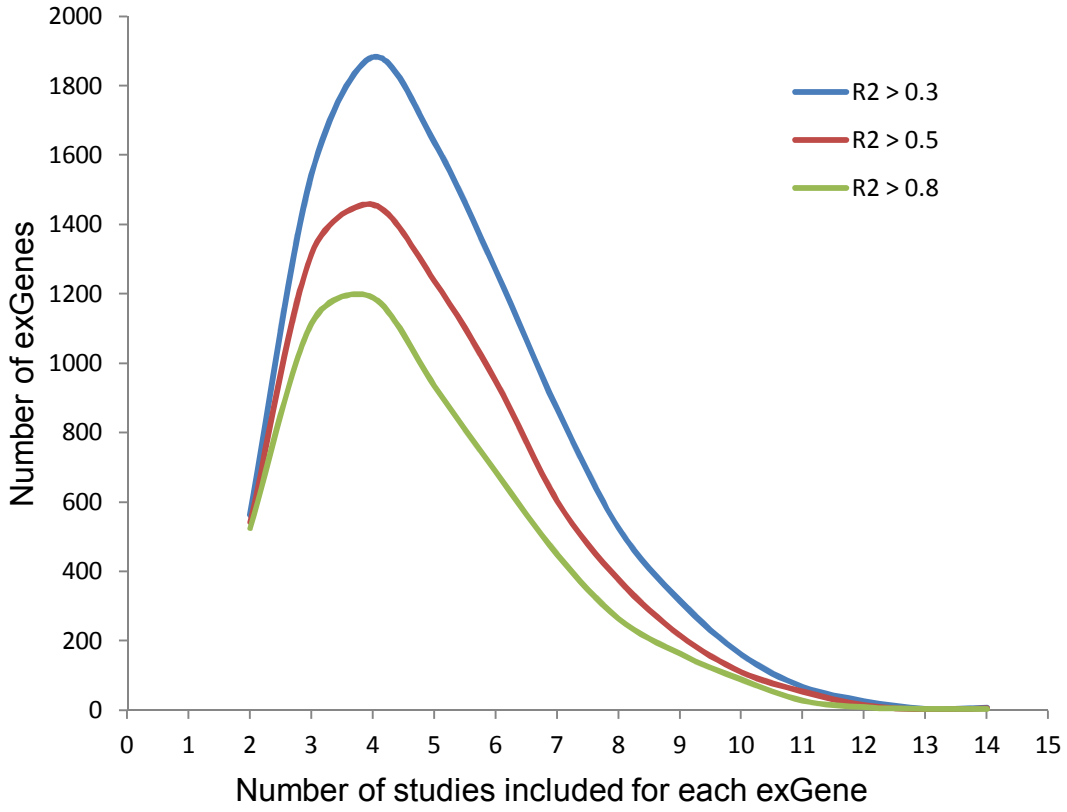


Figure 2.7. Number of HC-exGenes with support from 1, 2, 3, ... studies at various LD threshold (r^2) in the AllCell_AllPop integrated set.

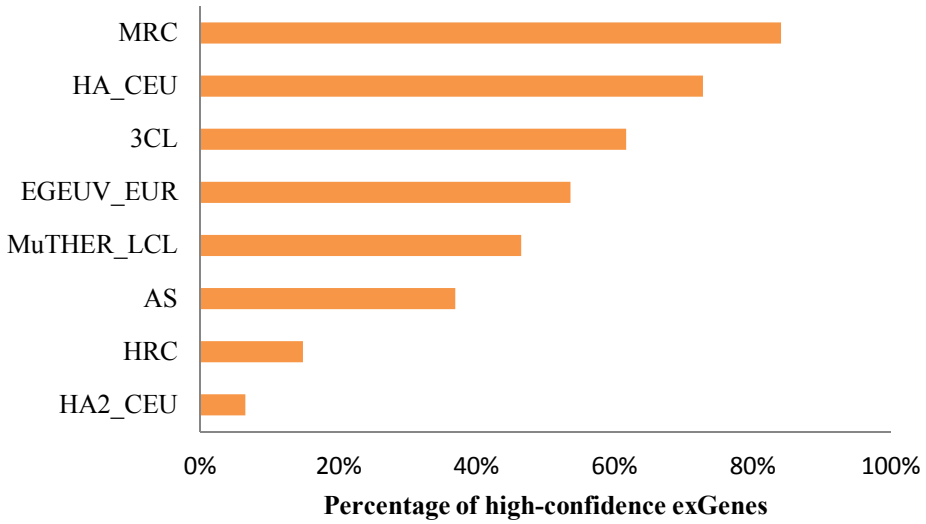


Figure 2.8. Approximate quality of each dataset, as reflected in the % of high-confidence exGenes relative to the LCL_CEU integrated set, at an $r^2 > 0.3$ LD threshold.

Tissue dependence of eQTL relationships

We made use of the data for different tissue types included in the 16 eQTL studies to perform limited testing on the extent to which eQTLs are conserved across tissue types. As noted earlier, only a fraction of eQTLs are found in multiple studies even when the same tissue and population have been used, so that simply looking at the fraction eQTLs common to studies in different tissues is not an adequate approach. To address this, we restricted the comparisons to situations where there are pairs of studies that share a tissue type, so providing a reference level of agreement, and that also have data on other tissues.

Two studies, each on LCLs and two other tissues, can each be used for this purpose: MuTHER with LCL, fat, and skin (Dimas et al., 2009), and 3C with LCL, fibroblast, and T-cell (Dimas et al., 2009). Both studies are in Caucasian populations, and so can be compared with the other LCL studies on that population. Figure 2.9 & 2.10 shows the fractions of common exGenes between pairs of datasets. Figure 2.9, shows the fraction of common exGenes between each of three MuTHER tissues and seven other studies conducted with LCL. The fraction of exGenes common to both LCL datasets varies widely, from 33-68%, reflecting the differing experimental and other factors discussed earlier. But in all seven of comparisons, the fraction of common exGenes is higher between LCL-LCL dataset pairs than for LCL to other tissue comparisons, indicating a level of tissue specificity. For the LCL-fat comparisons, the common exGene fraction is between 27 and 39% lower than for LCL-LCL, and for LCL - T-cell comparisons it is 29 -50% lower. Similar levels of tissue conservation were found within the 3C study. Figure 2.10 shows similar comparisons between the seven reference LCL sets and the LCL, fibroblast and T-cell data for the 3C study. Here the differences between cell types appear generally rather small: 16-32% fewer for LCL to fibroblast comparisons, and 15-27% less for LCL to T-cell comparisons.

These are very limited comparisons, but suggest that generally the level of conservation of eQTLs across tissues is fairly high, allowing extrapolation between tissue types, albeit at the expense of some false positives.

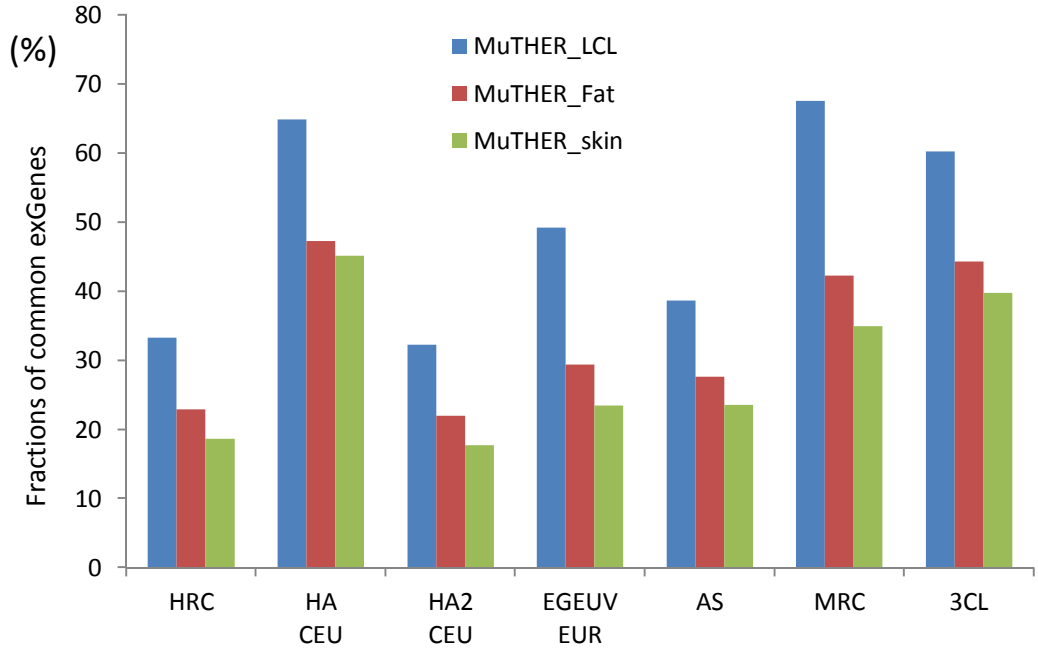


Figure 2.9. Comparisons of fractions of common exGenes between pairs of eQTL datasets of the same cell type and pairs with different cell types for the MuTHER study.

The blue bar shows the fractions of common exGenes between various LCL datasets and the MuTHER_LCL dataset. The red and green bars show the fractions of common exGenes between the other LCL datasets and the MuTHER_Fat and MuTHER_skin datasets, respectively.

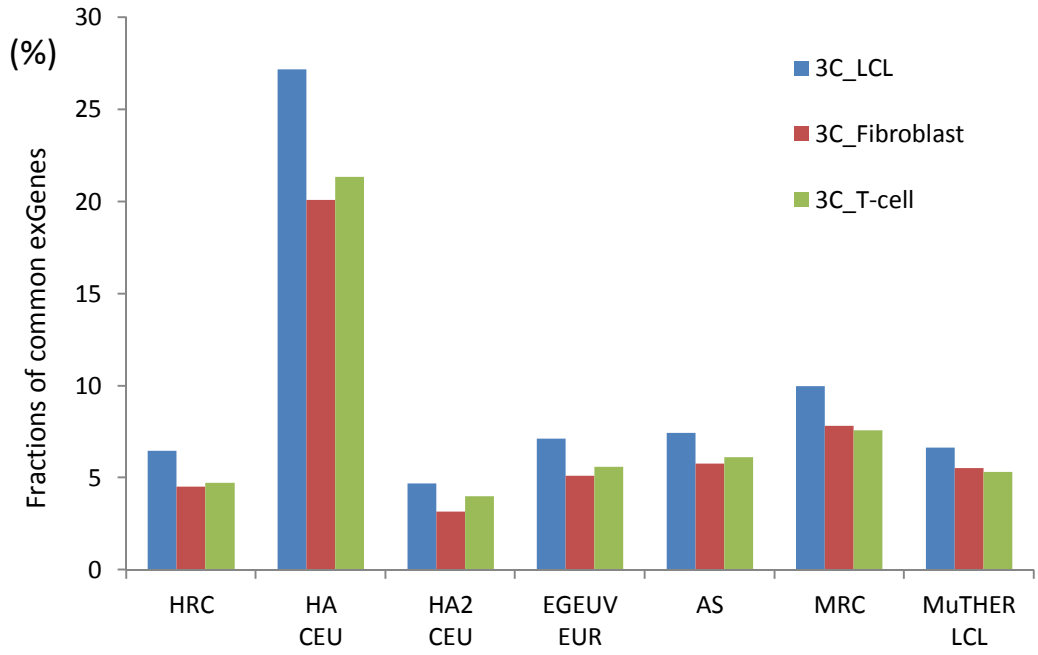


Figure 2.10. Comparisons of fractions of common exGenes between pairs of eQTL datasets of the same cell type and pairs with different cell types for the 3C study. The blue bar shows the fractions of common exGenes between the LCL datasets and the 3C_LCL dataset. The red and green bars show the fractions of common exGenes between the LCL datasets and the 3C_Fibroblast and 3C_T-cell datasets, respectively. In both sets of comparisons, there is evidence of limited tissue specificity.

Population dependence of eQTL relationships

Since there are a number of studies using LCLs in Caucasian and African populations, we can also examine the extent of population dependent eQTL relationships. Here we used the data from the HA study, conducted in these two populations, and compared the fraction of common exGenes between those datasets and those in other studies on Caucasian populations.

Figure 2.11 shows the fractions of common exGenes between seven datasets conducted in LCL in Caucasian populations and the Caucasian and African datasets in the HA study. Differences between within and across population fractions are usually small, with the exception of the 3C comparison, where the fraction of common exGenes is about 25% smaller for across the populations than within Caucasian. Although this is a limited comparison, it supports the view that the differences in eQTLs across populations are not large.

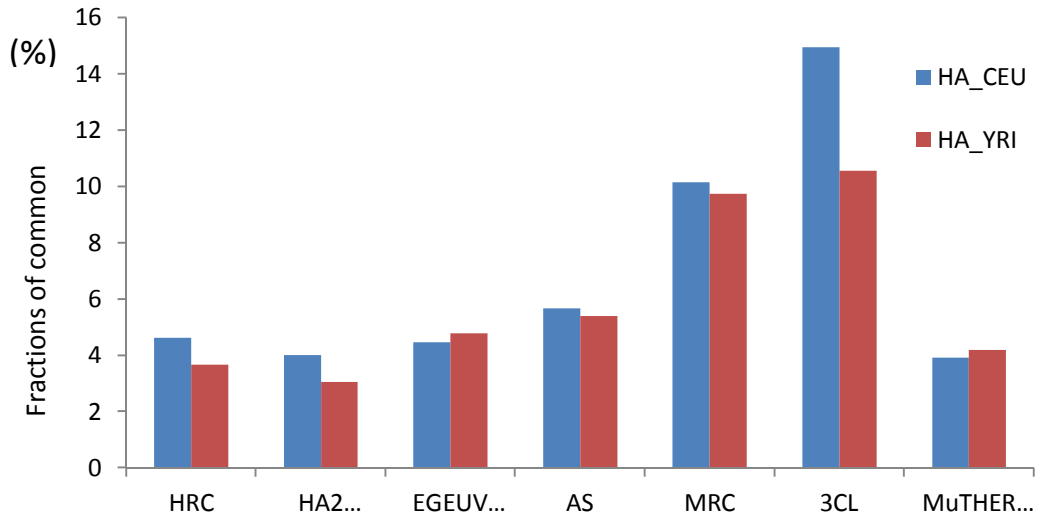


Figure 2.11. Comparisons of fractions of common exGenes between datasets in the same population versus datasets from different populations.

The blue bar shows the fractions of common exGenes between various Caucasian datasets in Caucasian data in the HA_CEU dataset. The red bars are the fractions of common exGenes between the other Caucasian datasets and HA_YRI dataset. The results indicate low population dependence of eQTLs.

Relationship to Functional elements

To explore the possible mechanisms underlying eQTL relationships, we examined the relationship between the locations of each exSNP and those of the following seven categories of functional element: microRNA target sites, transcription factor binding sites, DNaseI hypersensitivity regions, conservation sensitivity regions, programmed -1 ribosomal frameshift (-1 PRF) regions, and splicing sites. For each integrated eQTL set, we identified any exSNP that falls within the chromosomal position range of the known representatives for microRNA target sites, transcription factor binding sites, DNaseI hypersensitivity regions, and conservation sensitivity regions (Table 2.6). For PRF and splicing SNPs, we searched for exSNPs that are in LD with these functional SNPs.

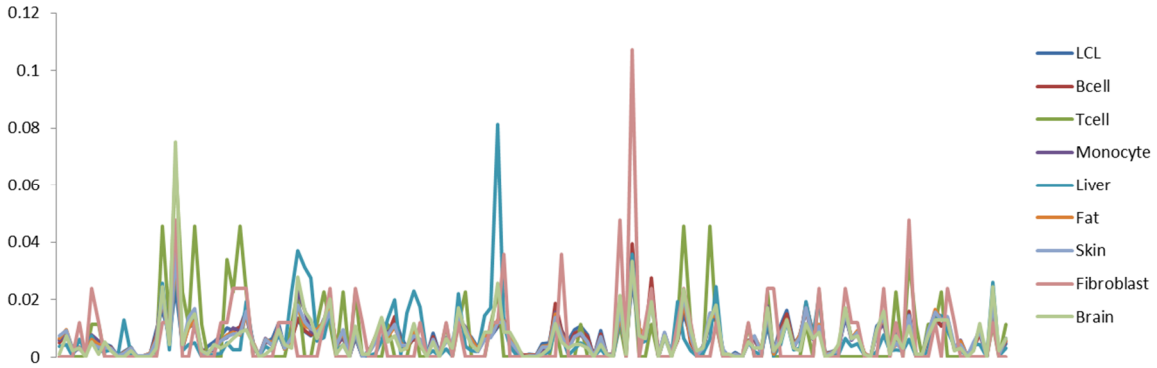
To assess potential tissue specific mechanisms, we then investigated the co-localization of tissue specific exSNPs and these functional elements and to see if there is any preference of functional elements for each tissue. In general counts for particular elements are small, so not statistically significant. But two transcription factor binding sites, Pol2 in Fibroblasts, and MafK (ab50322) in Liver, do appear much higher than background (Figure 2.12).

Table 2.6. Number of exSNPs that fall on each type of functional element, and number of associated exGenes, for each integrated eQTL set

	AllCell AllPop	LCL CEU	LCL YRI	LCL ASN	LCL	Bcell	Tcell	Monocyte	Liver	Fat	Skin	Fibroblast	Brain
miRNA target													
SNPs	6738	5169	388	126	5267	491	14	1149	119	1835	1587	13	163
exGenes	4077	2576	173	77	2630	408	14	964	137	1098	938	13	105
miRNAs	752	748	411	236	748	524	24	650	241	711	703	13	285
DNase													
SNPs	113290	85831	4467	1554	87763	5690	152	15238	1007	27041	21632	139	2207
exGenes	10870	5853	1107	294	6303	2610	141	4343	1425	2649	2189	132	563
DNases	118	118	118	112	118	118	46	118	110	118	118	53	113
TFBS													
SNPs	71243	56056	3558	1004	57351	3296	82	8482	591	15397	12317	88	1221
exGenes	9380	5300	830	254	5644	1829	81	3354	928	2359	1958	88	377
TFs	148	148	147	142	148	148	110	148	140	147	146	107	139
Sensitive													
SNPs	2412	1921	91	32	1958	112	4	286	21	499	393	3	42
exGenes	1797	1196	80	26	1243	103	4	274	28	429	317	3	29
Ultra-Sensitive													
SNPs	435	390	22	5	396	17	0	39	2	48	49	0	6
exGenes	219	141	7	3	146	17	0	37	3	46	36	0	7
prfDB													
SNPs	32	17	4	1	19	1	0	11	2	11	14	0	0
exGenes	70	33	5	1	36	1	0	14	3	14	17	0	0
Splicing													

SNPs	4741	3318	275	192	3456	459	100	1242	578	1910	1609	98	227
exGenes	4420	2410	204	116	2555	340	68	1032	615	1102	938	77	171

A



B

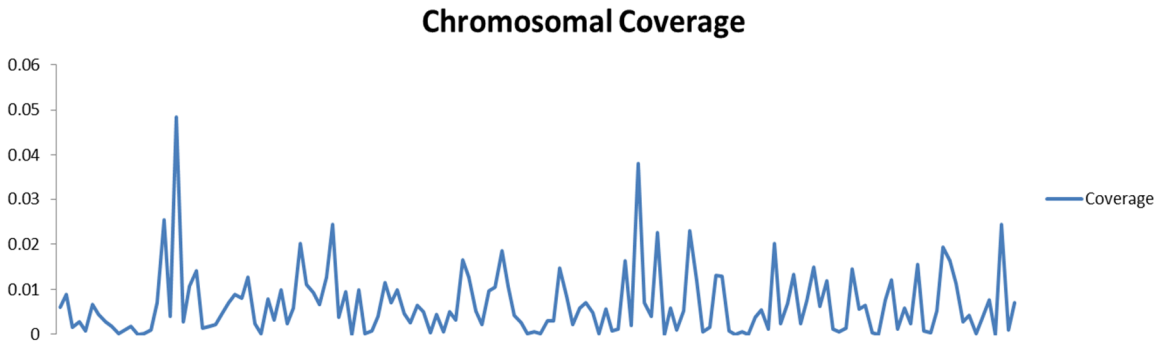


Figure 2.12. Distribution of tissue-specific exGenes associated transcription factor binding sites.

A. Fraction of exGenes for each tissue that are associated with the binding sites for each of 148 different transcription factors.Transcription factors are arranged sequentially along the X axis. Different colors represent different tissues (see key at right). Note the single large peaks for Liver (light blue) and Fibroblast (pink). B. Expected distribution of exGene coverage of each transcription factor, based on the fraction of bases in the genome included in sites for that factor. Transcription factors are arranged sequentially along the X axis.

2.3: Methods

Data sources

All eQTL association data in this study were collected from 16 publicly available studies that had been performed on various human tissues and populations. Table 2.1 lists the studies included. The statistical models and acceptance thresholds used are as follows:

HA study (Stranger et al., 2007) : Both significant cis-eQTLs, identified by a linear regression model or a Spearman rank correlation at a 0.001 permutation P-value threshold for each individual population, and significant trans-eQTLs, identified by linear regression model at a 0.001 permutation threshold per population, were included in this study.

BR study (Myers et al., 2007) : Both cis- and trans- eQTLs, identified using the PLINK analysis toolset, with a one-degree-of-freedom allelic test of association at an empirical P-value < 0.05 threshold, were included in this study.

AS study (Dixon et al., 2007) : Both cis- and trans-eQTLs with LOD > 6 (corresponding to a 'after Bonferroni correction' P-value threshold of ~ 1.2×10^{-7}) were selected as significant eQTL associations and included in this study.

LV study (Eric E Schadt et al., 2008) : Both significant cis- and trans-eQTLs, determined by Kruskal-Wallis test (Kruskal and Wallis, 1952), were included in this study.

HA2 study (Duan et al., 2008) : With the P-value of 2×10^{-8}, all significant TC-eQTL associations were observed in the CEU and the YRI samples, respectively. Both significant cis- and trans-eQTLs were included in this study.

3C study (Dimas et al., 2009) : Spearman rank correlation was used to test for cis-associations between SNP genotypes and probe expression levels for each tissue type. All significant cis-eQTLs for each of three cell types were included in this study.

MO study (Zeller et al., 2010) : The association tests were performed by "TAMU ANOVA" and further checked by a Kruskal-Wallis test. Both significant cis- and trans- eQTLs sets were included in this study.

HRC study (Montgomery et al., 2010) : All significant cis-eQTL at a 0.001 permutation P-value thresholds were included in this study.

HRY study (Pickrell et al., 2010) : Significant cis- eQTLs with genes or putative new exons at a FDR of 10% (corresponding to $P = 2.4 \times 10^{-5}$) were selected in this study.

BR2 study (Gibbs et al., 2010) : All significant cis-eQTLs for each of the four tissue regions were included in this study.

SKN study (Ding et al., 2010) : All significant cis-eQTLs with P-value threshold ($P < 9 \times 10^{-7}$) from normal human skin were included in this study.

LV2 study (Innocenti et al., 2011) : Both significant cis- and trans- eQTLs with Bayes Factor > 5 were collected in this study.

IM study (Fairfax et al., 2012) : Both cis-and trans-eQTLs at a permuted P-value threshold ($P<0.001$) from primary monocytes and B-cells were selected in this study.

MuTHER study (Grundberg et al., 2012; Nica et al., 2011) : All significant (<1% FDR) cis- eQTL association for three tissue types (Fat, LCL, and Skin) were included in this study.

MRC study (Liang et al., 2013) : Both significant (<5% FDR) cis-and trans-eQTLs from a meta-analysis of British children with asthma or atopic dermatitis were selected in this study.

E-GEUV study (Lappalainen et al., 2013) : All significant (below false discovery rate 5%) gene cis-eQTLs for EUR and YRI populations from the 1000 Genomes Project (Abecasis et al., 2010) were selected in this study.

Data preparation

To efficiently analyze significant exSNP-exGene association pairs between these studies, all transcript names, probe IDs, or alias gene names were converted to current unique Entrez Gene IDs and Gene names (NCBI build 37.2). Ambiguities in alias gene names were resoved using chromosome location information. Transcript clusters (TCs) identified in the HA2 study were converted to Entrez gene IDs by mapping the region of each TC to gene ranges on the human genome assembly hg19. In addition, retired and discontinued SNP IDs were filtered out and all SNP IDs were converted to the current dbSNP IDs (dbSNP build 134). Retired or unmappable gene names were eliminated from the study. Any SNP with multiple chromosome coordinates on NCBI reference assembly 37.2 (dbSNP b134) were removed from each dataset.

Linkage disequilibrium

Linkage Disequilibrium (LD) information between pairs of SNPs was acquired from the HapMap project phase III (release 27) (The International HapMap 3 Consortium, 2010) or derived from 1000 Genomes Project (phase1 release) (The 1000 Genomes Project Consortium, 2010) for several ethnic populations (CEU, YRI, CHB, and JPT

for Hapmap; EUR and YRI for YRI). For 1000 Genomes LD data, the r^2 values for pairs of SNPs with MAFs > 5% and located within 200,000 bp of each other were calculated using PLINK (v. 1.07) (Purcell et al., 2007). Figure 2.13 shows the distribution of differences in LD value obtained from the two sources. The spearmen correlation between LD values from Hapmap project and 1000 Genomes is 0.89.

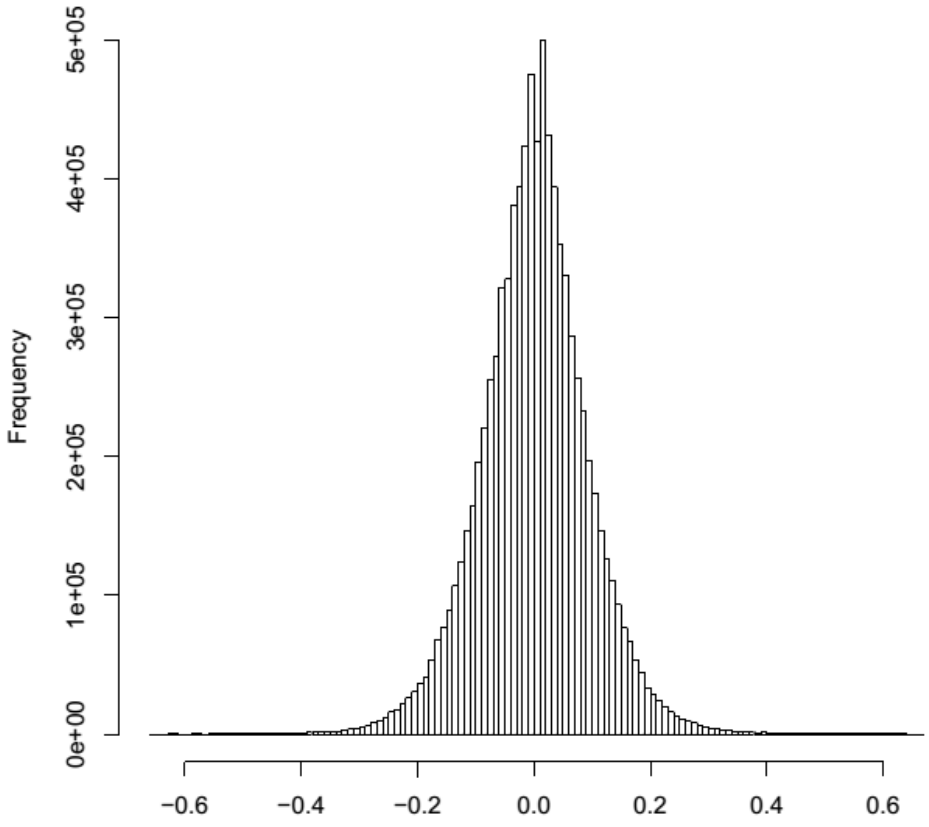


Figure 2.13. Histogram of differences in LD values for pairs of SNPs derived from Hapmap and 1000 Genomes data.

Where both Hapmap and 1000 Genomes provided LD values for a SNP pair, the Hapmap value was used. Where possible, appropriate populations were used for obtaining LD values. HA_CEU, HA2_CEU, HRC, AS, BR, LV, 3C, BR, and BR2 datasets are from Caucasian (CEU) populations and HA_YRI, HA2_YRI, and HRY

datasets are from Yoruba (YRI) populations. HA_CHB and HA_JPT datasets are Chinese (CHB) and Japanese (JPT) population, respectively. No clear ethnic identity is available for the MO and LV2 sets. For the LV2 dataset, individuals are mostly from the mixture of Caucasian and African populations. Therefore, we generated an intersection LD set occurring in both CEU and YRI populations. For the MO study, we generated an intersection of LD set among all four population CEU, CHB, JPT, and YRI populations.

Hierarchical clustering

The distance between each pair of datasets was defined as (1-f), where f is fraction of common exGenes between the two sets. The hclust module in R was used.

Functional elements

Date from several publicly available databases of annotated functional regions was used to identify eQTLs that fall on known transcriptional regulatory sites. MicroRNA gene regions were acquired from NCBI refGene. Data from Targetscan (5.1) (Grimson et al., 2007), microcosm from miRBase (v5) (Kozomara & Griffiths-Jones, 2011), and microRNA.org (Aug 2010 release) (Betel, Wilson, Gabow, Marks, & Sander, 2008) were combined for the predicted microRNA binding sites. TargetScan predicts biological targets of miRNAs by searching for the presence of conserved 8mer and 7mer sites that match the seed region of each miRNA. Other databases, microcosm and microRNA, computationally predict targets for microRNAs across many species by several methods, for example the degree of complementarity to the

miRNA, and conservation across multiple species. Transcription Factor binding sites and DNaseI hypersensitivity regions were downloaded from the Integrated regulation tracks of ENCODE Project (Dunham et al., 2012). Conservation sensitive and ultra-sensitive sites were collected from the 1000 Genome Project (Phase 1) (Khurana et al., 2013). Potential programmed -1 ribosomal frameshift (-1 PRF) regions were collected from PRFdb (Plant, Wang, Jacobs, & Dinman, 2004). PRFdb used multiple algorithms to identify potential -1 PRF signals as defined by a heptameric slippery site followed by an mRNA pseudoknot in eukaryotic genes or sequences of interest. SNPs on potential splicing site were collected from SplicePort (Dogan, Getoor, Wilbur, & Mount, 2007). SplicePort implements a feature generation algorithm for the classification of potential splice sites, scoring each GT or AG dinucleotide using features within a window of 162 nucleotides (80 nt. on either side of any splice site region) to identify deleterious effects of genetic variation on splicing. The chromosome coordinates of all binding sites were converted to hg19 assembly and all eQTLs that locate on these binding sites were identified.

2.4: Discussion

There have now been a number of high-throughput studies for finding eQTLs in human populations and tissues, providing a wealth of data about the relationship between genetic variation and the level of gene expression. At present, though, reproducibility between studies is low (Dixon et al., 2007; Göring et al., 2007; Myers et al., 2007; Stranger et al., 2007; Veyrieras et al., 2008). We are interested in obtaining a conservative but relatively reliable set of eQTLs for use in other

applications, particularly identifying those human complex disease loci where a genetic variant affecting expression of a gene may be contributing to disease susceptibility. To this end, we compared the results of 16 independent eQTL studies, to find those variant/expression relationships that have been observed more than once. Across the 16 studies considered, more than 15,000 different genes have been reported as involved in an eQTL relationship, usually with a nearby (cis) variant. The number of human genes that are expressed at a high enough level for eQTL associations to be detected is probably not much larger than this, so at face value, almost every human gene has its expression affected by at least one variant. This remarkable observation may be misleading however - only a little over a quarter of these genes have been found to be involved in the same eQTL more than once, across the included studies. Most commonly, each gene is found to involve in a single eQTL relationship.

The assumption that consensus eQTLs are more reliable than those only observed once requires statistical independence of each study. Each of the studies was performed by different investigators, and in general, different genotyping and transcription profiling technologies were used. Additionally, a third factor affecting reliability, the statistical analysis technique used, varies across studies. Thus, the condition of independence between studies is largely met.

The inclusion of studies with data derived from different tissue types allowed us to estimate the extent to which eQTLs are conserved. The data are limited, and the

presence of large amounts of noise also restricts analysis, but nevertheless, the available comparisons suggest a substantial amount, larger than 50%, of at least partially tissue independent eQTL. It should be noted that although a study may be tissue specific, that tissue will often include a range of cell types. For instance, in eQTL studies of brain tissue (Gibbs et al., 2010; Myers et al., 2007) various types of cells are included, such as blood cells, subtypes of neuronal cells, and different glial cells, so it not possible to distinguish the eQTL relationships for each specific cell type. We also examined the limited data on conservation of eQTLs between Caucasian and African populations. Here the degree of population independence appears higher than across tissues.

We have used the consensus eQTL results to generate several integrated datasets for use in other applications. In Chapter 3, we describe the use of one of these for analysis of complex disease loci.

Chapter 3: The role of Human expression quantitative Traits in complex trait disease

3.1: Introduction

A main challenge in interpreting personal genomes is to identify the causal variants underlying human complex traits and their functional consequences. In the past decade, Genome-wide association (GWA) studies have successfully identified thousands of genetic variants associated with numerous human complex traits, including diseases. So far, the GWAS catalog of the National Human Genome Research Institute lists ~12500 single nucleotide polymorphisms (SNPs) associated with one of more complex traits, gathered from ~1800 GWA studies (www.genome.gov/gwastudies/). Each of these disease associated loci must harbor some underlying mechanism whereby the presence of a causal variant alters some molecular level process, and in turn, that perturbation affects higher level processes and pathways. A number of different mechanisms may be involved, including altered protein folding, half-life and function through missense SNPs (Sunyaev, Ramensky, & Bork, 2000; Z Wang & Moult, 2001), SNPs that affect splicing (G.-S. Wang & Cooper, 2007), and SNPs affecting RNA expression level (Nicolae et al., 2010). The majority of disease-associated SNPs are located in non-coding intergenic or intronic regions of the genome, including promoter regions, enhancers, or non-coding RNA genes (Hindorff et al., 2009; Ricaño-Ponce & Wijmenga, 2013), but generally there is little direct evidence on how these variants affect molecular level processes. One major source of difficulty in identifying mechanism is that genetic variants in a locus

found to be associated with disease (the markers) are a small part of a larger set, all in linkage disequilibrium (LD) with each other, and any one of these might be causal.

Genome-wide association studies have also been used to discover expression quantitative trait loci (eQTLs), by finding correlations between transcript expression levels and the presence of genetic variants (Jansen and Nap 2001). The recent emergence of high-throughput technologies, particularly transcription microarrays and RNA-sequencing, provide an efficient way to simultaneously measure the expression levels of thousands of genes. Microarray technology has also been used for large scale genotyping, and comparison of these two types of data then allows eQTL mapping in a large number of individuals (Lappalainen et al., 2013; Liang et al., 2013; Montgomery et al., 2010). Initially, data derived from Epstein-Barr virus transformed immortalized lyphobalstoid cell lines (LCLs) were used for population-wide eQTL analysis in humans (Dixon et al., 2007; Duan et al., 2008; Stranger et al., 2007). Recently, a number of studies have performed eQTL mapping on various human tissues, such as brain (Gibbs et al., 2010; Myers et al., 2007), liver (Greenawalt et al., 2011; Innocenti et al., 2011; Eric E Schadt et al., 2008), adipose (Emilsson et al., 2008; Greenawalt et al., 2011; Nica et al., 2011), fibroblasts (Dimas et al., 2009), and skin (Ding et al., 2010; Grundberg et al., 2012; Nica et al., 2011). So far, thousands of cis- and trans- regulatory eQTLs have been discovered in a variety of human tissues and populations.

A number of studies have combined information from eQTL association results and disease GWAS findings to improve the functional interpretation of disease associated loci (Chu et al., 2011; Ertekin-Taner, 2011; Heid et al., 2010; Hrdlickova et al., 2011; Hsu et al., 2010; Lango Allen et al., 2010; Moffatt et al., 2007; Richards et al., 2012; Speliotes et al., 2010; Wu et al., 2012). Several studies have shown that SNPs associated with human traits and chemotherapeutic drug susceptibility are in general enriched for eQTLs (Cookson et al., 2009; Gamazon, Huang, et al., 2010; Nicolae et al., 2010). Although most studies have used eQTL data from the most accessible cell type, LCL, it is not clear how good a proxy these are for human cells and tissues relevant to non-immune related disease, such as psychiatric traits or cancers (Choy et al., 2008; Nicolae et al., 2010). Some studies have used eQTL results from tissues partially appropriate to the disease of interest when linking to disease-associated SNPs (Ding et al., 2010; Fransen et al., 2010; Innocenti et al., 2011; Kang, Morgan, et al., 2012; Kang, Yang, Chen, & Zhang, 2012; Liu, 2011; Maranville et al., 2011; Eric E Schadt et al., 2008; Zhong et al., 2010). For example, Ding et al. (Ding et al., 2010) reported an eQTL study of human skin that aimed to elucidate the role of regulation of gene expression in psoriasis. Richards et al. (Richards et al., 2012) assigned eQTL status to schizophrenia susceptibility alleles based on eQTL data derived from adult human brain (Myers et al., 2007).

A genetic variant may affect the expression level of a gene in a number of different ways, for example altering the affinity of a transcription factor to its cognate DNA binding site; altering affinity of a microRNA, or other factors that affect message

half-life; altering the relative propensity of different splicing isoforms, sometimes leading to nonsense mediated decay (Brogna & Wen, 2009; Maquat, 2004) and altering chromosome structure or other aspects of epigenetic control. In turn, altered expression may lead to altered disease susceptibility in variety of ways. Whatever the underlying mechanism, any SNP that alters expression sufficiently, as well as other SNPs in LD with it, will be detectable by an eQTL GWAS experiment. Thus, it is in principle possible to find which disease associated loci harbor an underlying expression mechanism by comparing the set of markers from a disease GWAS with the set of markers from an eQTL study: if the cause of disease risk is a change in expression discovered in an eQTL, the two sets of markers should be identical.

In practice, a number of factors complicate relating disease GWAS and eQTL results. There is substantial noise in GWAS and eQTL measurements, so that some SNPs that should be markers will be not identified, and some that are considered markers may be false positives, and P-values for both types of association may not be reliable. Because of sparse sampling of SNPs on current microarrays (typically only about one million of the approximately 40 million common SNPs are assayed), it is unlikely a causal variant will be directly assayed for association in either the disease or eQTL studies, and only a few markers will be detected, making comparisons of the two marker distributions difficult, especially if different genotyping microarrays have been used. Imputation methods (Howie, Donnelly, & Marchini, 2009) may be used to obtain estimated association P values for many SNPs not directly measured, solving this latter issue. Imputation requires the full genotyping data for each study

participant. In principle it is possible to obtain these full data for any study, but in practice, the need to deal with the human subjects issues for the disease data makes this effectively impossible for an analysis that makes use of many data sets. Also, available data from expression association studies often do not provide P-values, only information on whether the P value for each genotyped SNP is above or below a threshold. In order to address these data issues, we made use of one set of disease GWAS data with complete genotype information to investigate the properties of full marker distributions, and on that basis devised a method that can be applied to cases where only microarray marker SNP information is available.

A further complication in relating eQTLs to disease GWAS is the apparent unreliability of individual eQTL studies, arising from a variety of issues in statistical analysis as well as experimental factors. So far, most eQTLs have not been reproducible in multiple studies, even within studies conducted on the same cell types in the same population (Dixon et al., 2007; Göring et al., 2007; Myers et al., 2007; Stranger et al., 2007; Veyrieras et al., 2008). In Chapter 2, we described integration of human genome-wide eQTL data from 16 publicly available studies to identify higher confidence eQTL relationships on the basis of consensus, both generally and within several specific cell types.

In this study, we sought to identify which loci associated with complex trait disease may harbor an underlying expression mechanism, making use of the consensus eQTLs. To this end, we examined each of a set of disease-associated loci to ascertain

whether any known eQTL relationship may have produced the disease association data.

3.2: Results

High-confident eQTLs

In Chapter 2, we integrated 16 publicly available human genome-wide eQTL studies and identified those eQTL relationships that are found in two or more independent studies, making use of linkage disequilibrium (LD) relationships. These consensus eQTLs are regarded as relatively high-confidence (HC) eQTL associations. Several such sets were built. The AllCell_AllPop set contains the high-confidence eQTL associations derived from comparisons across all 16 eQTL studies, and is used here. The number of consensus relationships discovered depends on the threshold for linkage disequilibrium used. Table 3.1 shows a summary of high-confident eQTLs at the most conservative LD level ($r^2 \geq 0.8$).

Table 3.1 Summary of high-confident eQTLs from 16 integrated sets

Integrated set	AllCell_AllPop
Unique exGenes	15,170
Unique eQTL relationships	240,785
HC unique exGenes	4,252
HC unique eQTL relationships	18,615

'Unique eQTL relationships' are the number of associations found in at least one included study. 'Unique exGenes' are the number of genes participating on one or more eQTL relationships in at least one included study. 'HC unique eQTL

relationships' and 'HC exGenes' are those found to be involved in the same eQTL relationship in at least two included studies.

Identification of Disease loci with a possible expression related mechanism

To investigate the role expression regulation plays in disease susceptibility, we compared results from disease GWA studies and those from eQTL GWA studies. For each identified disease risk locus in a set of common diseases, we estimated whether there is an eQTL consistent with an underlying expression mechanism driving altered disease risk. We assume that in each disease risk locus, an underlying causal/mechanism variant affects disease risk. Because of linkage disequilibrium, that usually result in a set of SNPs (marker SNPs), including the causal one if that is a SNP, occurring at a different frequency in disease populations than in control populations, and so being detectable in GWA studies. If the disease causal variant affects the expression level of a gene, there should also be a set of overlapping marker SNPs discovered in eQTL studies. Thus, comparison of the location of disease markers and of nearby eQTL markers in a locus provides a means of estimating whether a known eQTL relationship provides a possible basis for the disease mechanism. The procedure for comparing disease and eQTL markers in described in Methods.

The diseases analyzed are Bipolar disorder (BD), Coronary artery disease (CAD), Crohn's disease (CD), Hypertension (HT), Rheumatoid arthritis (RA), Type 1 diabetes (T1D), and Type 2 diabetes (T2D)). 21 disease risk associated loci reported

in the seminal WTCCC1 GWA study of these diseases and a further 316 risk loci from meta-analyses and subsequent studies, extracted from the GWAS catalog (www.genome.gov/gwastudies/), were included.

For each disease-associated locus in each set, we collected all disease marker SNPs and all neighboring marker SNPs involved in high confidence eQTLs that are within 200,000 bps distance of any disease marker. The centiMorgan (cM) distance between each disease marker and each eQTL marker SNP was estimated using the Caucasian HapMap genetic map (A distance of 1 cM between locations corresponds to a recombination frequency of 1% per generation, and provides the measure of genetic linkage).

Figure 3.1 shows the percentage of loci for each disease type where disease markers match high confidence eQTL markers from the AllCell_AllPop set, as a function of cM threshold. The number of loci included raises steeply at low cM values, but less steeply above 0.005 cM. The steep slope at low values is likely a consequence of different tag SNPs used on the microarray chips for disease and expression association studies – often the exact disease marker SNP is not present on the expression chip, but there is one very close in cM space. Above 0.05 cM, the curves begin to plateau, but extra loci do continue to be added as the distance increases. Saturation of the number of loci covered is between 45 -73%, depending on the disease.

Matches between disease and eQTL markers were collected for three thresholds, cM distances of zero, less than 0.005, and less than 0.05, based on the analysis described in Methods. Table 3.2 shows the number of disease loci that meet these criteria. 15%-32% of the disease risk loci for each disease have putative expression mechanisms, based on the 0 cM threshold, and that increases to 23%-52% at a threshold of 0.005 cM, and 29%-61% at a 0.05 cM threshold. There is considerable variation in the fraction of putative expression loci across the seven diseases, with Type 2 diabetes having the lowest values (31% at the 0.05 threshold), and Rheumatoid arthritis and Crohn's disease having the highest (62% and 57% respectively at the 0.05 threshold). Appendix Table S5. shows all candidate expression loci for the seven diseases at a cM threshold 0.05 and the eQTL-associated genes for each locus. Each of these genes is a candidate for involvement in disease mechanism, based on the eQTL data.

Table 3.2. Number of disease risk loci with possible underlying expression mechanisms in seven common diseases.

Data at three thresholds of agreement between disease and expression markers are included – where at least one disease and expression SNP are identical (0 cM), where a disease and expression marker are less than 0.005 cM apart, and where the markers are less than 0.05 cM apart.

Disease set	BD	CAD	CD	HT	RA	T1D	T2D
Loci included	65	45	84	17	34	50	42
0 cM	13	8	24	3	12	12	7
0.005 cM	20	15	37	6	18	19	10
0.05 cM	26	23	48	8	21	24	13

To place these results in the context of previous studies, we defined three categories of eQTL-associated disease candidate genes. Genes in category A are those where expression change has already been related to the relevant disease. Those in category B are cases where the eQTL candidate gene has already been proposed as disease involved, usually from a GWA study, but an expression mechanism has not previously been suggested. The genes in category C are those that have not previously been proposed as disease relevant. (Genes in the strong LD immune protein region on chromosome 6 are not included because of ambiguous candidate gene assignments). Table 3.3 shows the number of loci with genes in each category for each disease. Only a small number of disease candidate genes have a previously proposed expression mechanism. There are 94 genes in Category B – previously disease associated genes where we have now identified a putative expression mechanism. False positives are most likely to be in Category C, but we do expect a substantial fraction of these new disease candidate genes will turn out to be correct. As illustrated below, in some cases, the new candidates are supported by circumstantial evidence of biological relevance. Appendix Table S6. lists the Category assignment for each eQTL disease candidate gene.

Table 3.3. Number of genes in each category for each disease.
Category A genes are those where an expression mechanism has previously been suggested, and the new analysis supports that finding. Category B genes are those where the disease candidate gene has previously been suggested, and we have now

identified a putative expression mechanism. . Category C genes are those where the expression related candidate genes have not previously been suggested as disease relevant.

Category	BD	CAD	CD	HT	RA	T1D	T2D
A	1	4	4	0	1	2	3
B	21	12	38	3	12	9	4
C	25	23	67	8	26	35	14

We compared the overlap of disease and eQTL loci with that expected by chance in the following way. We selected all the WTCCC1 GWAS disease loci that have a single identified candidate gene, to provide the most confident subset of likely causal genes. We then determined how many of these genes are also exGenes in high confidence (HC) eQTL relationships at the most conservative eQTL LD level ($r^2 \geq 0.8$). We then used a chi-squared test on the 2 x 2 table of overlap and non-overlap between these HC-eQTL and disease GWAS genes (Table 3.4) to determine the probability the overlap is significantly different from chance. For disease GWAS genes, we selected the genes which are the only one gene reported in single locus for all disease sets of WTCCC1 data. And high-confidence eQTL genes were determined from AllCell_AllPop set at the most conservative LD level ($r^2 \geq 0.8$). In this test, we calculated four numbers, including the number of GWAS reported genes which are overlapped with eQTL genes, the number of GWAS reported genes which are not overlapped with eQTL genes, the number of eQTL genes which are not reported in GWAS, and the number of genes which are neither eQTL genes nor GWAS reported genes. A Chi-squared test with Yates correction and 1 degree of freedom returns a P

value of 0.462 and a two-tailed value of 0.50. Thus the overlap of eQTL and disease genes is not significantly different from random.

Table 3.4**.** The 2x2 table of numbers of genes in that are disease candidates and/or involved in in high-confidence eQTL relationships.

	GWAS gene	non GWAS gene	Total
HC-exGene	41	4211	**4252**
non HC-exGene	119	10629	**10748**
Total	**160**	**14840**	**15000**

Examples of disease associated eQTL relationships

ADAM15 for Crohn's disease

A GWA study identified a region with a marker SNP, rs1142287, in chromosome region 1q22, that is significantly associated with Crohn's disease risk (Franke et al., 2010). This SNP is a synonymous variant located on the exon region of SCAMP3, and this gene and a neighboring one, MUC1, were reported as candidate genes for Crohn's disease. MUC1 encodes a key constituent of mucus, the physical barrier that protects the intestinal epithelium from gut bacteria. MUC1 overexpression and hypoglycosylation have been reported in irritable bowel disease (IBD) (Campbell, Yu, & Rhodes, n.d.). Secretory carrier membrane protein 3 (encoded by SCAMP3) regulates EGFR trafficking within endosomal membranes. SCAMP3 is manipulated by intracellular salmonellae to acquire nutrients and influence host immune responses (Mota, Ramsden, Liu, Castle, & Holden, 2009). Thus, there is circumstantial evidence supporting both genes as Crohn's disease relevant. In our analysis, there is one Crohn's related eQTL relationship for MUC1, with expression data in two studies

included in the integrated eQTL set (Fairfax et al., 2012; Grundberg et al., 2012) but none for SCAMP3 (using the 0.05cM threshold). There is also a Crohn's related eQTL relationship associated with the expression level of ADAM15, supported by expression data from the same two eQTL studies. ADAM15 encodes a member of the disintegrin and metalloproteinase (ADAM) protein family of type I transmembrane glycoproteins, involved in cell adhesion and proteolytic ectodomain processing of cytokines and adhesion molecules. Although no genome-wide association studies have suggested ADAM15 as a candidate for involvement in Crohn's disease, Mosnier et al. (Mosnier et al., 2006) showed differential expression of ADAM15 in epithelial cells during inflammatory bowel disease compared with the normal colon and suggested a role of ADAM15 in leukocyte-endothelial cells transmigration associated with acute inflammatory changes in inflammatory bowel disease. On the basis of those results and our analysis, we suggest ADAM15 may be the candidate mechanism gene underlying the association between this chromosome region, 1q22, and Crohn's disease.

TSPAN3 and PSTPIP1 for Type 2 Diabetes

A marker SNP, rs7178572, in chromosome region 15q24, was identified as associated with Type 2 Diabetes risk by two GWA studies (Perry et al., 2012; Sim et al., 2011). This SNP is located on the intron region of HMG20A, and that gene was reported as a candidate gene for Type 2 Diabetes. HMG20A encodes a high mobility group (HMG)-domain protein that activates REST (RE-1 silencing transcription factor)-responsive genes that play a key role in the initiation of neuronal differentiation,

making it an unlikely candidate for involvement in Crohn's. From our analysis, two alternative genes are suggested by the eQTL data. First, two studies included in the integrated eQTL set (Fairfax et al., 2012; Lappalainen et al., 2013) have many marker SNPs that lie within 0.05 cM of the disease marker and are significantly associated with the expression level of TSPAN3 (tetraspanin 3). Second, three of eQTL studies (Grundberg et al., 2012; Lappalainen et al., 2013; Zeller et al., 2010) have found that some of these SNPs are also significantly associated with the expression level of PSTPIP1, (proline-serine-threonine phosphatase interacting protein 1).

The protein encoded by TSPAN3 is a member of the transmembrane 4 superfamily, which are cell-surface proteins. These proteins mediate signal transduction events that play a role in the regulation of cell development, activation, growth and motility. PSTPIP1 gene encodes a protein which binds to the cytoplasmic tail of CD2, an effector of T cell activation and adhesion, negatively affecting CD2-triggered T cell activation, and so regulates the actin cytoskeleton. The latter gene is relevant to immune processes. Although there is still no direct evidence to show the relationship between these two genes and Type 2 Diabetes, the eQTL analysis suggests further investigation is warranted.

GALNT4 for Hypertension

A marker SNP rs2681472, on chromosome region 12q21.3, is significantly associated with Hypertension in European origin and East Asian populations (Cho et al., 2012; Hong et al., 2010) and these GWA studies have proposed the ATP2B1 gene (ATPase,

Ca++ transporting, plasma membrane 1) as a nearby candidate gene for involvement in hypertension. A recent study has shown that ATP2B1 is involved in calcium homeostasis, related to essential hypertension (Hirawa, Fujiwara, & Umemura, 2013). From our eQTL analysis, we found no eQTL SNPs in this region associated with ATP2B1 expression. However, two studies included in the integrated set (Grundberg et al., 2012; Zeller et al., 2010) have several SNPs that are within 0.005 cM of the disease marker and that are significantly associated with the expression level of another near-by gene, GALNT4 (polypeptide N-acetylgalactosaminyltransferase 4). Although there is no GWA study showing an association between GALNT4 and Hypertension, one recent GWA study suggested GALNT4 plays a causal role in susceptibility to atherosclerosis, related to high blood pressure (Erbilgin et al., 2013). The GALNT4 gene encodes the N-acetyl galactosaminyl transferase 4 enzyme and is thought to involve in endothelial-platelet interactions by O-glycosylating the threonine residues of the P-selectin glycoprotein ligand (PSGL-1). We suggest that the underlying mechanism in the 12q21.3 region associated with Hypertension likely involves altered expression of GALNT4.

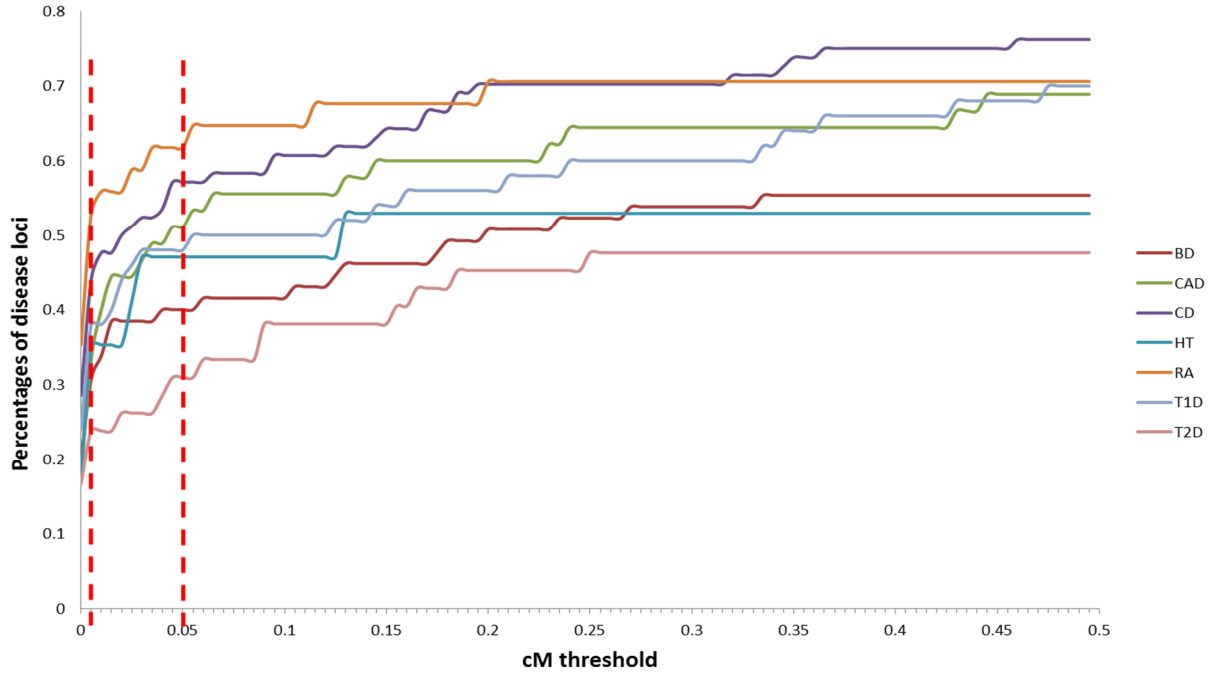


Figure 3. 1. Percentage of disease loci with possible expression mechanisms as a function of the cM distance between the closest disease and expression marker SNPs.

The AllCell_AllPop eQTL set was used. Two vertical dotted lines indicate the cM thresholds, 0.005 and 0.05. The maximum threshold used in this study is 0.05 cM.

3.3: Methods

High-confidence eQTL Data

High-confidence eQTL data were derived on the basis of consensus within the included 16 independent human genome wide eQTL studies, as described in Chapter 2. Briefly, for disease analysis, a high-confidence eQTL relationship is defined as one that is identified in at least two studies of these 16. The number of high confidence eQTL relationships so defined varies with the Linkage disequilibrium criterion used. For the disease analysis, the most conservative linkage disequilibrium level ($r^2 \geq 0.8$) was used, providing a total of 4,252 unique genes with an expression level associated with the presence of at least one high-confidence eQTL SNP.

Genome-wide associations studies of human common diseases

Loci significantly associated with disease susceptibility for seven specific human common diseases (Bipolar disorder (BD), Coronary artery disease (CAD), Crohn's disease (CD), Hypertension (HT), Rheumatoid arthritis (RA), Type 1 diabetes (T1D), and Type 2 diabetes (T2D)) were collected from the Wellcome Trust Case Control Consortium (WTCCC1) GWA study (The Wellcome Trust Case Control Consortium, 2007) and from other related meta analyses and follow-up studies in the GWAS catalog (www.genome.gov/gwastudies/). Appendix Table S4. lists all the GWA studies included.

LD Relationships

LD relationships were taken directly from data in the Hapmap project (The International HapMap 3 Consortium, 2010) and also derived from the 1000 Genomes

project data (Abecasis et al., 2010), using PLINK (v. 1.07) (Purcell et al., 2007). The complete microarray genotype data were downloaded for the WTCCC1 study of seven complex trait diseases and the probabilities of each genotype for the SNPs in each disease locus from WTCCC1 GWA study not represented on the microarray were imputed using IMPUTE2 (Howie et al., 2009) and then the disease association P-value of each SNP was calculated using SNPTEST (Ferreira & Marchini, 2011).

CentiMorgan distance calculation

The genetic map data of all human chromosomes, calculated from HapmapII data with LDhat (ldhat.sourceforge.net/instructions.shtml) was acquired from NCBI FTP (ftp.ncbi.nlm.nih.gov/hapmap/recombination/2011-01_phaseII_B37/). Where necessary, the centiMorgan co-ordinates of disease associated marker SNPs and expression associated eQTLs were interpolated from those of the closest SNPs with defined centiMorgan values, based on chromosomal distance.

Comparison of disease and eQTL markers

We require a procedure that estimates whether or not the detected disease and eQTL markers in a locus arise from the same underlying causal variant. The model used for this purpose is shown in Figure 3.2.

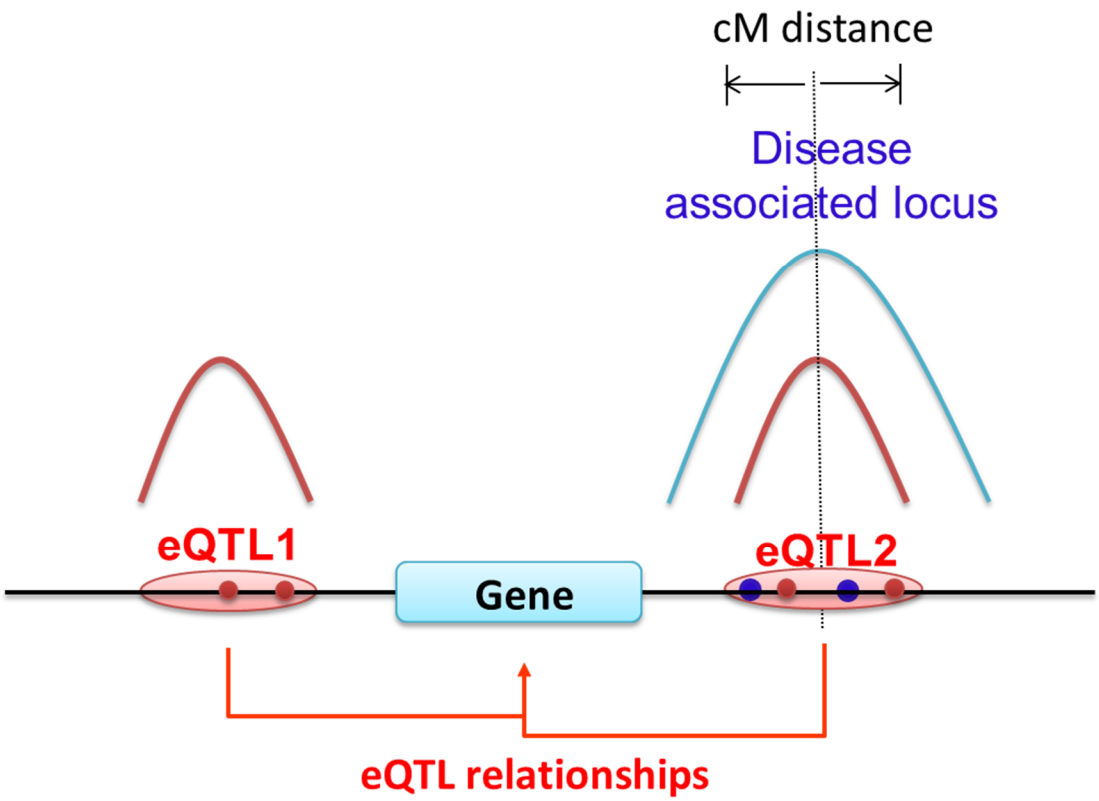


Figure 3.2. Model for identifying those disease associated loci with a probable underlying expression mechanism.

In this hypothetical case, a causal variant, at the position of the vertical dotted line, is related to disease susceptibility as a result of altering the expression level of the nearby gene. Because of LD, the presence of the causal variant will usually result in one or more nearby SNPs being associated with disease risk, and the blue curve represents the expected P-value distribution of these. Sparse sampling with a microarray and noise factors result in only one or a few of these associations being detected (blue dots). Since the causal variant affects expression, the same SNPs will be associated with expression level of the gene, with a co-located expected P value distribution, represented by the red curve, and, again because of noise and other factors, only some markers will be identified (red dots). In this example, there is

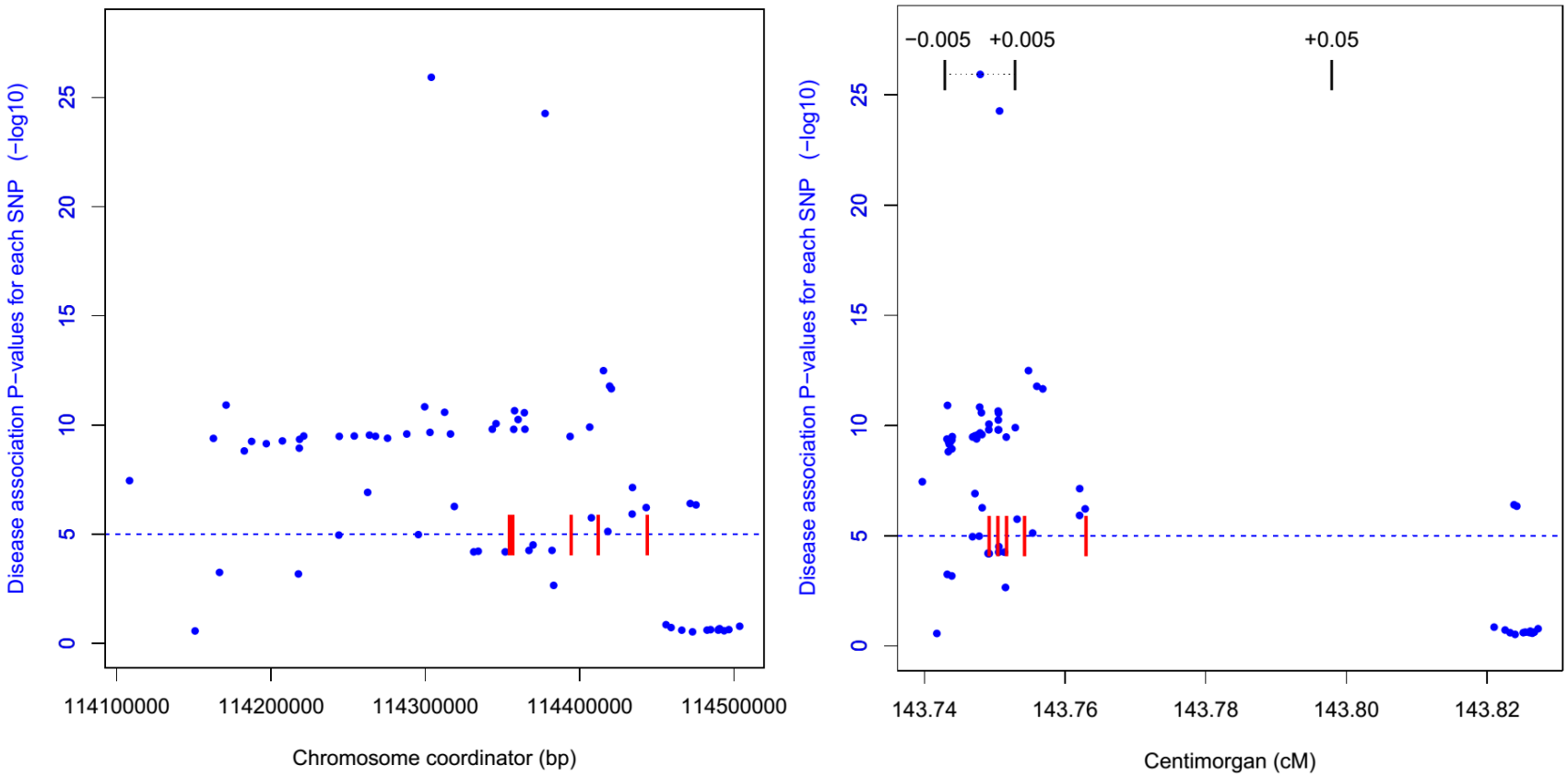
another eQTL in this region (eQTL1) where SNPs are associated with the expression level of the same gene, but unrelated to disease susceptibility, and so its eQTL P-value distribution does not overlap with that for disease association.

As noted earlier, imputation methods allow estimation of P-values for SNPs not present on the disease association microarray, given complete genotype information for all individuals in a study. We have used compete genotype data for the WTCCC1 study of seven complex trait diseases (The Wellcome Trust Case Control Consortium, 2007) in order to examine the relationship between disease association P-value distributions and eQTL markers. P-values for SNPs not on the microarray were obtained using SNPTEST (Ferreira & Marchini, 2011). Full imputed disease association P value distributions were compared with marker SNPs for high-confidence eQTL relationships derived across the 16 eQTL studies (AllCell_AllPop).

Figure 3.3 shows Manhattan plots of these data for a region where SNPs are significantly associated with the risk of Type 1 diabetes in the WTCCC1 study, and that also contains eQTL associations. The left hand plots show the distribution of disease association P-values and the location of the expression marker SNPs as a function of the chromosome coordinate, in base-pair units. In these plots, it is often not possible to judge whether or not the disease and expression signals share a causal variant. The right hand figures show the same data, but as a function of the cross-over event probability, measured in centiMorgans (cM). For the example in Figure 3.3, the cM scale allows a clear distinction between situations where the underlying causal

variant for the disease and expression signals are the same (Figure 3.3A) and where they are different (Figure 3.3B).

A. RCL1


Disease association P−values for each SNP (−log10)
0
5
10
15
20
25
114100000
114200000
114300000
114400000
114500000
Chromosome coordinator (bp)
−0.005
+0.005
+0.05
143.74
143.76
143.78
143.80
143.82
Centimorgan (cM)

B. DCLRE1B

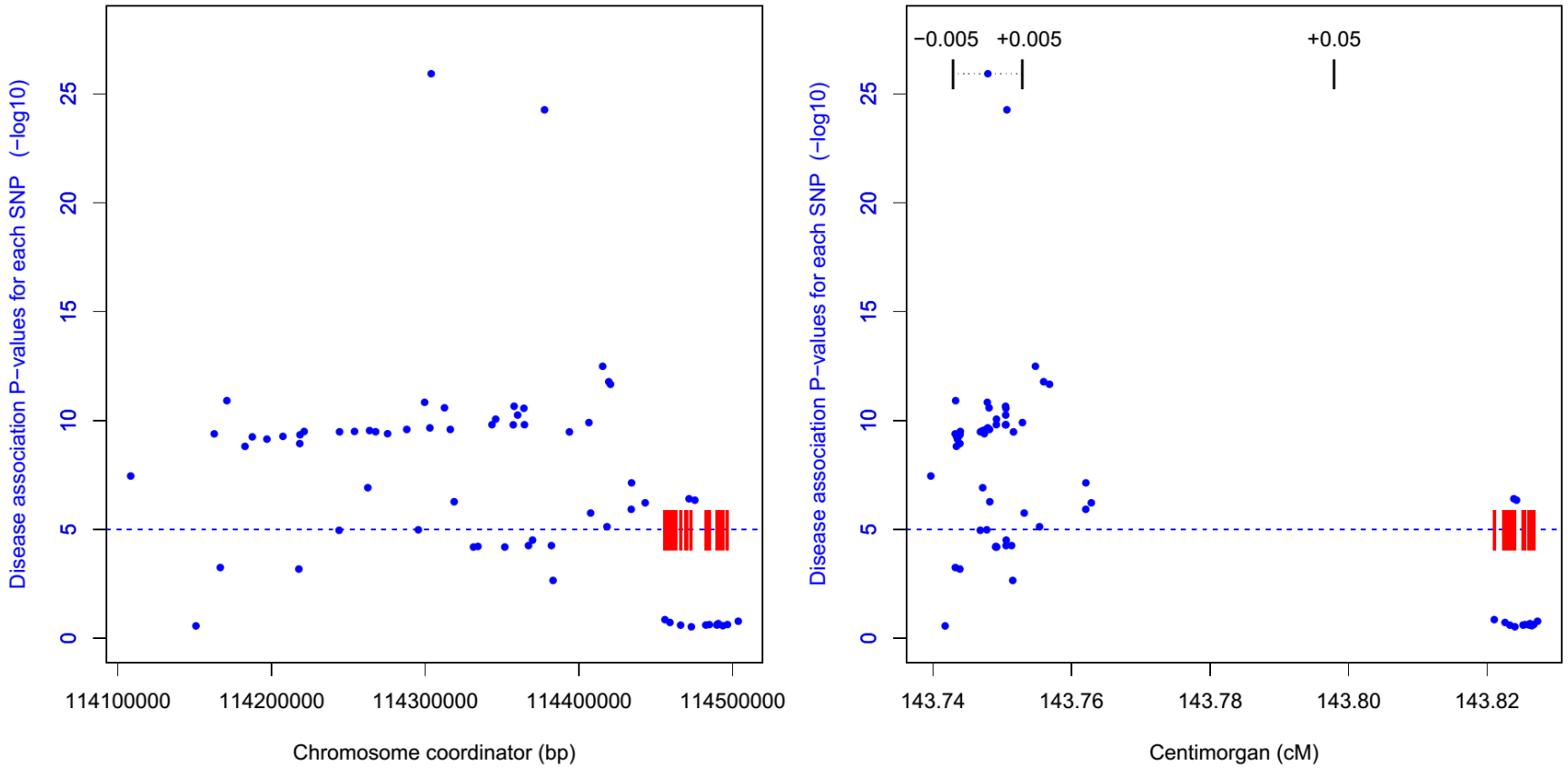


Figure 3.3. Manhattan plots for a locus associated with Type 1 diabetes in the WTCCC1 data.

These plots show the relationship between disease association P-value for all SNPs in the region (blue points) and the location of high-confidence expression associated SNPs (red dashes). There are two separate high confidence eQTL relationships in this region, each involving a different gene. The dotted line indicates the significance threshold for disease P values (10^{-5}). The left plots show the P-value distribution of disease and expression SNPs as a function of chromosome coordinates and the right plot show the same data as a function of genetic map position, in centiMorgans (cM). (A) Disease associations and high-confidence eQTL SNPs associated with the expression level of RCL1 (RNA terminal phosphate cyclase-like 1). In chromosome co-ordinates (left), the disease markers appear widely spread, and there is no clear distinction between these and eQTL markers. On the cM scale (right), it is clear that the disease marker SNPs and eQTL SNPs occupy the same narrow range in the cross-over coordinate. (B) High confidence eQTL SNPs associated with DCLRE1B (DNA cross-link repair 1B) in the same locus. In chromosome co-ordinates (left) it is unclear whether these markers overlap with the disease markers or not. On the cM scale (right) there is a clear separation between expression and disease markers, reflecting low linkage disequilibrium between the two sets of markers, and so that it is unlikely the same causal variant generates both signals. Together, these plots show that the data are consistent with a disease susceptibility causal variant affecting the expression of RCL1, and inconsistent with an expression effect on DCLRE1B.

Appendix Figure S1 shows more plots for the other 16 of the 21 WTCCC1 loci that contain at least one high-confidence eQTL relationship. 14 of the 17 loci have eQTL markers that overlap with the disease marker SNPs. Consistently, in this set, where there is overlap, the shortest distance between a disease marker and an eQTL marker is less than 0.05 cM, and in no case where there is not overlap is there a distance less than 0.05. On that basis, we adopted three thresholds for confidence that the disease and expression signals arise from a common underlying variant: When there is an exact match between a disease marker and an expression marker (i.e. these are the same SNP), when the closest disease and expression markers are with 0.005 cM, and when the two closest markers are within 0.05 cM.

3.4: Discussion:

It has long been appreciated that expression mechanisms may play a major role in complex trait disease, and some studies have already provided data to support this idea (Cookson et al., 2009; Nicolae et al., 2010). Up to now, though, it has not been possible to determine how generally this is the case, or which disease associated loci may harbor expression related mechanisms. In this study, by combining current eQTL data and disease GWAS data, we have been able to address these questions on a relatively large scale. We find that, conservatively, approaching 50% of disease loci have a high confidence eQTL relationship consistent with an underlying expression mechanism. The fraction of loci with putative expression mechanisms ranges from 30 to 60%, depending on the disease. We have illustrated that these data are useful for better identifying disease relevant genes in particular loci. Each proposed expression mechanism defines possible follow-up experiments.

Like all computational models, this one is not a perfect replica of the real world, and there are both false positive and false negative aspects to the results. Where possible, we have taken a conservative approach, and placed most emphasis on minimizing potential false positives, particularly those that might arise from the apparent unreliability of large scale eQTL results, as evidenced by poor agreement between independent studies, discussed in Chapter 2. Thus, we include only consensus eQTLs – those found in more than one independent study. Only approximately a quarter of the genes found to be involved in eQTL relationships in single studies qualify as high confidence as a result of having been observed in two or more studies, so it is likely

this filter prevents the identification of some bona fide eQTL-disease relationships. In this sense, there are a substantial fraction of false negatives.

Decisions as to whether eQTL relationships found in different studies are identical are based on LD relationships between SNPs, and so are sensitive to the LD threshold used – the more lenient the LD threshold, the more relationships appear the same. We used three different LD thresholds for this, but for the purpose of establishing relationships to disease, considered only the most conservative ($r2 > 0.8$). There are 68% fewer high confidence eQTLs than if the most lenient ($r2 > 0.3$) criterion weas used, possibly leading to the omission of some disease loci with underlying eQTLs. eQTL relationships may vary depending on cell type and also cell state– whether an immune system cell is active, for example. In complex trait disease, it is often difficult to know which cell type is implicated in each disease locus, and even if this is clear, expression data for that cell in that state are unlikely to be available. Typically, it has been assumed that these differences are secondary, and most disease/expression studies have used eQTLs from LCLs (Cookson et al., 2009; Nicolae et al., 2010). One study across multiple tissue types has suggested the degree of tissue dependence is large (69%-80%) (Dimas et al., 2009). Analysis across studies in Chapter 2 suggests this is not the case. There, between 20 and 40% fewer relationships are found comparing data for different tissues across studies as when comparing data for the same tissues across studies. The Chapter 2 results suggest that conservation across populations is higher than that across tissues, although the data are limited for both.

How does the variation of eQTLs across tissues affect the disease results? Random expectation is that in about half the cases where an eQTL is present in one tissue and not the other, it will be the eQTL study tissue that does not exhibit it and the disease tissue that does, leading to false negatives, consistent with our conservative strategy. In the rest of the cases, the eQTL will not be present in the disease relevant cell type, but is in the reference eQTL datasets. In some of these instances, an eQTL present in a reference tissue will not apply to the disease tissue because in the latter the gene is not significantly expressed. We saw evidence of this in Chapter 2, where eQTLs specific to liver were for genes only expressed in that tissue. If a gene is not expressed in the disease relevant cell type, it cannot be disease relevant, so such cases are not of concern. Finally there will be a fraction of cases where a gene is, say, controlled by a different transcription factor in the two tissues, and so a variant may affect transcription factor binding and so expression in one cell (the eQTL reference) will not affect expression in the other, leading a false positive. At present, there are insufficient data to reliably estimate how common that situation is. Overall, given the 20 – 40 % non-tissue transferability seen in Chapter 2, it's likely there is between 5 and 20% false positives from this cause.

Most disease GWA studies have been conducted in Caucasian or closely related populations, and the majority of the eQTL data are from that source. But a few are not so there is also a consideration of transferability here too. From Chapter 2, the limited data suggest across population consistency is relatively large, and, following the

analysis of the previous paragraph for cell type consistency, and considering the dominance of appropriate populations in the eQTL data, false positives from this cause are expected to be minor.

A key step in identifying which disease loci have a potential underlying expression related mechanism is comparing markers from eQTL studies with those from disease GWAS. As described earlier, because of the absence of complete genotyping for individuals in both type of studies, this depends on the threshold used. For 79 of loci, there is exact agreement between at least one disease marker and one eQTL marker. Where that is not the case, in a further 125 of loci, the two markers are very close in LD space, less 0.005 cM. The remaining 51% included, out to a separation of 0.05 cM, are still within a conservative threshold, so some eQTL disease mechanisms may have been missed for this reason.

It may be that in some cases where there is an eQTL mechanism underlying a disease association, that is not the dominant mechanism contributing disease susceptibility. Other work in the lab (Ray and Moult, unpublished) has shown that a significant fraction of these disease loci have a potential high impact missense SNP disease mechanism. Expression effects for the data analyzed in Chapter 2 are usually relatively small, with a median value of 1.14 fold change in the level of expression, and few greater than two-fold. In contrast, high impact missense typically change in vivo activity of a protein by 5 to 10 fold, sometimes more (Yampolsky & Stoltzfus,

2005). Where both mechanisms are present in a locus, the missense one will likely dominate.

GWAS disease studies typically find high frequency SNPs with modest phenotype effects (Hindorff et al., 2009). That suggests that genes involved in disease mechanism may not make good drug targets – reversing the effect of the disease associated variant will only have a small impact on disease. Consistent with this, a previous study in our group has shown that very few GWAS disease candidate genes are known drug targets for the corresponding disease (Cao & Moult, 2014).From this point of view, disease candidate genes where a subtle expression effect is enough to produce a detectable consequence in disease susceptibility may be worth closer examination for drug target potential – if a small effect is detectable, a bigger effect achieved by a drug may be useful.

Nicolae et al. (Nicolae et al., 2010) has reported a higher than random relationship between eQTL SNPs and GWAS disease risk SNPs. However, in this study, we did not find a greater than random coincidence of the genes involved in high confidence eQTL relationships and proposed candidate genes for involvement increased disease risk. Our test is straightforward, and, we believe, reasonably robust. So why is the difference in findings? Without extensive investigation of the earlier results, we cannot be completely certain, but there is one apparent strong bias in the earlier SNP based tests. In that work, sets of random disease markers were generated by randomly choosing SNPs from amongst all those represented on the microarray used for the

GWAS study. Microarray SNPs are chosen considering a number of criteria, and aim to approximately span the whole of the genome. But real disease markers are heavily biased to be close to genes. And, as shown in Figure 2.2, so are cis-eQTL marker SNPs. Thus, markers chosen from a broadly randomly distributed set of SNPs are much less likely to overlap with eQTL SNPs than real markers will. To perform the SNP based test properly, markers should be chosen based on the observed distance of real markers from genes, in manner similar to that we used in testing our high confidence high eQTL relationships (see Appendix).

At first glance a lack of enhanced overlap of eQTL and disease sites might seem surprising. In fact, we do expect to see this. The expectation of such enhancement rests on the assumption that eQTLs are fairly rare, so that only a subset of genes with the property could be involved in GWAS detectable disease risk. The 11 eQTL studies included in our work purported to find eQTL relationships for over 15000 unique genes, probably essentially all genes with high enough expression to be relevant to eQTL. While many of these are likely false positives, there are also many false negatives and even with this limited set of studies, our very conservative high confidence set covers over 4000 genes. So, in reality most genes probably do have an eQTL that potentially could contribute to disease risk. The determining factor is not whether an eQTL is there but rather whether perturbing the activity level of the gene product is relevant to the disease phenotype.

Chapter 4: Web-based database for query and visualization of human genome-wide expression quantitative trait loci

4.1: Introduction

Recent large-scale investigation of the relationship between human genetic variation and transcriptional regulation of gene expression is providing new biological insights into the mechanisms by which altered expression contributes to disease pathogenesis. With the rapid improvement of high-throughput technology, a number of genome-wide expression quantitative trait loci (eQTL) mapping studies have together generated hundreds of thousands of associations between the presence of a SNP and altered expression of a gene for various human tissues in different populations. The analysis of the underlying mechanisms of these eQTL relationships is expected to be of great help in dissecting the relationship between genome variability and human complex traits. However, as a result of a number of factors, there is often substantial disagreement between the results of different eQTL studies (Dixon et al., 2007; Göring et al., 2007; Myers et al., 2007; Stranger et al., 2007; Veyrieras et al., 2008). To address this issue, we have implemented a procedure to identify the more high-confidence regulatory eQTLs, based on consistency across multiple studies.

Currently, there are a number of web-based databases and software for eQTL data interpretation and analysis. eQTL Explorer (web.bioinformatics.ic.ac.uk/eqtlexplorer/) facilitates mining of results from genome-wide linkage analyses and provides visualization to aid interpretation of the eQTL

data through a Java graphical interface (Hubner et al., 2005; Mueller et al., 2006). eQTL Viewer (statgen.ncsu.edu/eQTLViewer/svgHome.html) is a web-based bioinformatics tool that generates a scalable two-dimensional graph for visualizing eQTL mapping results (Gelfond, Ibrahim, & Zou, 2007). SNPexp (tinyurl.com/snpexp) is a web-based tool for visualization of eQTL mapping results (Holm, Melum, Franke, & Karlsen, 2010). In addition, several online databases that collect data from multiple human genome-wide eQTL studies are also available. SCAN (www.scandb.org/newinterface/) is primarily designed for accessing functional annotation related to SNPs and includes results from one eQTL study conducted in Lymphoblastoid cell lines using individuals from HapMap populations and several eQTL studies in additional human tissues, such as brain and liver (Gamazon, Zhang, et al., 2010). eQTL Browser (eqtl.uchicago.edu/) is a database that allows a user to navigate eQTLs from several recent studies in multiple tissues. seeQTL (www.bios.unc.edu/research/genomic_software/seeQTL/) provides reanalyzed eQTL associations from several studies and display the results in the genome browser (Xia et al., 2011). It also provides a consensus association score for each eQTL across all LCL studies that have used HapMap populations. GTEx (www.gtexportal.org/home/), part of the NIH GTEx roadmap project, currently provide a central resource to archive and display association between genetic variation and high-throughput molecular-level phenotypes (The GTEx Consortium 2013). Genevar (www.sanger.ac.uk/resources/software/genevar/) is a platform of database and web services designed for integrative analysis and visualization of SNP-gene associations in eQTL studies (Yang et al., 2010).

The resource described in this study introduces additional capabilities not presently available. It provides comprehensive access to the results of work described in our previous two chapters. Briefly in Chapter 2, we integrated 29 eQTL datasets from 16 publicly available genome-wide studies for various human tissues spanning a number of different populations. We made use of linkage disequilibrium (LD) information acquired from the HapMap project (Altshuler et al., 2010) and also LD data derived from the 1000 Genomes project (Abecasis et al., 2010) to develop a method for identifying more reliable eQTL relationships, based on consensus across studies. We also compared eQTLs across different tissues and populations to find currently cell type-dependent or population-dependent eQTLs. To illuminate possible mechanisms underlying the eQTL associations, we mapped eQTL SNPs to some annotated functional elements. In Chapter 3, we described mapping of eQTL relationships to the results of disease GWA studies on seven complex trait diseases so as to identify those disease loci with a putative eQTL mechanism. The ExSNP resource not only facilitates the querying of these useful eQTL data but also provides a comprehensive genome browser to visualize the relative positions of SNPs to their associated genes and other neighboring genes.

4.2: Construction and content

Database construction

The ExSNP database contains the following components: original eQTLs, high-confidence eQTLs, cell type-dependent eQTLs, population-dependent eQTLs, disease associated eQTLs, and functionally annotated eQTLs (Figure 4.1).

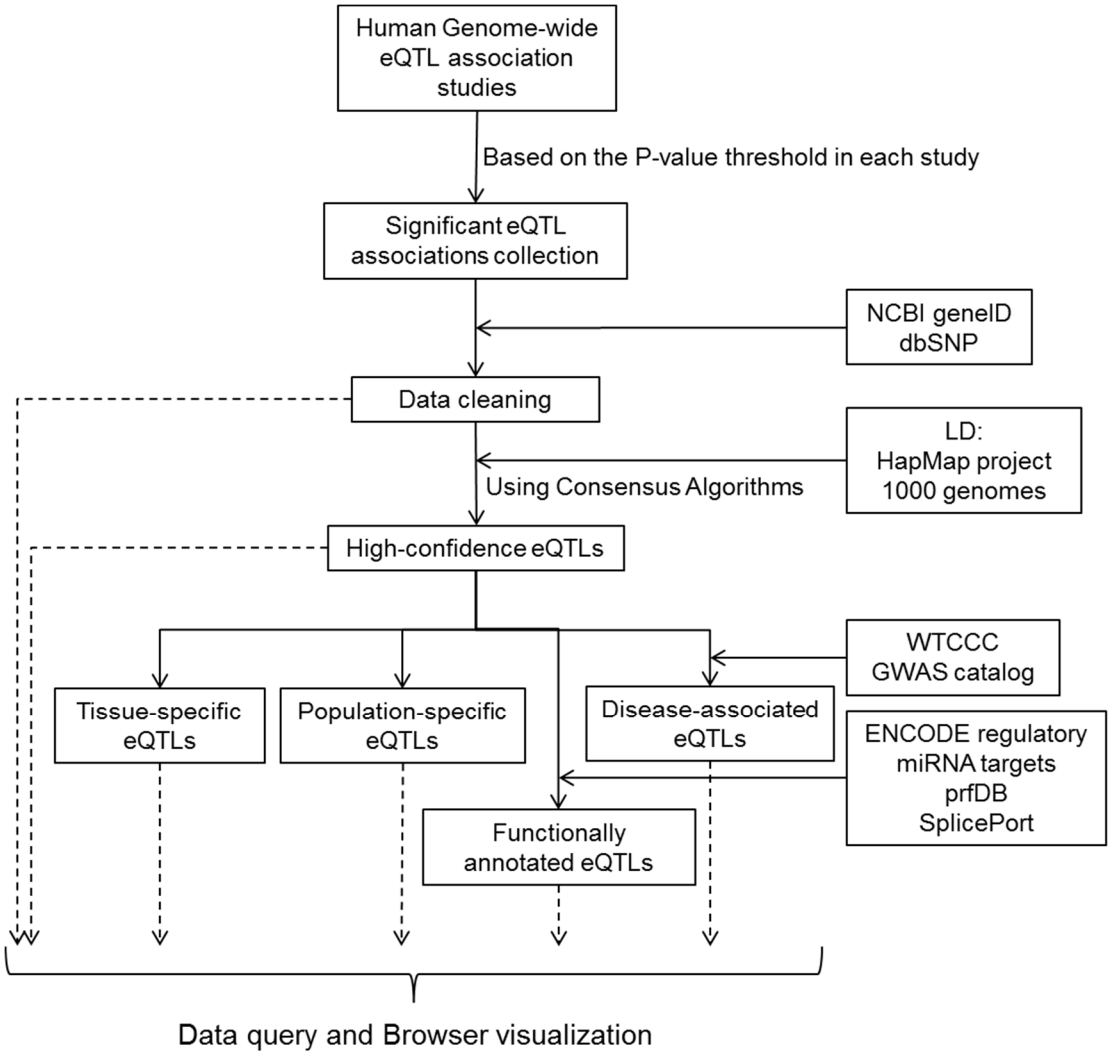


Figure 4.1. Workflow for the construction of the ExSNP database.

eQTL data sources and processing

All genome-wide human eQTL association data were collected from 16 publicly available studies, covering various tissues and human populations. Nine of these studies were conducted in the most accessible cell type, lymphoblastoid cell lines (Dimas et al., 2009; Dixon et al., 2007; Duan et al., 2008; Grundberg et al., 2012; Lappalainen et al., 2013; Liang et al., 2013; Montgomery et al., 2010; Pickrell et al., 2010; Stranger et al., 2007). Four of these LCL studies mapped eQTLs of individuals from the HapMap Project (Duan et al., 2008; Montgomery et al., 2010; Pickrell et al., 2010; Stranger et al., 2007) and one study mapped eQTLs of individuals from the 1000 Genomes Project (Lappalainen et al., 2013). Two studies performed eQTL mapping on LCLs in a Childhood Asthma sibling cohort (Dixon et al., 2007; Liang et al., 2013). In addition to those on LCLs, we included several studies on other single tissue types, specifically two on brain (Gibbs et al., 2010; Myers et al., 2007) and one each on liver (Innocenti et al., 2011; Eric E Schadt et al., 2008), monocytes (Zeller et al., 2010), and skin (Ding et al., 2010). Three of the included studies covered multiple tissue types. One of these (Dimas et al., 2009), investigated and compared eQTLs from three cell types: LCLs, primary fibroblasts, and primary T-cells. Another study (Fairfax et al., 2012) focused on two circulating immune cells, primary monocytes and B-cells. One study (Grundberg et al., 2012) discovered eQTLs in three cell types, LCLs, skins, adipose, derived from a subset of well-phenotyped healthy female twins in the MuTHER resource (Nica et al., 2011). Most of the 16 studies used transcript microarrays to measure RNA expression level. In contrast, three studies (Lappalainen et al., 2013; Montgomery et al., 2010; Pickrell et al., 2010) estimated transcript levels

of genes by using RNA-Seq methods. All but the 1000 genomes study used data from genotyping microarrays. Since the individuals in that study were fully sequenced, full SNP information was already available.

All transcript names, probe IDs and alias gene names used in each study were converted to current unique Entrez Gene IDs and Gene names (NCBI build 37.2). In addition, retired and discontinued SNP IDs were filtered out and all valid SNPs were converted to dbSNP rsIDs (build134). In total, there are 796,908 eQTL associations, comprising 548,344 unique SNPs and 15,170 unique genes. Linkage Disequilibrium (LD) data for pairs of SNPs were directly gathered from the International HapMap Project (Altshuler et al., 2010) and also derived from the 1000 Genomes Project (Abecasis et al., 2010) using PLINK (Purcell et al., 2007).

Identification of consensus eQTLs

We first identified the set unique eQTL relationships within each of the 29 datasets. Typically, as a result of linkage disequilibrium, each underlying unique eQTL relationship is associated with the presence of multiple SNPs (marker SNPs). Marker SNPs for each gene involved in an eQTL relationship were grouped into sub-sets based on linkage disequilibrium between them. We then compared these unique eQTL relationships across studies, to determine which pairs are compatible with the same underlying eQTL. The confidence score for each eQTL relationship is defined as the number of studies that have identified that relationship. Three different levels

of LD relationships were used, corresponding to r^2 thresholds of 0.8, 0.5, and 0.3. Further details of this procedure were given in Chapter 2.

Population-dependent and cell type-dependent eQTLs

In order to examine eQTLs for various population and cell types, we divided all studies into 12 integrated sets. LCL has so far been the most commonly used cell type, allowing us to form three population specific datasets for the Caucasian, African, and Asian populations with this cell type. Eight studies were included in the Caucasian set (Dimas et al., 2009; Dixon et al., 2007; Duan et al., 2008; Grundberg et al., 2012; Lappalainen et al., 2013; Liang et al., 2013; Montgomery et al., 2010; Pickrell et al., 2010; Stranger et al., 2007). Four studies were included in the African set (Duan et al., 2008; Lappalainen et al., 2013; Pickrell et al., 2010; Stranger et al., 2007), and one study was included in Asian set (Stranger et al., 2007). We then identified the population-dependent eQTL relationships which appear only in each single population set or across multiple populations..

In addition, we constructed nine integrated sets, one for each cell type. These are LCL (Dimas et al., 2009; Dixon et al., 2007; Duan et al., 2008; Grundberg et al., 2012; Lappalainen et al., 2013; Liang et al., 2013; Montgomery et al., 2010; Pickrell et al., 2010; Stranger et al., 2007), brain (Gibbs et al., 2010; Myers et al., 2007), liver (Innocenti et al., 2011; Eric E Schadt et al., 2008), monocyte (Fairfax et al., 2012; Zeller et al., 2010) , B-cell (Fairfax et al., 2012), T-cell (Dimas et al., 2009), fibroblast (Dimas et al., 2009), adipose (Grundberg et al., 2012), and skin (Ding et al.,

2010; Grundberg et al., 2012). The cell type-dependent eQTL relationships which appear only in each single cell-type set or across multiple cell types have also been identified.

Functional interpretation of eQTLs

To identify possible mechanisms underlying these eQTL associations, we mapped marker SNPs for each eQTL relationship to several types of functional elements annotated on the human genome. Functional element data were acquired from several publicly available resources. Information on microRNA binding sites was taken from TargetScan (5.1) (Grimson et al., 2007), MicroCosm Targets (v5) (Kozomara & Griffiths-Jones, 2011), and microRNA.org (Aug 2010 release) (Betel et al., 2008). Transcription Factor binding sites and DNaseI hypersensitivity regions were extracted from the Integrated regulation tracks of the ENCODE Project (Dunham et al., 2012). The sensitive and ultra-sensitive sites discovered in the 1000 Genomes Project (Phase 1) (Khurana et al., 2013) were used for sequence conservation analysis. Predicted potential programmed -1 ribosomal frameshift (-1 PRF) regions were collected from PRFdb (Belew, Hepler, Jacobs, & Dinman, 2008; Jacobs, Belew, Rakauskaite, & Dinman, 2007; Plant et al., 2004). Data of SNPs potentially affecting splicing were retrieved from SplicePort (Dogan et al., 2007).

Disease associated eQTLs

To investigate which disease risk loci discovered by GWAS may be related to an underlying eQTL mechanism, we compared SNPs that have been found to be

associated with disease risk with those involved in high confidence eQTL relationships. GWAS results for seven common human diseases (Bipolar disorder, Crohn's disease, Coronary artery disease, Hypertension, Rheumatoid arthritis, Type 1 diabetes, and Type 2 diabetes) were obtained from the GWAS catalog (www.genome.gov/gwastudies). Detailed data for the WTCCC1 study of these diseases (The Wellcome Trust Case Control Consortium, 2007) were used to derive a relationship between disease and eQTL marker SNPs consistent with the same underlying causal variant, and on this basis, marker SNPs for eQTLs that are within a threshold centi-Morgan (cM) distances (0.005 or 0.05) of disease risk markers were considered to represent the same mechanism. In addition to these seven diseases, we also investigated the relationship between high-confidence eQTL relationships and complex traits where eQTL data for the appropriate cell types are available, again taking data from the NHGRI GWAS Catalog (www.genome.gov/gwastudies/). The cell types are Brain, LCL, Liver, and Skin.

Web implementation

ExSNP is implemented using a LAMP (including Linux, Apache, MySQL, and PHP/Perl) platform. The web utility is supported and structured by a relational model using MySQL, and the web interface is executed in PHP-HTML. The server-side script is written in Perl (5.10). It provides a web-based interface for data query and search. The ExSNP browser leverages the power of Scalable Vector Graphics (SVG) to display a vector-based graph on the Website.

4.3: Use of the resource

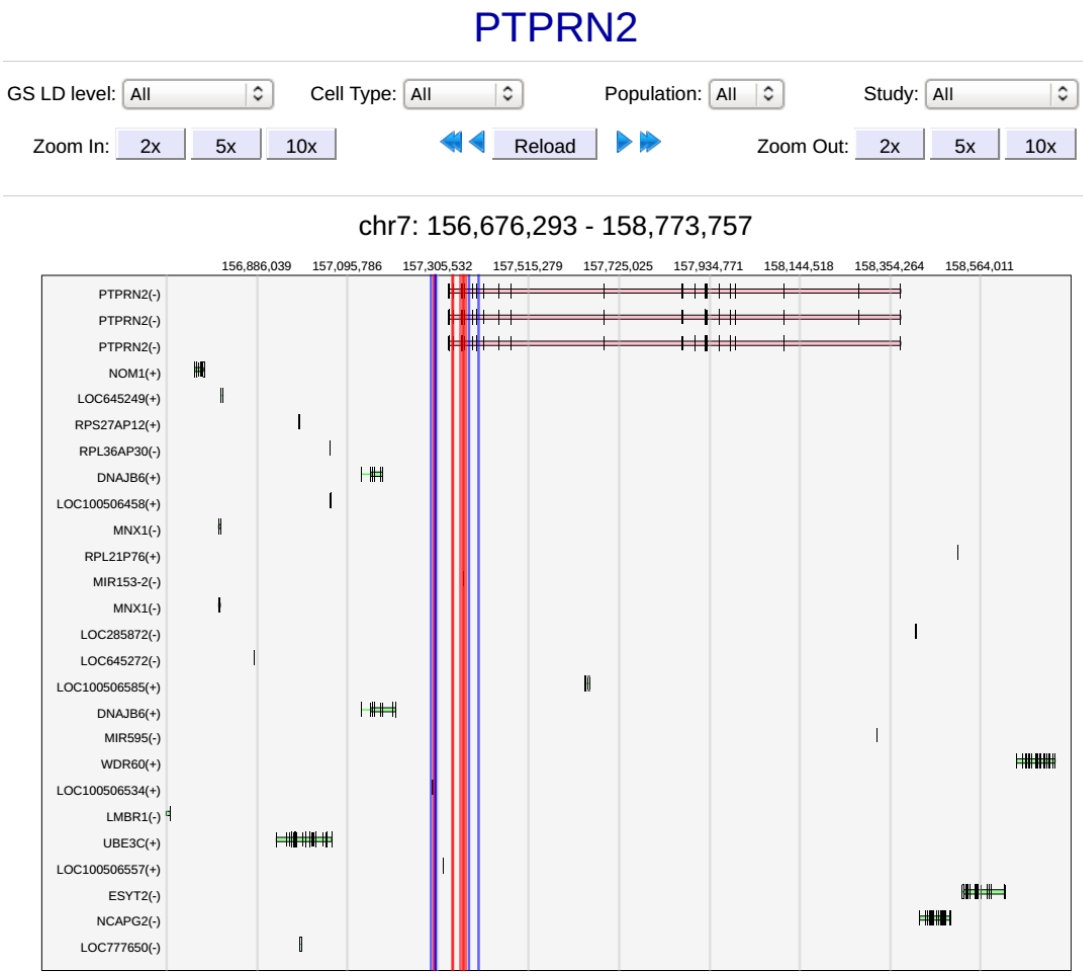
eQTL related Queries

ExSNP provides three different approaches for data searching: by SNP ID, by Gene ID, and by Gene name. A user can search against all included eQTLs or just high-confidence eQTLs. One can also ask if one or more SNPs of interest are in an LD relationship with any SNP associated with gene expression at various LD levels ($r^2 \geq$ 0.8, 0.5, or 0.3). For retrieval of tissue-dependent or population-dependent eQTL relationships, a user can query by SNP ID, Gene ID, or Gene name and also select specific tissue/population sets. Disease associated eQTLs can be queried by SNP ID, Gene ID, or Gene name for a selected disease and specific eQTL set at various LD thresholds. SNPs located on functional regions can also be queried by selected functional element type.

eQTL browsing

The ExSNP browser is able to display the relative position of SNPs involved in eQTL relationships and the associated genes together with other near-by genes. The interactive interface allows a user to zoom in on a gene of interest and select specific sets of high-confident eQTLs for display. The browser also displays functional element regions for microRNA binding sites, transcription factor binding sites, and programmed -1 ribosomal frameshift signals, so as to facilitate identification of the possible functional impact of relevant SNPs. For example (Figure 4.2), there are total 14 SNPs associated with the expression level of PTPRN2. 13 of these SNPs are located near 3'end of gene region and one falls on an distant upstream inter-genic

region (not shown in the figure). These SNPs are identified in up to four independent studies (Fairfax et al., 2012; Grundberg et al., 2012; Lappalainen et al., 2013; Zeller et al., 2010) for three different cell types, Monocytes, B-cells, and LCLs. At a threshold of $r^2 \geq 0.3$, 12 out of the 14 SNPs form a single high-confidence eQTL relationship.



Position	SNP	Probe	Study ▾	Population	Cell	LD level	Overlap studies
chr7:157,289,684	rs887626	NM_130842.1	IM_B	CEU	Bcell	0	1
chr7:157,295,625	rs3735198	PTPRN2	MO	CEU	Monocyte	8	2
chr7:157,363,490	rs221296	PTPRN2	MO	CEU	Monocyte	5	3
chr7:157,370,054	rs221304	PTPRN2	MO	CEU	Monocyte	5	3
chr7:157,357,671	rs221289	PTPRN2	MO	CEU	Monocyte	5	3
chr7:157,361,734	rs221294	PTPRN2	MO	CEU	Monocyte	5	3
chr7:157,339,589	rs123776	PTPRN2	MO	CEU	Monocyte	5	3
chr7:157,338,366	rs221269	PTPRN2	MO	CEU	Monocyte	5	3
chr7:157,295,879	rs17837771	PTPRN2	MO	CEU	Monocyte	8	2
chr7:157,376,644	rs4716760	ILMN_1728844	MuTHER_LCL	CEU	LCL	0	1
chr7:157,300,517	rs11761751	ILMN_1728844	MuTHER_LCL	CEU	LCL	0	1
chr7:157,399,690	rs2024099	ILMN_1728844	MuTHER_LCL	CEU	LCL	0	1
chr7:157,300,845	rs11761845	ILMN_1728844	MuTHER_LCL	CEU	LCL	0	1

Figure 4.2. Sample screenshot from the ExSNP browser.

This example shows all the eQTL associations of one gene, PTPRN2. The red bars in the plot show the transcripts of the target gene, PTPRN2. The green bars are the

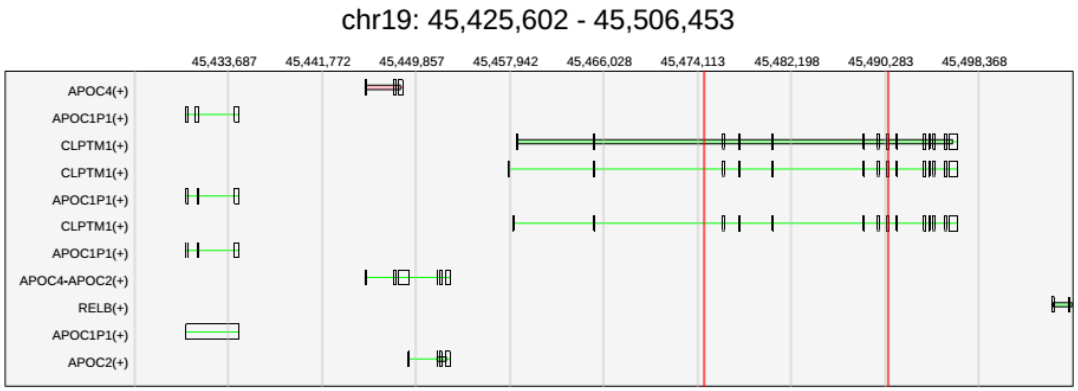
transcripts of neighboring genes. The blue lines represent the location of each SNP associated with the expression level of the PTPRN2 gene. The red lines represent SNPs in high confidence eQTL relationships. The table section lists part of the set of SNPs associated with the expression level of PTPRN2, the studies in which each SNP was found, and the population and cell types of that study. The last column gives the number of studies in which that eQTL association was found.

An example of a tissue dependent high-confidence eQTL relationship in liver

There are a total of 28 genes that are found to be involved in eQTL relationship in both studies of liver (Innocenti et al., 2011; Eric E Schadt et al., 2008), but not reported in any other tissue. Although the molecular functions of many of these genes are still unknown, we found a few genes that are involved in lipid metabolism and that are primarily expressed in liver, for example APOC4.

APOC4, apolipoprotein C-IV, encodes a lipid-binding protein that plays a role in lipid metabolism. Several GWA studies have demonstrated that APOC4 gene is associated with the level of blood low-density lipoprotein (LDL) cholesterol and the risk of coronary artery disease (Waterworth et al., 2010; Willer et al., 2008). Figure 4.3 shows that two SNPs, located on the gene region of another gene, CLPTM1, are associated with the expression level of APOC4 (shown as red vertical lines). One SNP, rs11668758, is located in the intron region of CLPTM1and the other, rs3786505, is a synonymous variant. These two SNPs are in the same high-

confidence eQTL relationship. This result illustrates that SNPs associated with the expression level of a gene may be located on another neighboring gene.



Position	SNP	Probe	Study	Population	Cell	LD level	Overlap studies
chr19:45,490,570	rs3786505	APOC4	LV2	CEU	Liver	5	17
chr19:45,474,680	rs11668758	HSG00307895	LV	CEU	Liver	8	2

Figure 4.3. Visualization of a Liver dependent high-confidence eQTL relationships for APOC4.

The red lines show the locations of the eQTL SNPs, rs3786505 and rs11668758. The red boxes represent the transcripts of the APOC4 gene, with which these eQTL SNPs are involved in a high confidence eQTL relationship. The green boxes are the transcripts of the neighboring genes.

An example of a high confidence eQTL relationship associated with human disease

Several marker SNPs, including rs2872507, rs2305480, and rs2290400, in chromosome region of 17q12 , have been identified as associated with the risk of human complex disease, especially immune related diseases, such as Crohn's disease (Barrett et al., 2008; Franke et al., 2010), Rheumatoid arthritis (Okada et al., 2014;

Stahl et al., 2010), Asthma (Bønnelykke et al., 2013; Moffatt et al., 2007), and Type 1 diabetes (Barrett et al., 2009).

Different studies have proposed different disease relevant candidate genes for this locus. For Crohn's disease, GSMDL, ZPBP2, ORMDL3, and IKZF3, were reported. In contrast, only IKZF3 was reported as a candidate for Rheumatoid arthritis, and only ORMDL3 for Asthma and Type 1 diabetes. Based on the eQTL analysis, six genes, GSDMA, GSDMAB, KRT222, ORMDL3, PGAP3, and ZPBP2, are found to have an eQTL association with these marker SNPs. Three of these eQTL genes, KRT222 (Montgomery et al., 2010), ZPBP2 (Grundberg et al., 2012) and PGAP3 (Grundberg et al., 2012) were discovered only in one eQTL study. Two genes, GSDMB and ORMDL3, are in high-confidence eQTL relationships at the highest LD threshold ($r^2 \geq 0.8$) (Figure 4.4). Thus, the eQTL analysis suggests that these two genes are highly possible to be the candidates for involvement in susceptibility to these immune related diseases. Previous studies have shown that changes in the binding of an insulator protein, CTCF, and related chromatin remodeling on this autoimmune associated locus, might lead to alter the cis-regulatory of these two genes (Verlaan et al., 2009). This result demonstrates how the addition of eQTL information can be useful in reducing ambiguities in disease GWA study results.

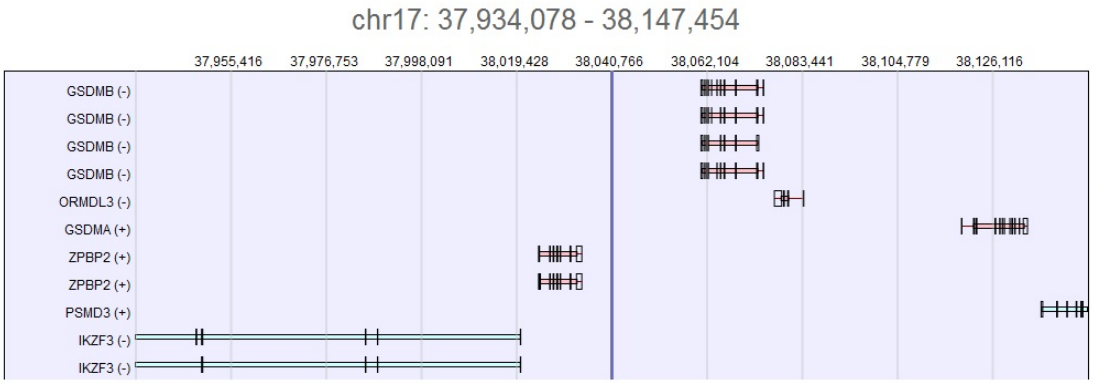


Figure 4.4. Visualization of eQTL relationships for the chromosome region 17q12 that is associated with the risk of Asthma and some autoimmune diseases.
This plot shows the related position of genes in this locus and several of these gene associated with rs2872507. The blue line is the position of rs2872507.

4.4: Conclusion

ExSNP is designed with the intention of being an interactive and user-friendly integrated web database to query and visualize available human eQTL data and consensus eQTLs. It, so far, covers the broadest range of eQTL studies for various cell types and human populations. Users can search for all eQTL data and high-confidence eQTLs by querying with a SNP ID or gene name. The eQTL browser also allows users to navigate and visualize the relative position of all cis-regulatory SNPs for each gene and to recognize what functional effects these eQTLs may be involved in. All analyzed data are available for users to download.

Chapter 5: Conclusions and perspectives

In this dissertation, we demonstrated a way to use expression quantitative trait loci (eQTLs) information for investigating putative expression mechanisms of human common diseases. Here we give a brief summary of the conclusions for our study and discuss future directions in this area.

5.1: High Confidence eQTL sets

In the first part of my dissertation, we overviewed available human genome-wide eQTL data and showed there is a high level of inconsistency among results from genome-wide eQTL association studies. Our objective was then to develop a method for integrating the results from the eQTL studies, so as to identify the high-confidence eQTLs. We integrated data from 16 publicly available genome-wide eQTL studies covering various human tissues and populations, and found consensus SNP-gene associations across these studies. We also compared eQTLs across different tissues and populations so as to estimate the proportions of tissue-dependent and population-dependent relationships. In order to help understand mechanisms underlying these eQTL associations, we mapped eQTLs to annotated functional elements, discovering two enrichments of tissue-specific transcription factor binding sites.

5.2: Disease associated eQTLs

In the second part of my dissertation, we used the high-confidence eQTL data identified from genome-wide eQTL studies among various human tissues and populations to identify which loci associated with a set of common diseases may have

an underlying expression mechanism contributing to disease susceptibility. We developed an algorithm using centiMorgan distance, instead of conventional Linkage disequilibrium (R^2), to estimate the overlap of disease associated loci and eQTL relationships. For that purpose, we used data from the WTCCC1 study to establish thresholds of centiMorgan distance and then applied the analysis to all associated loci for seven human common diseases. In the end, we identified a large number of disease loci containing high-confidence eQTL relationships. We not only re-discovered many genes that had previously been suggested to have altered expression contributing to disease susceptibility, but were also able to propose expression mechanisms for many genes previously suggested as disease relevant but for which no mechanisms has been proposed, as well as finding a set of new candidate genes for disease involvement, based on expression relationships.

5.3: Web-based resource

In the third part of my dissertation, we introduced a comprehensive web-based database, ExSNP, which incorporates all the analysis data from our studies, including original eQTLs, high-confidence eQTLs, cell type-dependent eQTLs, population-dependent eQTLs, disease associated eQTLs, and functionally annotated eQTLs. The ExSNP resource not only facilitates the querying of these eQTL data but also provides a comprehensive genome browser to visualize the relative positions of SNPs to their associated genes and other neighboring genes.

5.4: Future perspectives

Currently more and more studies are performing eQTL mapping by applying next-generation sequencing technology, RNA-Seq, and whole genome sequencing, using larger sample populations. This will provide a relatively accurate estimate for the expression levels of different transcripts and much more complete SNP genotypes. There is a need to develop a better statistical model to reliably identify significant eQTLs, for example, incorporating other information such as pathway relationships.

Undoubtedly, hundreds of thousands eQTLs in various tissues and populations will be identified in the near future. Currently the GTEx project (www.gtexportal.org/home/), funded by NHGRI, is performing the collection of eQTL studies towards an end goal of 900 donors and around 20,000 tissues samples. This resource will enable studies of expression quantitative trait loci (eQTLs), alternative splicing, and the tissue specificity of gene regulatory mechanisms, and aid in the interpretation of genome-wide association studies (GWAS).

Establishing the presence of eQTL in a disease locus is indirect evidence of the disease relevance of the associated expression change. Further evidence is required. The eQTL relationships do not provide direct explanations of the underlying mechanisms of human common traits. Therefore, it is necessary to perform additional experiments to validate the eQTLs associated with human common traits. For example, one could examine the differences in gene expression level in an appropriate cell type between disease case and control populations. More information

on functional elements, such as enhancers and DNA methylation sites, would also facilitate understanding of the mechanisms that give rise to the eQTLs.

In the post-GWAS era, the greatest challenge is to combine GWAS findings with additional molecular data to functionally characterize the associations. Since the etiologies of human common diseases are complex, no single molecular analysis is expected to fully unravel the disease mechanism. Multiple molecular levels may interact according to different physiological conditions, cell types and disease stages. As the various 'Omics' techniques advance, there are increasing size and complexity of high-throughput data, including Transcriptomics, Proteomics, Metabolomics, and Epigenomics (Bauer, Glintschert, & Schuchhardt, 2014; Di Girolamo, Lante, Muraca, & Putignani, 2013). This information makes it possible to investigate the effect of risk variants on multiple molecular levels. Therefore, the next step is to develop new integrative approaches/algorithms that can combine 'Omics' data from these different molecular levels and prior knowledge of pathways and ontologies to facilitate the deduction of causal processes from gene-disease associations from GWAS.

Appendix

Figure S1. Hierarchical cluster

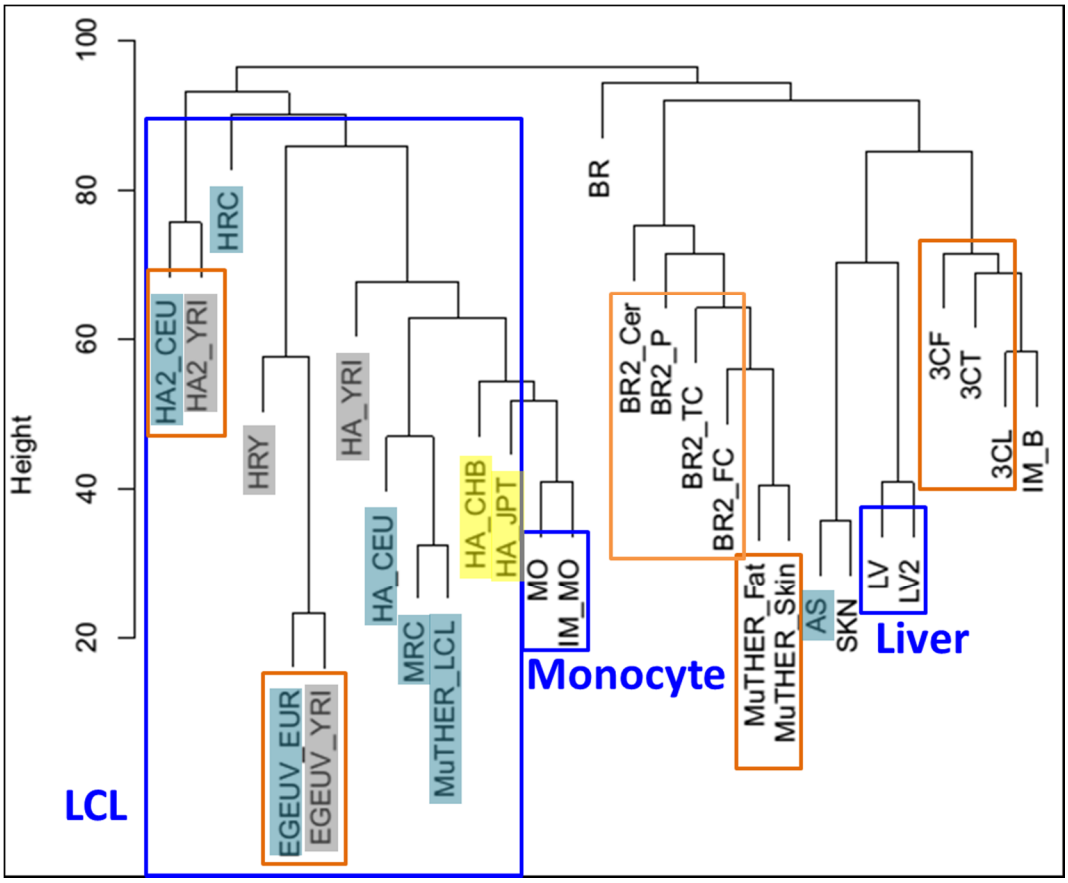


Figure S1.Hierarchical clustering of the fraction of common exGenes between pairs of eQTL datasets. Distance scale is based on the % of common exGenes between pairs of datasets. Blue boxes surround clustered datasets conducted in the same cell type, for example, LCL, monocytes, and liver. Orange boxes are for clustered datasets from the same study but conducted in different cell types or populations. The blue highlighted datasets are those performed in LCL for Caucasian populations. The grey highlighted datasets are those performed in LCL for African populations. The yellow highlighted datasets are those performed in LCL for Asian populations.

Method for the simulation of High-confidence eQTL datasets

To generate the pseudo high-confidence eQTL data for each integrated set, we simulated a 1000 random versions of the eQTL data for each dataset. There are a total 11 eQTL studies included in this simulation (Table S1). For each dataset of each eQTL study, we randomly generated 'NULL' eQTL association data (exSNP-exGene pairs) as follows. For each exGene with one or more associated exSNPs, we first randomly selected a pseudo exGene from the set of all RefGen genes and then randomly selected the same number of SNPs as there are exSNPs for the original exGene. These pseudo exSNPs were selected from the set of SNPs within 1MB of the selected gene, with a probability derived from the distribution of the distances between exSNPs and the associated exGenes. (Table S2).

We continually generated pseudo eQTL associations until we had simulated the number of exGenes and exSNPs in each dataset, and repeated the procedure 1000 times. For each run of the random simulation, we then calculated the average number of high-confidence eQTL associations, SNPs, and genes by using the same high-confidence algorithm (Table S3). Assuming the distribution of the number of simulated high-confidence eQTL data follows a normal distribution, we calculated the P-value of the number of the observed high-confidence eQTL data and found that all these P-values are too small to calculate.

Table S1. Microarray Chip used in each study

Study	Genotype Microarray
HA	HapMap Array
BR	Affymetrix GeneChip Human Mapping 500K
AS	Sentrix Human-1 & Illumina HumanHap300

LV	Illumina 650Y & Affymetrix 500K
HA2	HapMap Array
3C	Illumina 550K
MO	Affymetrix 6.0
HRC	HapMap Array
HRY	HapMap Array
BR2	Illumina HumanHap 550
LV2	Illumina 610 (UC) & Illumina HumanHap 550 (UW)

Table S2. Distribution of exSNP –exGene distances (in base pair units) for all cis-eQTL associations in all 11 studies

Range	Counts	Frequency
-1000000 ~ -950000	104	0.001
-950000 ~ -900000	128	0.002
-900000 ~ -850000	145	0.002
-850000 ~ -800000	136	0.002
-800000 ~ -750000	145	0.002
-750000 ~ -700000	184	0.003
-700000 ~ -650000	215	0.003
-650000 ~ -600000	300	0.004
-600000 ~ -550000	266	0.004
-550000 ~ -500000	306	0.004
-500000 ~ -450000	322	0.004
-450000 ~ -400000	440	0.006
-400000 ~ -350000	591	0.008
-350000 ~ -300000	813	0.011
-300000 ~ -250000	1144	0.016
-250000 ~ -200000	1358	0.018
-200000 ~ -150000	2128	0.029
-150000 ~ -100000	3260	0.044
-100000 ~ -50000	5314	0.072
-50000 ~ 0	12134	0.165
0	18675	0.254
0 ~ 50000	11699	0.159
50000 ~ 100000	4599	0.063
100000 ~ 150000	2744	0.037
150000 ~ 200000	1639	0.022
200000 ~ 250000	1033	0.014
250000 ~ 300000	769	0.010
300000 ~ 350000	537	0.007
350000 ~ 400000	438	0.006
400000 ~ 450000	324	0.004
450000 ~ 500000	216	0.003
500000 ~ 550000	187	0.003
550000 ~ 600000	163	0.002
600000 ~ 650000	156	0.002
650000 ~ 700000	181	0.002
700000 ~ 750000	138	0.002
750000 ~ 800000	157	0.002
800000 ~ 850000	152	0.002
850000 ~ 900000	105	0.001
900000 ~ 950000	133	0.002
950000 ~ 1000000	96	0.001

Table S3. Data for the randomly simulated high-confidence (HC) eQTL data.

	AllCell AllPop	LCL CEU	LCL ASN	LCL YRI	LCL	Liver	Brain
HC-eQTL relationships							
Min.	285	0	0	0	34	0	6
1stQ	620	35	1	0	153	0	111
Median	754.5	67	6.5	0	224.5	0	165
Mean	795.1	131.7	28.75	0.86	291.1	0.23	176.7
3rdQ	922	133	20	0	351	0	232
Max.	2430	1234	443	29	1559	4	584
Observed	**26626***	**6079***	**3028***	**774***	**11704***	**507***	**5068***
HC-eQTL SNPs							
Min.	285	0	0	0	34	0	6
1stQ	617	35	1	0	153	0	111
Median	750.5	67	6.5	0	224.5	0	165
Mean	791.4	131.4	28.75	0.86	290.6	0.23	176.4
3rdQ	914.2	133	20	0	351	0	231.2
Max.	2366	1190	443	29	1559	4	584
Observed	**23741***	**5849***	**2985***	**773***	**11312***	**502***	**4779***
HC-eQTL genes							
Min.	47	0	0	0	8	0	2
1stQ	62	5	1	0	18	0	8
Median	67	6	1	0	21	0	10
Mean	67.25	6.63	1.49	0.27	21.24	0.2	9.754
3rdQ	72	8	2	0	24	0	12
Max.	90	15	7	3	35	3	19
Observed	**1685***	**277***	**134***	**105***	**595***	**390***	**134***

* indicates the significant P-value. For each integrated dataset, the median and mean number of HC-eQTL relationships, HC-eQTL SNPs, and HC-eQTL genes found is given as well as the minimum, maximum, and 1st and 3rd quantile values found in the1000 simulations. Bold numbers are the observed in the real data.

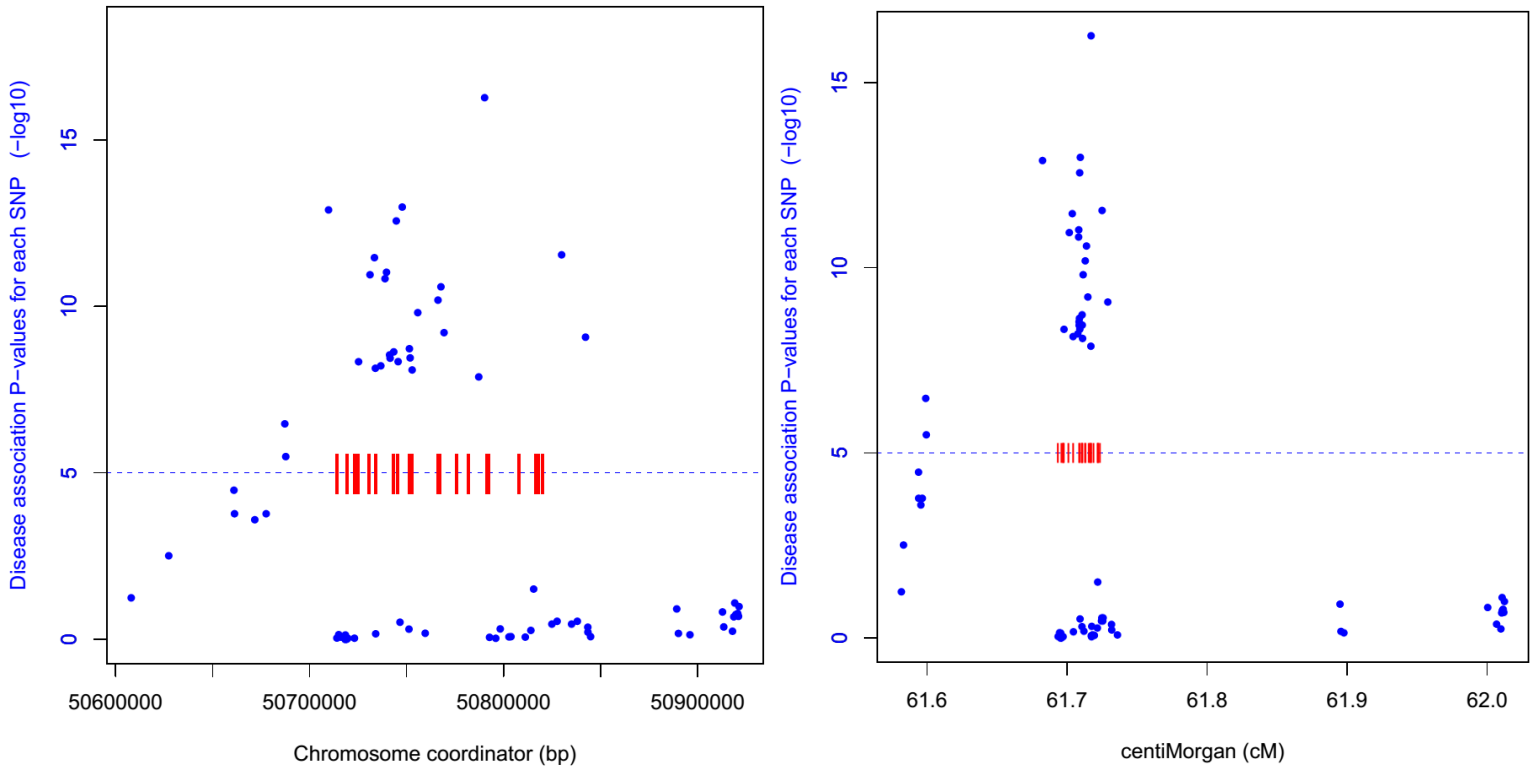


Figure S2. Additional examples of eQTL and disease data for WTCCC1 loci: Crohn's Disease risk locus, Chr16, eQTL gene: NOD2 (nucleotide-binding oligomerization domain containing 2)

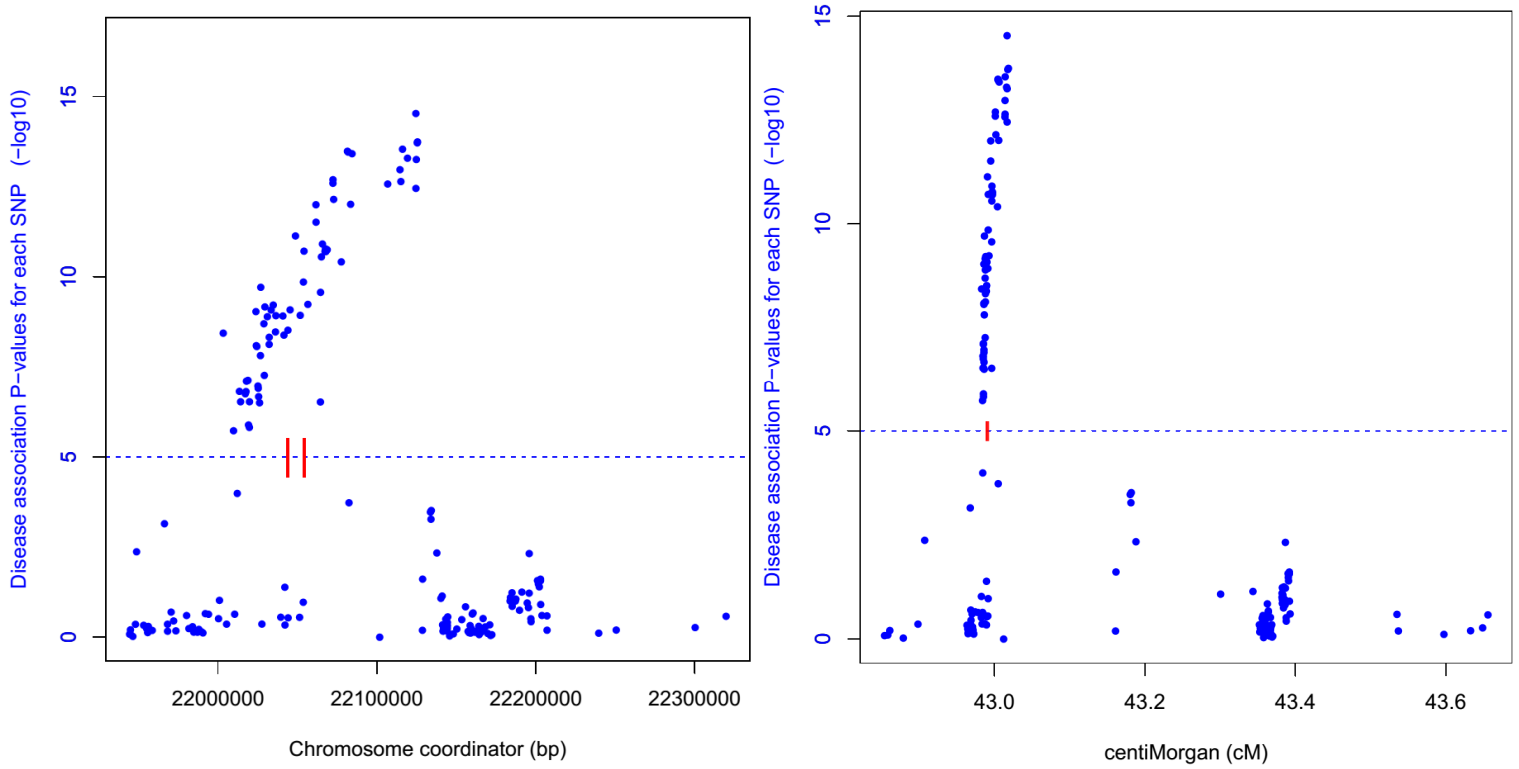


Figure S3. Additional examples of eQTL and disease data for WTCCC1 loci: Coronary artery disease risk locus, Chr 9, eQTL gene: CDKN2B (cyclin-dependent kinase inhibitor 2B)

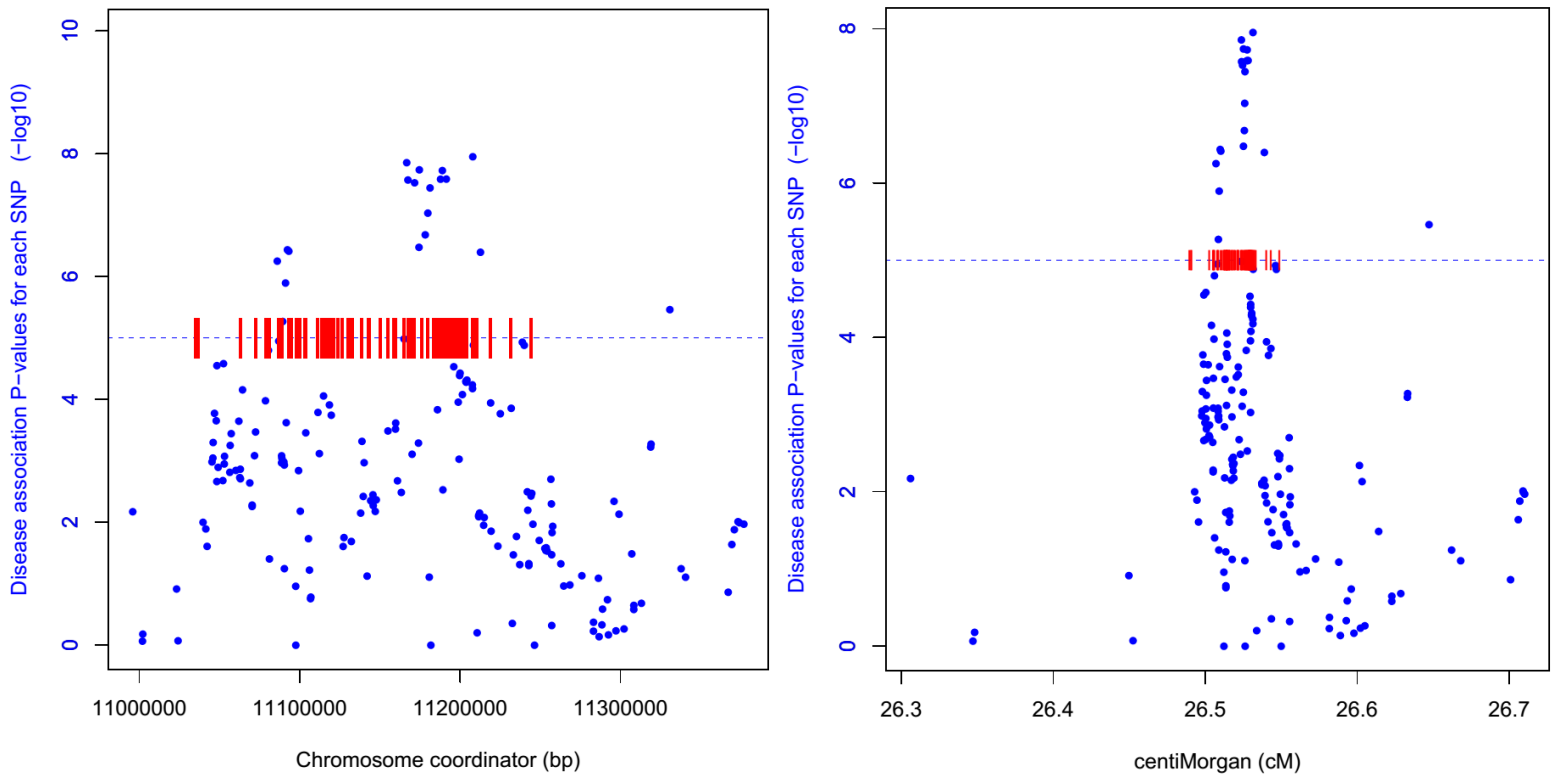


Figure S4. Additional examples of eQTL and disease data for WTCCC1 loci: Type 1 diabetes risk locus, Chr16, eQTL gene: DEXI (Dexi homolog)

Table S4. WTCCC and included follow-up GWA studies for seven human common diseases

Disease Set	Disease/Trait	PUBMED ID	Loci count
BD	Bipolar disorder	22925353, 21771265, 21926972, 18711365, 19416921, 21738484, 19488044, 21353194, 17554300, 17486107, 22688191	65
CAD	Coronary artery disease	17634449, 19198611, 21378990, 19198612, 22319020, 21088011, 21239051, 21606135, 17554300, 21378988	45
CD	Crohn's disease	17684544, 20570966, 17554261, 21102463, 18587394, 17554300, 22936669, 17435756, 23128233, 22293688, 17804789, 17447842	84
HT	Hypertension	19430479, 19304780, 17554300, 21909115, 21082022	17
RA	Rheumatoid arthritis	17804836, 20453842, 19503088, 18668548, 21653640, 17554300, 17982456, 18794853	34
T1D	Type 1 diabetes	17632545, 19966805, 19430480, 18840781, 18978792, 17554260, 17554300, 22293688, 21980299, 18198356	51
T2D	Type 2 diabetes	17293876, 19056611, 20581827, 20418489, 18372903, 17668382, 17554300, 17463246, 17463248, 22693455, 17463249, 22101970, 22293688, 17460697	43

Table S5. exGenes associated with high-confident eQTLs in the AllCell_AllPop integrated set at a 0.05 cM threshold for the WTCCC1 seven diseases.

Disease	Chromosome	Loci	exGenes
BD	2	2q11.2	CIAO1, LIPT1, TSGA10, UNC50
BD	2	2q37.3	ANKMY1
BD	3	3p21.1	GLT8D1, GNL3, ITIH4, NT5DC2, TMEM110
BD	3	3p22	CMTM8, LRRFIP2, TRANK1
BD	3	3q27	MCCC1
BD	4	4q22	PPM1K
BD	5	5q15	ANKRD32, MCTP1
BD	7	7p22	MAD1L1
BD	8	8q24.3	SLC45A4
BD	9	9p13	NUDT2
BD	9	9p22	TTC39B
BD	9	9q33	RALGPS1, RPL12, SLC2A8, ZNF79
BD	11	11q13.2	CCS, CTSF, LRFN4, RIN1
BD	11	11q24	SPA17
BD	12	12q13.1	CACNB3
BD	12	12q23	CMKLR1
BD	14	14q11.2	HNRNPC, RPGRIP1
BD	14	14q32.3	TDRD9
BD	15	15q14	C15orf53
BD	15	15q25	CTSH
BD	16	16p12	COG7, DCTN5, GGA2
BD	19	19p13.1	ATP13A1, LPAR2, MAU2
BD	19	19q13.1	LOC400684
BD	19	19q13.2	RABAC1

BD	20	20p13	CDC25B
BD	20	20q13.1	PARD6B
CAD	1	1p13.3	CELSR2, PSRC1, SORT1
CAD	1	1q21.3	UBE2Q1
CAD	2	2q33	FAM117B
CAD	3	3p25	ANKRD28, COLQ
CAD	3	3q22	CEP70, FAIM
CAD	6	6p21.3	CCHCR1, DDR1, DEF6, DPCR1, HCG22, HCG27, HLA-B, HLA-C, HLA-DQB1, LST1, MICB, TCF19, TCP11, UHRF1BP1, VARS2
CAD	6	6q14	FAM46A
CAD	6	6q25.3	SLC22A3
CAD	7	7q22	GPR22
CAD	9	9p21	CDKN2B
CAD	9	9q34.2	SURF1, SURF6
CAD	10	10q23.3	LIPA
CAD	10	10q24.3	AS3MT, C10orf26, C10orf32, NT5C2, SFXN2, USMG5
CAD	11	11q22.3	PDGFD
CAD	11	11q23.3	TAGLN
CAD	11	11q24	FOXRED1, ST3GAL4
CAD	12	12q24.1	FAM109A, SH2B3, TMEM116
CAD	12	12q24.31	C12orf43, SPPL3
CAD	15	15q25	ADAMTS7, CTSH
CAD	17	17p11.2	PEMT, RASD1
CAD	17	17p13	SRR
CAD	17	17q21.3	ATP5G1, UBE2Z
CAD	19	19p13.2	SMARCA4
CD	1	1p13.2	AP4B1
CD	1	1p31.1	DNAJB4, GIPC2, NEXN

CD	1	1p31.3	SLC35D1
CD	1	1p36.2	PER3
CD	1	1q22	ADAM15, MUC1, RIT1
CD	1	1q23	CD244, LY9, SLAMF7
CD	1	1q32.1	IL19
CD	2	2p16	AHSA2, KIAA1841, LOC339803, PUS10
CD	2	2p23	C2orf28, GPN1, KRTCAP3, SLC5A6
CD	2	2q37.1	DGKD, SP110, SP140
CD	3	3p21.3	AMT, HYAL3, IP6K2, KLHDC8B, NCKIPSD, NICN1, P4HTM, RBM6, UBA7, USP4, WDR6
CD	4	4q24	BANK1
CD	5	5p13.1	PTGER4
CD	5	5q15	ERAP1, ERAP2, LNPEP
CD	5	5q31.1	PDLIM4, SLC22A4, SLC22A5
CD	5	5q31.3	FGF1, NDFIP1
CD	5	5q35.2	CPEB4
CD	6	6p21.3	AIF1, ATP6V1G2, CCHCR1, CLIC1, CSNK2B, DOM3Z, GPANK1, HCG22, HCG27, HLA-B, HLA-C, HLA-DOB, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB4, HLA-DRB5, HSPA1B, LST1, LY6G5C, MICB, PRRC2A, TAP2, TCF19, VARS2
CD	6	6q22.3	THEMIS
CD	6	6q25.3	RSPH3
CD	6	6q27	FGFR1OP, RNASET2, RPS6KA2
CD	7	7p15.2	SKAP2
CD	8	8q21.3	RIPK2
CD	9	9p24	JAK2
CD	9	9q34.3	CARD9, DNLZ, INPP5E, SDCCAG3
CD	10	10p11.2	CREM

CD	10	10q21.1	CISD1
CD	10	10q21.2	ADO
CD	11	11q12	C11orf10, FADS1, FADS2
CD	11	11q13.1	CCDC88B, FKBP2, PRDX5, RPS6KA4, TRMT112, TRPT1
CD	11	11q13.5	C11orf30
CD	12	12q12	SLC2A13
CD	13	13q14.1	LACC1, TNFSF11
CD	14	14q24.1	ZFP36L1
CD	14	14q31	GALC, GPR65
CD	15	15q14	RASGRP1
CD	16	16p11.2	APOBR, CCDC101, EIF3C, SPNS1, TUFM
CD	16	16q12.1	NOD2, SNX20
CD	17	17q12	GSDMA, GSDMB, ORMDL3, ZPBP2
CD	17	17q21.2	CNTNAP1
CD	19	19p13.2	CDC37, ICAM3, ICAM4
CD	19	19p13.3	GPX4
CD	20	20q13.3	STMN3
CD	21	21q22.1	GART, IFNGR2, ITSN1, TMEM50B
CD	22	22q11.2	UBE2L3
CD	22	22q12.2	MTMR3, UQCR10
CD	22	22q13.1	PDGFB, SYNGR1
CD	22	22q13.2	CCDC134, EP300, L3MBTL2, MEI1, PMM1
HT	1	1p13.2	ST7L
HT	4	4q24	SLC39A8
HT	6	6p21.3	AIF1, ATP6V1G2, CSNK2B, DOM3Z, GPANK1, HCG27, HLA-C, HLA-DQA1, HLA-DQB1, HLA-DRB1, HLA-DRB5, HSPA1B, LST1, LY6G5C, MICB
HT	6	6p22.2	BTN3A2, BTN3A3, HIST1H2BD, TRIM38
HT	8	8p23.1	XKR6

HT	12	12q21.3	GALNT4
HT	15	15q26.1	FES
HT	20	20q13.3	CTSZ, TH1L
RA	1	1p13.2	AP4B1
RA	1	1p36.3	MMEL1, TNFRSF14
RA	2	2p14	SPRED2
RA	2	2p16	AHSA2, LOC339803, PUS10
RA	2	2q11.2	AFF3
RA	4	4p15.2	ANAPC4, ZCCHC4
RA	5	5q21	PAM, PPIP5K2
RA	6	6p21.3	HCG22, HCG27, HLA-B, HLA-C, HLA-DOB, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRA, HLA-DRB1, HLA-DRB4, HLA-DRB5, HLA-DRB6, LOC285835, LST1, MICB, PRRC2A, PSMB9, TAP2, VARS2
RA	6	6p22.1	BTN3A2, GABBR1, HCG4, HCG4B, HLA-A, HLA-F, HLA-G, VARS2, ZFP57, ZNRD1
RA	6	6q27	FGFR1OP, RNASET2, RPS6KA2
RA	7	7q32	IRF5, TNPO3
RA	8	8p23.1	BLK, FAM167A, FDFT1, XKR6
RA	9	9p13	NUDT2
RA	9	9q33	C5, GSN, MEGF9
RA	12	12q13.3	METTL21B, TSFM
RA	12	12q24.1	FAM109A, SH2B3, TMEM116
RA	14	14q24.3	BATF
RA	17	17q12	GSDMA, GSDMB, ORMDL3, ZPBP2
RA	20	20q13.1	CD40, PLTP
RA	21	21q22.3	UBASH3A
RA	22	22q12.3	IL2RB
T1D	1	1p13.2	AP4B1

T1D	1	1p31.3	PGM1
T1D	1	1q32.1	IL19
T1D	2	2p23	ADCY3, POMC
T1D	2	2q11.2	AFF3
T1D	6	6p21.3	HCG27, HLA-DOB, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB4, HLA-DRB5, PRRC2A, TAP2
T1D	7	7p15.2	SKAP2
T1D	8	8q24.1	TNFRSF11B
T1D	12	12p13.3	CLEC2B, CLEC2D, CLECL1
T1D	12	12q13.2	RPS26, SPRYD4, STAT2, SUOX
T1D	12	12q24.1	ALDH2, FAM109A, SH2B3, TMEM116
T1D	14	14q24.1	ZFP36L1
T1D	15	15q14	RASGRP1
T1D	15	15q25	CTSH
T1D	16	16p11.2	APOBR, CCDC101, EIF3C, SPNS1, TUFM
T1D	16	16p13.1	DEXI, RMI2
T1D	16	16q23	CFDP1
T1D	17	17q12	GSDMA, GSDMB, ORMDL3, ZPBP2
T1D	17	17q21.2	CCR7, SMARCE1
T1D	19	19p13.2	CDC37, ICAM3, ICAM4
T1D	19	19q13.3	FKRP, PRKD2, STRN4
T1D	21	21q22.3	UBASH3A
T1D	22	22q12.2	MTMR3, UQCR10
T1D	22	22q12.3	C1QTNF6
T2D	2	2q36	IRS1
T2D	3	3p25	PPARG
T2D	4	4q27	CCNA2, EXOSC9
T2D	6	6p21.3	HLA-DOB, HLA-DPB1, HLA-DQA1, HLA-DQA2, HLA-DQB1, HLA-DQB2, HLA-

			DRB1, HLA-DRB5, TAP2
T2D	8	8q22.1	TP53INP1
T2D	9	9p21	CDKN2B
T2D	10	10p13	CAMK1D
T2D	11	11p15.1	B7H6, NUCB2, SNORD14A
T2D	11	11q13.4	STARD10
T2D	12	12q13.1	CERS5, DIP2B, SLC11A2, TFCP2
T2D	12	12q24.31	C12orf43
T2D	15	15q24	PSTPIP1, TSPAN3
T2D	15	15q26.1	HDDC3, RCCD1, UNC45A

Table S6. Categories of the proposed eQTL related disease candidate genes.

Disease	Chromosome	Locus	eQTL gene	Category
BD	2	2q11.2	CIAO1	C
BD	2	2q11.2	LIPT1	B
BD	2	2q11.2	TSGA10	B
BD	2	2q11.2	UNC50	C
BD	2	2q37.3	ANKMY1	C
BD	3	3p21.1	GLT8D1	B
BD	3	3p21.1	GNL3	B
BD	3	3p21.1	ITIH4	B
BD	3	3p21.1	NT5DC2	B
BD	3	3p21.1	TMEM110	B
BD	3	3p22	CMTM8	C
BD	3	3p22	LRRFIP2	C
BD	3	3p22	TRANK1	C
BD	3	3q27	MCCC1	C
BD	4	4q22	PPM1K	C
BD	5	5q15	ANKRD32	C
BD	5	5q15	MCTP1	B
BD	7	7p22	MAD1L1	B
BD	8	8q24.3	SLC45A4	C
BD	9	9p13	NUDT2	C
BD	9	9p22	TTC39B	B
BD	9	9q33	RALGPS1	C
BD	9	9q33	RPL12	C
BD	9	9q33	SLC2A8	C
BD	9	9q33	ZNF79	C
BD	11	11q13.2	CCS	B
BD	11	11q13.2	CTSF	B
BD	11	11q13.2	LRFN4	B
BD	11	11q13.2	RIN1	B
BD	11	11q24	SPA17	C
BD	12	12q13.1	CACNB3	A
BD	12	12q23	CMKLR1	B
BD	14	14q11.2	HNRNPC	B
BD	14	14q11.2	RPGRIP1	B
BD	14	14q32.3	TDRD9	C
BD	15	15q14	C15orf53	B
BD	15	15q25	CTSH	B
BD	16	16p12	COG7	C
BD	16	16p12	DCTN5	B
BD	16	16p12	GGA2	C

BD	19	19p13.1	ATP13A1	C
BD	19	19p13.1	LPAR2	C
BD	19	19p13.1	MAU2	C
BD	19	19q13.1	LOC400684	C
BD	19	19q13.2	RABAC1	C
BD	20	20p13	CDC25B	C
BD	20	20q13.1	PARD6B	B
CAD	1	1p13.3	CELSR2	A
CAD	1	1p13.3	PSRC1	A
CAD	1	1p13.3	SORT1	A
CAD	1	1q21.3	UBE2Q1	C
CAD	2	2q33	FAM117B	C
CAD	3	3p25	ANKRD28	C
CAD	3	3p25	COLQ	C
CAD	3	3q22	CEP70	C
CAD	3	3q22	FAIM	C
CAD	6	6q14	FAM46A	C
CAD	6	6q25.3	SLC22A3	B
CAD	7	7q22	GPR22	C
CAD	9	9p21	CDKN2B	A
CAD	9	9q34.2	SURF1	C
CAD	9	9q34.2	SURF6	C
CAD	10	10q23.3	LIPA	B
CAD	10	10q24.3	AS3MT	C
CAD	10	10q24.3	C10orf26	C
CAD	10	10q24.3	C10orf32	C
CAD	10	10q24.3	NT5C2	B
CAD	10	10q24.3	SFXN2	C
CAD	10	10q24.3	USMG5	C
CAD	11	11q22.3	PDGFD	B
CAD	11	11q23.3	TAGLN	C
CAD	11	11q24	FOXRED1	C
CAD	11	11q24	ST3GAL4	C
CAD	12	12q24.1	FAM109A	C
CAD	12	12q24.1	SH2B3	B
CAD	12	12q24.1	TMEM116	C
CAD	12	12q24.31	C12orf43	B
CAD	12	12q24.31	SPPL3	C
CAD	15	15q25	ADAMTS7	B
CAD	15	15q25	CTSH	C
CAD	17	17p11.2	PEMT	B
CAD	17	17p11.2	RASD1	B
CAD	17	17p13	SRR	B

CAD	17	17q21.3	ATP5G1	B
CAD	17	17q21.3	UBE2Z	B
CAD	19	19p13.2	SMARCA4	C
CD	1	1p13.2	AP4B1	C
CD	1	1p31.1	DNAJB4	C
CD	1	1p31.1	GIPC2	C
CD	1	1p31.1	NEXN	C
CD	1	1p31.3	SLC35D1	C
CD	1	1p36.2	PER3	B
CD	1	1q22	ADAM15	A
CD	1	1q22	MUC1	A
CD	1	1q22	RIT1	B
CD	1	1q23	CD244	B
CD	1	1q23	LY9	C
CD	1	1q23	SLAMF7	C
CD	1	1q32.1	IL19	B
CD	2	2p16	AHSA2	C
CD	2	2p16	KIAA1841	C
CD	2	2p16	LOC339803	C
CD	2	2p16	PUS10	C
CD	2	2p23	C2orf28	C
CD	2	2p23	GPN1	B
CD	2	2p23	KRTCAP3	C
CD	2	2p23	SLC5A6	C
CD	2	2q37.1	DGKD	C
CD	2	2q37.1	SP110	C
CD	2	2q37.1	SP140	B
CD	3	3p21.3	AMT	C
CD	3	3p21.3	HYAL3	C
CD	3	3p21.3	IP6K2	C
CD	3	3p21.3	KLHDC8B	C
CD	3	3p21.3	NCKIPSD	C
CD	3	3p21.3	NICN1	C
CD	3	3p21.3	P4HTM	C
CD	3	3p21.3	RBM6	C
CD	3	3p21.3	UBA7	C
CD	3	3p21.3	USP4	C
CD	3	3p21.3	WDR6	C
CD	4	4q24	BANK1	C
CD	5	5p13.1	PTGER4	A
CD	5	5q15	ERAP1	C
CD	5	5q15	ERAP2	B
CD	5	5q15	LNPEP	B

CD	5	5q31.1	PDLIM4	C
CD	5	5q31.1	SLC22A4	B
CD	5	5q31.1	SLC22A5	B
CD	5	5q31.3	FGF1	C
CD	5	5q31.3	NDFIP1	B
CD	5	5q35.2	CPEB4	B
CD	6	6q22.3	THEMIS	C
CD	6	6q25.3	RSPH3	C
CD	6	6q27	FGFR1OP	B
CD	6	6q27	RNASET2	B
CD	6	6q27	RPS6KA2	C
CD	7	7p15.2	SKAP2	C
CD	8	8q21.3	RIPK2	B
CD	9	9p24	JAK2	B
CD	9	9q34.3	CARD9	B
CD	9	9q34.3	DNLZ	C
CD	9	9q34.3	INPP5E	C
CD	9	9q34.3	SDCCAG3	C
CD	10	10p11.2	CREM	B
CD	10	10q21.1	CISD1	C
CD	10	10q21.2	ADO	C
CD	11	11q12	C11orf10	C
CD	11	11q12	FADS1	B
CD	11	11q12	FADS2	C
CD	11	11q13.1	CCDC88B	C
CD	11	11q13.1	FKBP2	C
CD	11	11q13.1	PRDX5	B
CD	11	11q13.1	RPS6KA4	C
CD	11	11q13.1	TRMT112	C
CD	11	11q13.1	TRPT1	C
CD	11	11q13.5	C11orf30	B
CD	12	12q12	SLC2A13	C
CD	13	13q14.1	LACC1	B
CD	13	13q14.1	TNFSF11	B
CD	14	14q24.1	ZFP36L1	B
CD	14	14q31	GALC	B
CD	14	14q31	GPR65	B
CD	15	15q14	RASGRP1	B
CD	16	16p11.2	APOBR	C
CD	16	16p11.2	CCDC101	C
CD	16	16p11.2	EIF3C	B
CD	16	16p11.2	SPNS1	C
CD	16	16p11.2	TUFM	C

CD	16	16q12.1	NOD2	A
CD	16	16q12.1	SNX20	C
CD	17	17q12	GSDMA	C
CD	17	17q12	GSDMB	C
CD	17	17q12	ORMDL3	B
CD	17	17q12	ZPBP2	B
CD	17	17q21.2	CNTNAP1	C
CD	19	19p13.2	CDC37	C
CD	19	19p13.2	ICAM3	B
CD	19	19p13.2	ICAM4	C
CD	19	19p13.3	GPX4	B
CD	20	20q13.3	STMN3	C
CD	21	21q22.1	GART	B
CD	21	21q22.1	IFNGR2	B
CD	21	21q22.1	ITSN1	C
CD	21	21q22.1	TMEM50B	B
CD	22	22q11.2	UBE2L3	B
CD	22	22q12.2	MTMR3	B
CD	22	22q12.2	UQCR10	C
CD	22	22q13.1	PDGFB	C
CD	22	22q13.1	SYNGR1	C
CD	22	22q13.2	CCDC134	C
CD	22	22q13.2	EP300	B
CD	22	22q13.2	L3MBTL2	C
CD	22	22q13.2	MEI1	C
CD	22	22q13.2	PMM1	C
HT	1	1p13.2	ST7L	B
HT	4	4q24	SLC39A8	B
HT	6	6p22.2	BTN3A2	C
HT	6	6p22.2	BTN3A3	C
HT	6	6p22.2	HIST1H2BD	C
HT	6	6p22.2	TRIM38	C
HT	8	8p23.1	XKR6	C
HT	12	12q21.3	GALNT4	C
HT	15	15q26.1	FES	B
HT	20	20q13.3	CTSZ	C
HT	20	20q13.3	TH1L	C
RA	1	1p13.2	AP4B1	C
RA	1	1p36.3	MMEL1	B
RA	1	1p36.3	TNFRSF14	B
RA	2	2p14	SPRED2	B
RA	2	2p16	AHSA2	C
RA	2	2p16	LOC339803	C

RA	2	2p16	PUS10	C
RA	2	2q11.2	AFF3	B
RA	4	4p15.2	ANAPC4	C
RA	4	4p15.2	ZCCHC4	C
RA	5	5q21	PAM	C
RA	5	5q21	PPIP5K2	C
RA	6	6q27	FGFR1OP	C
RA	6	6q27	RNASET2	C
RA	6	6q27	RPS6KA2	C
RA	7	7q32	IRF5	B
RA	7	7q32	TNPO3	C
RA	8	8p23.1	BLK	B
RA	8	8p23.1	FAM167A	C
RA	8	8p23.1	FDFT1	C
RA	8	8p23.1	XKR6	C
RA	9	9p13	NUDT2	C
RA	9	9q33	C5	B
RA	9	9q33	GSN	C
RA	9	9q33	MEGF9	C
RA	12	12q13.3	METTL21B	C
RA	12	12q13.3	TSFM	B
RA	12	12q24.1	FAM109A	C
RA	12	12q24.1	SH2B3	B
RA	12	12q24.1	TMEM116	C
RA	14	14q24.3	BATF	B
RA	17	17q12	GSDMA	C
RA	17	17q12	GSDMB	C
RA	17	17q12	ORMDL3	C
RA	17	17q12	ZPBP2	C
RA	20	20q13.1	CD40	A
RA	20	20q13.1	PLTP	C
RA	21	21q22.3	UBASH3A	B
RA	22	22q12.3	IL2RB	B
T1D	1	1p13.2	AP4B1	C
T1D	1	1p31.3	PGM1	B
T1D	1	1q32.1	IL19	C
T1D	2	2p23	ADCY3	B
T1D	2	2p23	POMC	B
T1D	2	2q11.2	AFF3	B
T1D	7	7p15.2	SKAP2	C
T1D	8	8q24.1	TNFRSF11B	B
T1D	12	12p13.3	CLEC2B	C
T1D	12	12p13.3	CLEC2D	C

T1D	12	12p13.3	CLECL1	C
T1D	12	12q13.2	RPS26	C
T1D	12	12q13.2	SPRYD4	C
T1D	12	12q13.2	STAT2	C
T1D	12	12q13.2	SUOX	C
T1D	12	12q24.1	ALDH2	C
T1D	12	12q24.1	FAM109A	C
T1D	12	12q24.1	SH2B3	A
T1D	12	12q24.1	TMEM116	C
T1D	14	14q24.1	ZFP36L1	C
T1D	15	15q14	RASGRP1	B
T1D	15	15q25	CTSH	A
T1D	16	16p11.2	APOBR	C
T1D	16	16p11.2	CCDC101	C
T1D	16	16p11.2	EIF3C	C
T1D	16	16p11.2	SPNS1	C
T1D	16	16p11.2	TUFM	C
T1D	16	16p13.1	DEXI	C
T1D	16	16p13.1	RMI2	C
T1D	16	16q23	CFDP1	C
T1D	17	17q12	GSDMA	C
T1D	17	17q12	GSDMB	C
T1D	17	17q12	ORMDL3	B
T1D	17	17q12	ZPBP2	C
T1D	17	17q21.2	CCR7	C
T1D	17	17q21.2	SMARCE1	C
T1D	19	19p13.2	CDC37	C
T1D	19	19p13.2	ICAM3	C
T1D	19	19p13.2	ICAM4	C
T1D	19	19q13.3	FKRP	C
T1D	19	19q13.3	PRKD2	C
T1D	19	19q13.3	STRN4	C
T1D	21	21q22.3	UBASH3A	B
T1D	22	22q12.2	MTMR3	C
T1D	22	22q12.2	UQCR10	C
T1D	22	22q12.3	C1QTNF6	B
T2D	2	2q36	IRS1	B
T2D	3	3p25	PPARG	A
T2D	4	4q27	CCNA2	B
T2D	4	4q27	EXOSC9	C
T2D	8	8q22.1	TP53INP1	A
T2D	9	9p21	CDKN2B	A
T2D	10	10p13	CAMK1D	B

T2D	11	11p15.1	B7H6	C
T2D	11	11p15.1	NUCB2	C
T2D	11	11p15.1	SNORD14A	C
T2D	11	11q13.4	STARD10	C
T2D	12	12q13.1	CERS5	C
T2D	12	12q13.1	DIP2B	C
T2D	12	12q13.1	SLC11A2	C
T2D	12	12q13.1	TFCP2	C
T2D	12	12q24.31	C12orf43	C
T2D	15	15q24	PSTPIP1	C
T2D	15	15q24	TSPAN3	B
T2D	15	15q26.1	HDDC3	C
T2D	15	15q26.1	RCCD1	C
T2D	15	15q26.1	UNC45A	C

Bibliography

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. a, … McVean, G. a. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–73. doi:10.1038/nature09534

Alberts, R., Fu, J., Swertz, M. a, Lubbers, L. A., Albers, C. J., & Jansen, R. C. (2005). Combining microarrays and genetic analysis. *Briefings in Bioinformatics*, *6*(2), 135–45. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15975223

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., … McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*(7311), 52–8. doi:10.1038/nature09298

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., … Rich, S. S. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics*, *41*(6), 703–7. doi:10.1038/ng.381

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., … Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genetics*, *40*(8), 955–62. doi:10.1038/ng.175

Bauer, C., Glintschert, A., & Schuchhardt, J. (2014). ProfileDB: a resource for proteomics and cross-omics biomarker discovery. *Biochimica et Biophysica Acta*, *1844*(5), 960–6. doi:10.1016/j.bbapap.2013.11.007

Belew, A. T., Hepler, N. L., Jacobs, J. L., & Dinman, J. D. (2008). PRFdb: a database of computationally predicted eukaryotic programmed -1 ribosomal frameshift signals. *BMC Genomics*, *9*, 339. doi:10.1186/1471-2164-9-339

Betel, D., Wilson, M., Gabow, A., Marks, D. S., & Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Research*, *36*(Database issue), D149–53. doi:10.1093/nar/gkm995

Bønnelykke, K., Matheson, M. C., Pers, T. H., Granell, R., Strachan, D. P., Alves, A. C., … Henderson, A. J. (2013). Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. *Nature Genetics*, *45*(8), 902–6. doi:10.1038/ng.2694

Brem, R. B., Yvert, G., Clinton, R., & Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science (New York, N.Y.)*, *296*(5568), 752–5. doi:10.1126/science.1069516

Brogna, S., & Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. *Nature Structural & Molecular Biology*, *16*(2), 107–13. doi:10.1038/nsmb.1550

Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M. T., Wiltshire, T., … de Haan, G. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using "genetical genomics". *Nature Genetics*, *37*(3), 225–32. doi:10.1038/ng1497

Campbell, B. J., Yu, L. G., & Rhodes, J. M. (n.d.). Altered glycosylation in inflammatory bowel disease: a possible role in cancer development. *Glycoconjugate Journal*, *18*(11-12), 851–8. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12820718

Cao, C., & Moult, J. (2014). GWAS and drug targets. *BMC Genomics*, *15*(Suppl 4), S5. doi:10.1186/1471-2164-15-S4-S5

Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., … Williams, R. W. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics*, *37*(3), 233–42. doi:10.1038/ng1518

Cheung, V. G., Spielman, R. S., Ewens, K. G., Weber, T. M., Morley, M., & Burdick, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, *437*(7063), 1365–9. doi:10.1038/nature04244

Cho, Y. S., Chen, C.-H., Hu, C., Long, J., Ong, R. T. H., Sim, X., … Seielstad, M. (2012). Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nature Genetics*, *44*(1), 67–72. doi:10.1038/ng.1019

Choy, E., Yelensky, R., Bonakdar, S., Plenge, R. M., Saxena, R., De Jager, P. L., … Altshuler, D. (2008). Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genetics*, *4*(11), e1000287. doi:10.1371/journal.pgen.1000287

Chu, X., Pan, C.-M., Zhao, S.-X., Liang, J., Gao, G.-Q., Zhang, X.-M., … Song, H.-D. (2011). A genome-wide association study identifies two new risk loci for Graves' disease. *Nature Genetics*, *43*(9), 897–901. doi:10.1038/ng.898

Cookson, W., Liang, L., Abecasis, G., Moffatt, M., & Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, *10*(3), 184–94. doi:10.1038/nrg2537

DeCook, R., Lall, S., Nettleton, D., & Howell, S. H. (2006). Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics*, *172*(2), 1155–64. doi:10.1534/genetics.105.042275

Deutsch, S., Lyle, R., Dermitzakis, E. T., Attar, H., Subrahmanyan, L., Gehrig, C., … Antonarakis, S. E. (2005). Gene expression variation and expression quantitative trait mapping of human chromosome 21 genes. *Human Molecular Genetics*, *14*(23), 3741–9. doi:10.1093/hmg/ddi404

Di Girolamo, F., Lante, I., Muraca, M., & Putignani, L. (2013). The Role of Mass Spectrometry in the "Omics" Era. *Current Organic Chemistry*, *17*(23), 2891–2905. doi:10.2174/13852728178881311181627 25

Dimas, A. S. A., Deutsch, S., Stranger, B. E. B., Montgomery, S. B., Borel, C., Attar-Cohen, H., … others. (2009). Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science*, *1246*(5945), 1246. doi:10.1126/science.1174148

Ding, J., Gudjonsson, J. E., Liang, L., Stuart, P. E., Li, Y., Chen, W., … Abecasis, G. R. (2010). Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *American Journal of Human Genetics*, *87*(6), 779–89. doi:10.1016/j.ajhg.2010.10.024

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., … Cookson, W. O. C. (2007). A genome-wide association study of global gene expression. *Nature Genetics*, *39*(10), 1202–7. doi:10.1038/ng2109

Dogan, R. I., Getoor, L., Wilbur, W. J., & Mount, S. M. (2007). SplicePort--an interactive splice-site analysis tool. *Nucleic Acids Research*, *35*(Web Server issue), W285–91. doi:10.1093/nar/gkm407

Doss, S., Schadt, E. E., Drake, T. A., & Lusis, A. J. (2005). Cis-acting expression quantitative trait loci in mice. *Genome Research*, *15*(5), 681–691. doi:10.1101/gr.3216905.

Duan, S., Huang, R. S., Zhang, W., Bleibel, W. K., Roe, C. A., Clark, T. A., … others. (2008). Genetic architecture of transcript-level variation in humans. *American Journal of Human Genetics*, *82*(5), 1101–1113. doi:10.1016/j.ajhg.2008.03.006.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., … Stefansson, K. (2008). Genetics of gene expression and its effect on disease. *Nature*, *452*(7186), 423–8. doi:10.1038/nature06758

Erbilgin, A., Civelek, M., Romanoski, C. E., Pan, C., Hagopian, R., Berliner, J. A., & Lusis, A. J. (2013). Identification of CAD candidate genes in GWAS loci and their expression in vascular cells. *Journal of Lipid Research*, *54*(7), 1894–905. doi:10.1194/jlr.M037085

Ertekin-Taner, N. (2011). Gene expression endophenotypes: a novel approach for gene discovery in Alzheimer's disease. *Molecular Neurodegeneration*, *6*(1), 31. doi:10.1186/1750-1326-6-31

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., … Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature Genetics*, *44*(5), 502–10. doi:10.1038/ng.2205

Farrall, M. (2004). Quantitative genetic variation: a post-modern view. *Human Molecular Genetics*, *13 Spec No*, R1–7. doi:10.1093/hmg/ddh084

Ferreira, T., & Marchini, J. (2011). Modeling interactions with known risk loci-a Bayesian model averaging approach. *Annals of Human Genetics*, *75*(1), 1–9. doi:10.1111/j.1469-1809.2010.00618.x

Franke, A., McGovern, D. P. B., Barrett, J. C., Wang, K., Radford-Smith, G. L., Ahmad, T., … Parkes, M. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, *42*(12), 1118–25. doi:10.1038/ng.717

Fransen, K., Visschedijk, M. C., van Sommeren, S., Fu, J. Y., Franke, L., Festen, E. a M., … Weersma, R. K. (2010). Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease. *Human Molecular Genetics*, *19*(17), 3482–8. doi:10.1093/hmg/ddq264

French, D., Yang, W., Hamilton, L. H., Neale, G., Fan, Y., Downing, J. R., … Relling, M. V. (2008). Concordant gene expression in leukemia cells and normal leukocytes is associated with germline cis-SNPs. *PloS One*, *3*(5), e2144. doi:10.1371/journal.pone.0002144

Gamazon, E. R., Huang, R. S., Cox, N. J., & Dolan, M. E. (2010). Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(20), 9287–92. doi:10.1073/pnas.1001827107

Gamazon, E. R., Zhang, W., Konkashbaev, A., Duan, S., Kistner, E. O., Nicolae, D. L., … Cox, N. J. (2010). SCAN: SNP and copy number annotation. *Bioinformatics*, *26*(2), 259–62. doi:10.1093/bioinformatics/btp644

Gelfond, J. a L., Ibrahim, J. G., & Zou, F. (2007). Proximity model for expression quantitative trait loci (eQTL) detection. *Biometrics*, *63*(4), 1108–16. doi:10.1111/j.1541-0420.2007.00778.x

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. a, Lai, S.-L., … Singleton, A. B. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genetics*, *6*(5), e1000952. doi:10.1371/journal.pgen.1000952

Gilad, Y., Rifkin, S. a, Bertone, P., Gerstein, M., & White, K. P. (2005). Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Research*, *15*(5), 674–80. doi:10.1101/gr.3335705

Gilad, Y., Rifkin, S. a, & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, *24*(8), 408–15. doi:10.1016/j.tig.2008.06.001

Göring, H. H. H., Curran, J. E., Johnson, M. P., Dyer, T. D., Charlesworth, J., Cole, S. a, … Blangero, J. (2007). Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genetics*, *39*(10), 1208–16. doi:10.1038/ng2119

Greenawalt, D. M., Dobrin, R., Chudin, E., Hatoum, I. J., Suver, C., Beaulaurier, J., … Kaplan, L. M. (2011). A survey of the genetics of stomach, liver, and adipose gene expression from a morbidly obese cohort. *Genome Research*, *21*(7), 1008–16. doi:10.1101/gr.112821.110

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*, *27*(1), 91–105. doi:10.1016/j.molcel.2007.06.017

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., … Spector, T. D. (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, *44*(10), 1084–9. doi:10.1038/ng.2394

Hamza, T. H., Chen, H., Hill-Burns, E. M., Rhodes, S. L., Montimurro, J., Kay, D. M., … Payami, H. (2011). Genome-Wide Gene-Environment Study Identifies Glutamate Receptor Gene GRIN2A as a Parkinson's Disease Modifier Gene via Interaction with Coffee. *PLoS Genetics*, *7*(8), e1002237. doi:10.1371/journal.pgen.1002237

Heap, G. a, Trynka, G., Jansen, R. C., Bruinenberg, M., Swertz, M. a, Dinesen, L. C., … Franke, L. (2009). Complex nature of SNP genotype effects on gene expression in primary human leucocytes. *BMC Medical Genomics*, *2*, 1. doi:10.1186/1755-8794-2-1

Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., … Lindgren, C. M. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nature Genetics*, *42*(11), 949–60. doi:10.1038/ng.685

Hindorff, L. a, Sethupathy, P., Junkins, H. a, Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. a. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(23), 9362–7. doi:10.1073/pnas.0903103106

Hirawa, N., Fujiwara, A., & Umemura, S. (2013). ATP2B1 and blood pressure: from associations to pathophysiology. *Current Opinion in Nephrology and Hypertension*, *22*(2), 177–84. doi:10.1097/MNH.0b013e32835da4ca

Holm, K., Melum, E., Franke, A., & Karlsen, T. H. (2010). SNPexp - A web tool for calculating and visualizing correlation between HapMap genotypes and gene expression levels. *BMC Bioinformatics*, *11*, 600. doi:10.1186/1471-2105-11-600

Hong, K.-W., Jin, H.-S., Lim, J.-E., Cho, Y. S., Go, M. J., Jung, J., … Oh, B. (2010). Non-synonymous single-nucleotide polymorphisms associated with blood pressure and hypertension. *Journal of Human Hypertension*, *24*(11), 763–74. doi:10.1038/jhh.2010.9

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6), e1000529. doi:10.1371/journal.pgen.1000529

Hrdlickova, B., Westra, H.-J., Franke, L., & Wijmenga, C. (2011). Celiac disease: moving from genetic associations to causal variants. *Clinical Genetics*, *80*(3), 203–313. doi:10.1111/j.1399-0004.2011.01707.x

Hsu, Y.-H., Zillikens, M. C., Wilson, S. G., Farber, C. R., Demissie, S., Soranzo, N., … Kiel, D. P. (2010). An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits. *PLoS Genetics*, *6*(6), e1000977. doi:10.1371/journal.pgen.1000977

Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., … Aitman, T. J. (2005). Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics*, *37*(3), 243–53. doi:10.1038/ng1522

Hunt, K. a, Zhernakova, A., Turner, G., Heap, G. a R., Franke, L., Bruinenberg, M., … van Heel, D. a. (2008). Newly identified genetic risk variants for celiac disease related to the immune response. *Nature Genetics*, *40*(4), 395–402. doi:10.1038/ng.102

Innocenti, F., Cooper, G. M., Stanaway, I. B., Gamazon, E. R., Smith, J. D., Mirkov, S., … Brown, C. D. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait Loci in primary human liver tissue. *PLoS Genetics*, *7*(5), e1002078. doi:10.1371/journal.pgen.1002078

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., … van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*(2), 149–55. doi:10.1038/ng.295

Jacobs, J. L., Belew, A. T., Rakauskaite, R., & Dinman, J. D. (2007). Identification of functional, endogenous programmed -1 ribosomal frameshift signals in the genome of Saccharomyces cerevisiae. *Nucleic Acids Research*, *35*(1), 165–74. doi:10.1093/nar/gkl1033

Jansen, R. C., & Nap, J. P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics*, *17*(7), 388–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11418218

Kang, H., Morgan, A., & Chen, R. (2012). Coanalysis of GWAS with eQTLs reveals disease-tissue associations. *AMIA Joint Summits on Translational Science Proceedings*, (2), 35–41. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3392070/

Kang, H., Yang, X., Chen, R., & Zhang, B. (2012). Integration of disease-specific single nucleotide polymorphisms, expression quantitative trait loci and coexpression networks reveal novel candidate genes for type 2. *Diabetologia*, 2205–2213. doi:10.1007/s00125-012-2568-3

Khurana, E., Fu, Y., Colonna, V., Mu, X. J., Kang, H. M., Lappalainen, T., … Gerstein, M. (2013). Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science (New York, N.Y.)*, *342*(6154), 1235587. doi:10.1126/science.1235587

Kozomara, A., & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, *39*(Database issue), D152–7. doi:10.1093/nar/gkq1027

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., … Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–8. doi:10.1038/nature09410

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. a. C., Monlong, J., Rivas, M. a., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. doi:10.1038/nature12531

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., … Liang, L. (2013). Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in Genetics*, *4*(May), 1–14. doi:10.3389/fgene.2013.00103

Li, Y., Alvarez, O. A., Gutteling, E. W., Tijsterman, M., Fu, J., Riksen, J. A. G., … Kammenga, J. E. (2006). Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. *PLoS Genetics*, *2*(12), e222. doi:10.1371/journal.pgen.0020222

Liang, L., Morar, N., Dixon, A. L., Lathrop, G. M., Abecasis, G. R., Moffatt, M. F., & Cookson, W. O. C. (2013). A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Research*, *23*(4), 716–26. doi:10.1101/gr.142521.112

Listgarten, J., Kadie, C., Schadt, E. E., & Heckerman, D. (2010). Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(38), 16465–70. doi:10.1073/pnas.1002425107

Liu, C. (2011). Brain expression quantitative trait locus mapping informs genetic studies of psychiatric diseases. *Neuroscience Bulletin*, *27*(2), 123–33. doi:10.1007/s12264-011-1203-5

Loo, L. W. M., Cheng, I., Tiirikainen, M., Lum-Jones, A., Seifried, A., Dunklee, L. M., … Le Marchand, L. (2012). cis-Expression QTL analysis of established colorectal cancer risk variants in colon tumors and adjacent normal tissue. *PloS One*, *7*(2), e30477. doi:10.1371/journal.pone.0030477

Manolio, T. a, Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. a, Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–53. doi:10.1038/nature08494

Maquat, L. E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Reviews. Molecular Cell Biology*, *5*(2), 89–99. doi:10.1038/nrm1310

Maranville, J. C., Luca, F., Richards, A. L., Wen, X., Witonsky, D. B., Baxter, S., … Di Rienzo, A. (2011). Interactions between Glucocorticoid Treatment and Cis-Regulatory Polymorphisms Contribute to Cellular Response Phenotypes. *PLoS Genetics*, *7*(7), e1002162. doi:10.1371/journal.pgen.1002162

Michaelson, J. J., Alberts, R., Schughart, K., & Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics*, *11*, 502. doi:10.1186/1471-2164-11-502

Moffatt, M. F., Kabesch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., … Cookson, W. O. C. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature*, *448*(7152), 470–3. doi:10.1038/nature06014

Monks, S. a, Leonardson, a, Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., … Schadt, E. E. (2004). Genetic inheritance of gene expression in human cell lines. *American Journal of Human Genetics*, *75*(6), 1094–105. doi:10.1086/426461

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., … Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, *464*(7289), 773–7. doi:10.1038/nature08903

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., & Cheung, V. G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature*, *430*(7001), 743–7. doi:10.1038/nature02797

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18516045

Mosnier, J.-F., Jarry, A., Bou-Hanna, C., Denis, M. G., Merlin, D., & Laboisse, C. L. (2006). ADAM15 upregulation and interaction with multiple binding partners in inflammatory bowel disease. *Laboratory Investigation; a Journal of Technical Methods and Pathology*, *86*(10), 1064–73. doi:10.1038/labinvest.3700465

Mota, L. J., Ramsden, A. E., Liu, M., Castle, J. D., & Holden, D. W. (2009). SCAMP3 is a component of the Salmonella-induced tubular network and reveals an interaction between bacterial effectors and post-Golgi trafficking. *Cellular Microbiology*, *11*(8), 1236–53. doi:10.1111/j.1462-5822.2009.01329.x

Mueller, M., Goel, A., Thimma, M., Dickens, N. J., Aitman, T. J., & Mangion, J. (2006). eQTL Explorer: integrated mining of combined genetic linkage and expression experiments. *Bioinformatics*, *22*(4), 509–11. doi:10.1093/bioinformatics/btk007

Myers, A. J., Gibbs, J. R., Webster, J. a, Rohrer, K., Zhao, A., Marlowe, L., … Hardy, J. (2007). A survey of genetic human cortical gene expression. *Nature Genetics*, *39*(12), 1494–9. doi:10.1038/ng.2007.16

Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I., & Dermitzakis, E. T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*, *6*(4), e1000895. doi:10.1371/journal.pgen.1000895

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., … Spector, T. D. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genetics*, *7*(2), e1002003. doi:10.1371/journal.pgen.1002003

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics*, *6*(4), e1000888. doi:10.1371/journal.pgen.1000888

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., … Plenge, R. M. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, *506*(7488), 376–81. doi:10.1038/nature12873

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–5. doi:10.1038/ng.259

Perry, J. R. B., Voight, B. F., Yengo, L., Amin, N., Dupuis, J., Ganser, M., … Cauchi, S. (2012). Stratifying type 2 diabetes cases by BMI identifies genetic risk

variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS Genetics*, *8*(5), e1002741. doi:10.1371/journal.pgen.1002741

Petretto, E., Mangion, J., Pravanec, M., Hubner, N., & Aitman, T. J. (2006). Integrated gene expression profiling and linkage analysis in the rat. *Mammalian Genome : Official Journal of the International Mammalian Genome Society*, *17*(6), 480–9. doi:10.1007/s00335-005-0181-1

Pickrell, J. K., Marioni, J. C., Pai, A. a, Degner, J. F., Engelhardt, B. E., Nkadori, E., … Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, *464*(7289), 768–72. doi:10.1038/nature08872

Plant, E. P., Wang, P., Jacobs, J. L., & Dinman, J. D. (2004). A programmed -1 ribosomal frameshift signal can function as a cis-acting mRNA destabilizing element. *Nucleic Acids Research*, *32*(2), 784–90. doi:10.1093/nar/gkh256

Price, A. L., Patterson, N., Hancks, D. C., Myers, S., Reich, D., Cheung, V. G., & Spielman, R. S. (2008). Effects of cis and trans genetic ancestry on gene expression in African Americans. *PLoS Genetics*, *4*(12), e1000294. doi:10.1371/journal.pgen.1000294

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. a R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–75. doi:10.1086/519795

Ricaño-Ponce, I., & Wijmenga, C. (2013). Mapping of immune-mediated disease genes. *Annual Review of Genomics and Human Genetics*, *14*, 325–53. doi:10.1146/annurev-genom-091212-153450

Richards, a L., Jones, L., Moskvina, V., Kirov, G., Gejman, P. V, Levinson, D. F., … O'Donovan, M. C. (2012). Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Molecular Psychiatry*, *17*(2), 193–201. doi:10.1038/mp.2011.11

Rotival, M., Zeller, T., Wild, P. S., Maouche, S., Szymczak, S., Schillert, A., … Blankenberg, S. (2011). Integrating Genome-Wide Genetic Variations and Monocyte Expression Data Reveals Trans-Regulated Gene Modules in Humans. *PLoS Genetics*, *7*(12), e1002367. doi:10.1371/journal.pgen.1002367

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Niu, X., Fengler, K. A., … Tuberosa, R. (2007). Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proceedings of the National*

Academy of Sciences of the United States of America, *104*(27), 11376–81. doi:10.1073/pnas.0704145104

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., Guhathakurta, D., … Lusis, A. J. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, *37*(7), 710–7. doi:10.1038/ng1589

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., … Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*, *6*(5), e107. doi:10.1371/journal.pbio.0060107

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., … others. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, *422*(6929), 297–302. doi:10.1038/nature01482.1.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., & Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(20), 10614–9. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=38202&tool=pmcentrez&rendertype=abstract

Schröder, a, Klein, K., Winter, S., Schwab, M., Bonin, M., Zell, a, & Zanger, U. M. (2011). Genomics of ADME gene expression: mapping expression quantitative trait loci relevant for absorption, distribution, metabolism and excretion of drugs in human liver. *The Pharmacogenomics Journal*, (January), 1–9. doi:10.1038/tpj.2011.44

Sim, X., Ong, R. T.-H., Suo, C., Tay, W.-T., Liu, J., Ng, D. P.-K., … Tai, E.-S. (2011). Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS Genetics*, *7*(4), e1001363. doi:10.1371/journal.pgen.1001363

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., … Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, *42*(11), 937–48. doi:10.1038/ng.686

Spielman, R. S., Bastone, L. a, Burdick, J. T., Morley, M., Ewens, W. J., & Cheung, V. G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics*, *39*(2), 226–31. doi:10.1038/ng1955

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., … Plenge, R. M. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature Genetics*, *42*(6), 508–14. doi:10.1038/ng.582

Storey, J. D., Madeoy, J., Strout, J. L., Wurfel, M., Ronald, J., & Akey, J. M. (2007). Gene-expression variation within and among human populations. *American Journal of Human Genetics*, *80*(3), 502–9. doi:10.1086/512017

Stranger, B. E., Forrest, M. S., Clark, A. G., Minichiello, M. J., Deutsch, S., Lyle, R., … Dermitzakis, E. T. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics*, *1*(6), e78. doi:10.1371/journal.pgen.0010078

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., … Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics*, *39*(10), 1217–24. doi:10.1038/ng2142

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., … Yaspo, M. (2008). of the Human Transcriptome, *685*(August), 956–960.

Sunyaev, S., Ramensky, V., & Bork, P. (2000). Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends in Genetics : TIG*, *16*(5), 198–200. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10782110

Teslovich, T. M., Musunuru, K., Smith, A. V, Edmondson, A. C., Stylianou, I. M., Koseki, M., … Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, *466*(7307), 707–13. doi:10.1038/nature09270

The Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, *447*(7145), 661–78. doi:10.1038/nature05911

Verlaan, D. J., Ge, B., Grundberg, E., Hoberman, R., Lam, K. C. L., Koka, V., … Pastinen, T. (2009). Targeted screening of cis-regulatory variation in human haplotypes. *Genome Research*, *19*(1), 118–27. doi:10.1101/gr.084798.108

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., & Pritchard, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics*, *4*(10), e1000214. doi:10.1371/journal.pgen.1000214

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., … Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–6. doi:10.1038/nature07509

Wang, G.-S., & Cooper, T. a. (2007). Splicing in disease: disruption of the splicing code and the decoding machinery. *Nature Reviews. Genetics*, *8*(10), 749–61. doi:10.1038/nrg2164

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. doi:10.1038/nrg2484

Wang, Z., & Moult, J. (2001). SNPs, protein structure, and disease. *Human Mutation*, *17*(4), 263–70. doi:10.1002/humu.22

Waterworth, D. M., Ricketts, S. L., Song, K., Chen, L., Zhao, J. H., Ripatti, S., … Sandhu, M. S. (2010). Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *30*(11), 2264–76. doi:10.1161/ATVBAHA.109.201020

Wayne, M. L., & McIntyre, L. M. (2002). Combining mapping and arraying: An approach to candidate gene identification. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(23), 14903–6. doi:10.1073/pnas.222549199

Webster, J. a, Gibbs, J. R., Clarke, J., Ray, M., Zhang, W., Holmans, P., … Myers, A. J. (2009). Genetic control of human brain transcript expression in Alzheimer disease. *American Journal of Human Genetics*, *84*(4), 445–58. doi:10.1016/j.ajhg.2009.03.011

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., … Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, *42*(Database issue), D1001–6. doi:10.1093/nar/gkt1229

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., … Abecasis, G. R. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, *40*(2), 161–9. doi:10.1038/ng.76

Wu, C., Miao, X., Huang, L., Che, X., Jiang, G., Yu, D., … Lin, D. (2012). Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nature Genetics*, *44*(1), 62–6. doi:10.1038/ng.1020

Xia, K., Shabalin, A. a, Huang, S., Madar, V., Zhou, Y.-H., Wang, W., … Wright, F. a. (2011). seeQTL: A searchable database for human eQTLs. *Bioinformatics*, 2–3. doi:10.1093/bioinformatics/btr678

Yampolsky, L. Y., & Stoltzfus, A. (2005). The exchangeability of amino acids in proteins. *Genetics*, *170*(4), 1459–72. doi:10.1534/genetics.104.039107

Yang, T.-P., Beazley, C., Montgomery, S. B., Dimas, A. S., Gutierrez-Arcelus, M., Stranger, B. E., … Dermitzakis, E. T. (2010). Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, *26*(19), 2474–6. doi:10.1093/bioinformatics/btq452

Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., … Kruglyak, L. (2003). Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics*, *35*(1), 57–64. doi:10.1038/ng1222

Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., … Cambien, F. (2010). Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One*, *5*(5), e10693. doi:10.1371/journal.pone.0010693

Zhang, M., Song, F., Liang, L., Nan, H., Zhang, J., Liu, H., … Han, J. (2013). Genome-wide association studies identify several new loci associated with pigmentation traits and skin cancer risk in European Americans. *Human Molecular Genetics*, *22*(14), 2948–59. doi:10.1093/hmg/ddt142

Zhang, W., Duan, S., Kistner, E., & Bleibel, W. (2008). Evaluation of genetic variation contributing to differences in gene expression between populations. *American Journal of Human Genetics*, (March), 631–640. doi:10.1016/j.ajhg.2007.12.015.

Zhong, H., Beaulaurier, J., Lum, P. Y., Molony, C., Yang, X., Macneil, D. J., … Schadt, E. E. (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genetics*, *6*(5), e1000932. doi:10.1371/journal.pgen.1000932