EFFECTS OF GENETIC VARIANTS ON GENE EXPRESSION VARIABILITY

A Dissertation

by

GANG WANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	James J. Cai
Co-Chair of Committee,	Yanan Tian
Committee Members,	Paul B. Samollow
	Joshua Yuan
Head of Department,	Evelyn Tiffany-Castiglioni

May 2016

Major Subject: Biomedical Sciences

ABSTRACT

Variation and variability of gene expression are central concepts in biology. Variation refers to differences among individuals, whereas variability refers to the potential of a population to vary. The advent of next-generation sequencing technology has lead to the accumulation of an ever-increasing number of population level, large-scale genotype and gene expression data sets, which provide excellent opportunities to identify the genetic loci that potentially affect gene expression variation and variability.

Over the last several years, much effort has been made to identify genetic loci that affect the mean differences in phenotypic expression between genotypes, but these studies have largely ignored loci that affect the variance of phenotypic expression within individual genotypes. Although studies of expression quantitative trait loci (eQTL) have established a convincing relationship between genotype and levels of gene expression, the impact of genetic variants on gene expression variance remains unclear. In addition, the analytical frameworks adopted by most eQTL studies have been based on population-level test statistics, which are powerful for assessing the effects of common genetic variants, but not rare or private genetic variants. Few frameworks or statistics are available for assessing the impacts of rare genetic mutations on gene expression. Thus, a new statistical method is required to address this issue.

In this dissertation, I aim to address these questions in humans using publically available large-scale, Next-generation RNA sequencing datasets and new experimental data from my own work. I first adopted a new statistical method called double

generalized linear model (DGLM) to study the effect of common genetic variants on gene expression variability, which I define as expression variability QTL (evQTL), using data from the TwinsUK study. I searched the whole genome to identify common genetic variants associated with variable expression at *cis*-acting genes and showed the contribution of both genetic and nongenetic factors to variable gene expression. I next examined two distinct modes of action of evQTLs: GxG interaction (the interaction between genotypes at different loci) and GxE interaction (the interaction between genotype and environment), which showed that common genetic variants work interactively or independently to influence gene expression variance. Lastly, I established a novel analytical framework to evaluate the effects of rare or private variants on gene expression variability. This method starts from the identification of outlier individuals that show markedly different gene expression from the majority of a population, and then reveals the contributions of private SNPs to the aberrant gene expression in these outliers.

ACKNOWLEDGEMENTS

It indicates a conclusion of my journey as a student when I am writing this dissertation. From the beautiful small town where I was born to College Station, I have stayed in several places in order to improve myself through education. It has been more than 20 years since I went to primary school. The past five years of graduate study at Texas A&M University are the most unforgettable moments in my life. Here, my heartfelt thanks go to my professors, classmates, friends, and family who have helped me grow and become the person I am today.

First and foremost, special thanks should be expressed to my committee chair, Dr. James Cai, for his great guidance and support through my study and research. His invaluable counsel enabled me to think and resolve many problems I was faced during my research. Dr. Cai is a great teacher and a great friend with a great personality, his diligence, honest and approachable, have set a good example not only for my career, but also for a human being. Thanks should be also expressed to my committee members, Dr. Yanan Tian, Dr. Paul Samollow and Dr. Joshua Yuan, for their valuable comments and suggestions.

I would also like to thank my lab-mates in Dr. Cai's lab, especially Dr. Ence Yang, a great lab-mate and friend, who gave me a big help during the difficult period at the beginning of my research.

Thanks also go to our collaborators, Dr. William Murphy, Dr. Beiyan Zhou, Dr. Nann Fangue and Dr. Richard Connon, for their kind help.

Lastly, and most importantly, I would express my deepest gratitude to my family, especially to my parents and my wife, who firmly stand behind me no matter where and when I am.

TABLE OF CONTESTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

NOMENCLATURE

CDFs	Cumulative distribution functions
cDNA	Complementary DNA
CEU	Utah residents with ancestry from northern and western Europe
ChIP	Chromatin immunoprecipitation
ChIP-seq	ChIP-sequencing
CTCF	CTCF-enriched element
CV	Coefficient of variation
DGLM	Double generalized linear model
dDNA	Deoxyribonucleic acid
DZ	Dizygotic
E	Predicted enhancer
EBI	European Bioinformatics Institute
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression QTL
evQTL	Expression variability QTL
FDR	False-discovery rate
FIN	Finnish in Finland
F-K	Fligner-Killeen
FPKM	Fragments Per Kilobase of transcript per Million reads
GBR	British in England and Scotland

GEO	Gene Expression Omnibus
GO	Gene Ontology
GTEx	Genotype-Tissue Expression project
GWAS	Genome-wide association studies
GxE	Interaction between gene (or genotype) and environment
IGV	Integrative Genomics Viewer
K-S	Kolmogorov-Smirnov
LCLs	Lymphoblastoid cell line
LD	Linkage Disequilibrium
MAF	Minor allele frequency
MCD	The minimum covariance determinant
MD	Mahalanobis distance
MISO	Mixture-of-isoforms
mRNA	Messenger RNA
MSigDB	Molecular Signatures Database
MuTHER	Multiple Tissue Human Expression Resource
MZ	Monozygotic
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
PBS	Phosphate-buffered saline
PCA	Principal component analysis
PEER	Probabilistic estimation of expression residuals

PF	Predicted promoter flanking region
PI	Propidium iodide
pre-mRNA	Precursor mRNA
qPCR	Quantitative PCR
R	Predicted repressed or low-activity region
RMD	Relative mean difference
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RT-qPCR	Quantitative reverse transcription PCR
SNP	Single nucleotide polymorphisms
SRA	Sequence Read Archives
SSMD	Sum of squared MD
SVLM	Squared residual value linear modeling
T	Predicted transcribed region
TSI	Toscani in Italia
TSS	Predicted promoter region including TSS
WE	Predicted weak enhancer or open chromatin *cis*-regulatory Element
YRI	Yoruba in Ibadan

CHAPTER I

INTRODUCTION

1.1 Human genetic variation

Human genetic variation comprises all of the differences in the genetic sequence both within and among populations. A larger amount of genetic variation is found associated with alter levels of gene expression and increased or decreased risk of disease through distinct mechanisms. Elucidating the contribution of genetic variation to human health and disease has become a major challenge for biology in the 21st century and promises significant benefits to human welfare [1]. Human genetic variation can be divided into two broad categories: single-nucleotide polymorphisms (SNP) and structural variations. Structural variations, in turn, include indels (insertions and deletions), copy number variants, inversions, and translocations. Among these genetic variations, SNP is the most common category of genetic variants in the human genome, accounting for more than 90% of known polymorphisms. Based on the minor allele frequency (MAF), SNPs can be classified into three groups: common (MAF> 5%), low-frequency (0.5% ≤ MAF < 5%) and rare SNPs (MAF < 0.5%). In this dissertation, I am primarily focused on the contribution of SNPs to gene expression variance.

In 2001, the Human Genome Sequencing Consortium [2] and Celera [3] published their first haploid human genome sequence in succession based on a very limited number of individuals. Despite the achievement of annotation of the human genome sequence, genetic variation was not noted in either of the human genome

references. One year later, the International HapMap Project was initiated to understand the common patterns of human genetic variation that may associate with disease risk. The first haplotype map (HapMap) of the human genome was released in 2005 [4]. More than one million SNPs were obtained using 269 human samples from four geographically diverse populations: 'Yoruba', 'Northern and Western European', 'Han Chinese' and 'Japanese'. Two years later (2007), a second generation human haplotype map was published with over 3.1 million SNPs reported [5]. In 2008, scientists expanded the number of samples to more than one thousand people, and the 1000 Genomes Project was launched with the objective of providing the most detailed catalog of human genetic variation available to study the relationship between genotypes and phenotypes. The pilot phase of the project was published in 2010 [6], with a description of approximately 15 million SNPs, 1 million short insertions and deletions, and 20,000 structural variants

1.2 The contribution of genetic variants to disease risk

An increasing number of studies are focused on genome variability due to its relevance to complex traits. Understanding the effects of genetic variants on disease risk has become a fundamental requirement for medical genetics. Genome-wide association (GWA) [7] is a powerful strategy for studying associations between common genetic variants and common complex traits, and is typically focused on the impact of common SNPs on complex human diseases, especially those related to major health conditions. Risch [8], who first proposed the idea of GWA in 1996, believed that there was no need to predict candidate genes in future complex human disease studies, and it would be possible to detect genome-wide genetic variants that linked to the disease through large-

scale (population based) genetic testing. In 2005, Klein et al. [9] published the first successful GWA study (GWAS), in which they found two SNPs to be strongly associated with age-related macular degeneration. A typical GWAS may involve the following four steps [10] : (1) selection of a large sample of individuals with a disease of interest and an equally large appropriate control group, (2) genotyping and quality control to ensure the high quality genotype data, (3) statistical analysis to identify significant associations between phenotypes and identifiable genetic variants, and (4) replication of identified associations using an independent sample group. Since 2006, with the completion of the Human Genome Project (HGP) and the HapMap, and especially the rapid development of next-generation sequencing (NGS) technology in the past few years, GWAS has resulted in an explosion of knowledge concerning associations between common SNPs and diseases. As of 2013, more than 1,900 human GWA studies had examined more than 300 common traits and diseases, including digestive system disease, cardiovascular disease, metabolic disease, immune system disease, nervous system disease, liver enzyme disease, lipid or lipoprotein disease, and cancers, and more [11]. One of the best known GWAS successes was the discovery of the FTO locus (fat mass and obesity-associated protein) located on chromosome 16. SNP rs9939609, located within this gene, is strongly associated with type 2 diabetes, where adults homozygous for the risk allele had 1.6-fold increased odds of obesity compared those without the risk allele [12].

Despite the success of GWA studies in identifying a large number of SNPs associated with distinct diseases, there are still several key limitations of the GWA

strategy. Firstly, although a limited number of SNPs that increase or decrease gene transcription activity are located in regulatory elements such as promoters, or that change the amino acid sequence by altering the base sequence in coding exons, the majority of SNPs associated with disease are located in noncoding regions, or are located in coding regions but do not change the translated amino acid sequence (synonymous base substitutions). Therefore, they are not informative for explaining how genetic variants contribute to the disease risk. To address this issue, much effort has been put into discovering the regulatory roles of noncoding sequences in the genome, as exemplified by the Encyclopedia of DNA Elements (ENCODE) project [13] and Roadmap epigenomics mapping consortium project [14]. Also, by the necessity of their experimental design, GWA studies focus on the impact of common SNPs on health conditions and ignore the effects of rare SNPs. However, a long-established idea believed that rare genetic variants may be the primary drivers of common diseases [15,16]. Increasing evidence support this idea with the identification of rare genetic variants associated with diseases, such as inflammatory bowel disease [17], prostate cancer [18] and Alzheimer's disease [19]. Lastly, and most importantly, the biggest challenge we are facing is that GWA studies provide no information to help us understand the underlying molecular mechanisms of the identified relationships between SNPs and the associated diseases condition(s). To overcome these drawbacks in GWA studies, the level of gene expression is introduced as an intermediate phenotype, which provides a link between genetic variants and disease processes.

1.3 Gene expression and regulation

Regulated gene expression in time and space is the most important process to determine and maintain the characteristics of cells. Dysregulation of gene expression may influence the cellular state and function of cells, resulting in abnormal development and diseases conditions [20,21]. Thus, understanding how the expression levels of different genes are determined in different cell types between and within species is the central goal of biology. Gene expression from DNA to mRNA is a multistep process that is regulated at different stages including, but not limited to, transcription and various post-transcriptional mechanisms, which collectively produce mature mRNA and regulated their concentration within the cell. In eukaryotes, transcription process can be divided into three different stages, initiation, elongation and termination [22,23]. The initial transcript generated from this process called pre-mRNA, which is then matured in the process of 5' capping, 3' polyadenylation and splicing [24]. The expression level of an mRNA is determined by rates of mRNA synthesis and degradation[25], which determine the steady-state level of mRNA. In this dissertation, unless otherwise specified, gene expression refers to the abundance of steady-state mRNA, which is influenced by both trans-factors and *cis*-elements.

A *cis*-regulatory element is a specific sequence in DNA that can regulate the expression of a gene to which it is physically linked to the same chromosome strand. *Cis*-regulatory elements can be subdivided into three general classes: promotor elements lie near the transcription starting site of the affected gene, at or near the binding site for RNA polymerase II: enhancer elements may be found upstream or downstream of the

coding region and work with the promoter to greatly enhance the efficiency of transcription: silencer elements can also be upstream or downstream of coding sequences and play a contrary role relative to enhancers by inhibiting the activation of transcription factors and decreasing the efficiency of transcription. Although different regulatory elements play different roles, they are all involved in the binding of trans-acting factors, which are described in the next paragraph.

Trans-acting factors, are molecules, usually proteins, produced at sites anywhere in the genome that can bind to *cis*-acting sequences (as defined above) to regulate gene expression. These can be classified into four general groups: (1) transcription factors: RNA polymerases produce primary RNA transcripts, but cannot bind directly with the promoter. RNA polymerases can enter the promoter region to start the transcription only after the combination of transcription factor and promoter to form a specific complex, (2) activatora are special regulatory proteins that can identify specific sequence elements, binding to the promoter or enhancer sequences in order to enhance the effectiveness of the promoter and increase the frequency of transcription, (3) coactivators provide a connection between activators and basic transcription elements by protein and protein interaction to promote gene transcription, (4) repressors bind to the upstream promoter or even at distant silencer locations where they inhibit transcription initiation through a variety of physical effects including changes in DNA conformation.

Thanks to the rapid development of DNA sequencing technology over the past two decades, we can now measure the global mRNA abundance quickly, efficiently, and accurately, making it possible to study the contribution of *trans*-factors and *cis*-elements

to the mRNA expression on a genome-wide scale. For example, the NGS technology has become an indispensable tool as it is the basis for all large-scale sequencing strategies including RNA-seq [26,27]. In addition, ChIP-seq [28], which is the combination of chromatin immunoprecipitation (ChIP) and the NGS technology provides an efficient means to study relationships between transcription factors and their binding *cis*-elements.

1.4 The influence of genetic variants on gene expression variation

Treating gene expression as a heritable, quantitative trait, and understanding how genetic variants influence gene expression variation and variability is the central topic of my dissertation research. Both genetic and nongenetic (e.g., environmental) factors contribute to gene expression. The most obvious example of the effect of a nongenetic factor is that there is a remarkable variation of gene expression between individuals with the identical genetic makeup, such as identical twins. In addition to nongenetic factors, genetic variants can also influence gene expression in distinct ways, either in *cis* or *trans* depending on the physical distance from the target gene they regulate. Usually, variants with 1 megabase (Mb) on either side of their target gene's translation start site (TSS) are considered *cis* elements , while those located on different chromosomes or more than 5 Mb up- or down-stream of the TSS are regarded as *trans* elements [29].

A relationship between genetic variation and gene expression has been recognized at least since Haldane [30] noted that a gene's activity could be the result of genetic variation in the gene itself. Since then many studies of associations between genetic variation and gene expression on a scale of a few loci at a time have made considerable achievements using approaches introduced by Jacob [31] and Damerval [32]

et al. For example, using Drosophila as a research model, Abraham [33] and Powell [34] et al. showed that *cis*-acting genetic variation affects gene expression in space. Despite such successes, many basic questions about the genetic variation and gene expression remain unknown. For example, how many loci underlie variation in gene expression? What is the magnitude of effects of these loci? Is there any genetic interaction between these loci to influence gene expression? To answer these kinds of questions, there is a critical need to study the contribution of genetic variation to gene expression in a much larger scale by mapping genetic variation to genome-wide gene expression. This need has been met to some degree by the emergence of DNA microarray technology [35], and more recently by the enormous power of NGS technology.

Since Jansen and Nap [36] introduced the concept "genetical genomics", also called "expression genetics", which is based on the genetic mapping of global gene expression through the use of high-throughput gene expression profiling technology. As a result, expression quantitative trait loci (eQTL) studies has been widely applied in different species. Brem et al. [37] published the first genome-wide study of gene expression in yeast *Saccharomyces cerevisiae* in 2002, which demonstrated the feasibility of this strategy. Subsequently, a number of eQTL studies were reported in other species such as *Arabidopsis*, *Saccharomyces cerevisiae*, *C. elegans*, and mouse [38-41]. Recently, more and more eQTL studies have focused on human populations [42-45], successfully mapping eQTL loci to many different characteristics showing cell-specific [46], tissue-specific [47], age-specific [48], and development-specific effects [49].

To identify the gene whose abundance is directly modified by a genetic variant, two types of data are required. First, genotypic data of multiple individuals. Second, the expression data of thousands of gene transcripts for the corresponding genotyped individuals. A statistical test is then applied to test if a given genetic variant is responsible for the expression of a given gene. Most available statistical methods for eQTL study are based on comparing the genotypes with gene expression levels using either linkage or association-based mapping [50,51]. The principle of a linkage mapping is to identify genetic variantswhose transmission patterns are associated with gene expression through families. The linkage mapping is an effective approach to do a genome-wide scan for a small number of SNPs; however, the limitation of this method is the low resolution. In contrast, the principle of an association mapping is that apply a correlation analysis on the expression of a gene across different individuals with different alleles of a genetic variant. For example, suppose we have two vectors, vector ***G*** contains genotypes for *n* individuals and vector ***E*** contains values of gene expression for the same individuals as in vector ***G***. And then, commonly used correlation analysis methods (e.g., Person correlation and Spearman rank correlation) or linear regression analysis can be performed for the two vectors to calculate a *p*-value to determine if the correlation is significant. Association mapping is far more powerful for detecting common genetic variants that contribute to the gene expression variation, and is more suitable for identifying eQTL with medium or small effect size. Unlike the traditional QTL mapping, eQTL mapping usually performed using thousands of genes and millions of SNPs simultaneously. Therefore, eQTL mapping required multiple tests not only for

millions of SNPs, but also for thousands of expression traits, by which the type I error will be greatly increased. To eliminate the effects of multiple tests, several commonly used statistical methods are used to control the type I error: (1) Bonferroni correction, the well-known method to correct for multiple-testing derived by observing Boole's inequality [52], permutation tests, for each linkage between expression trait and marker, we can assess the significance of the association by shuffling the phenotypes [53,54], (3) false discovery rate (FDR), which is the expected proportion of false positives in all claimed significant results. FDR is more powerful than Bonferroni correction as FDR-controlling procedures provide less stringent control of Type I errors. Over the last few years, a number of tools are designed and published for eQTL analysis, such as R/qtl [55], Plink [56], and Matrix eQTL. In addition, several new frameworks are designed for specific eQTL mapping. For example, a statistical framework was introduced by Flutre et al. [57] to take advantage of the richness of the data across multiple tissues by joint analysis of among tissues. Although each of these tools with distinct technical details and has their own drawbacks, they have the general trend to provide a genome-wide, fast and efficient tool for eQTL detection.

1.5 The influence of genetic variants on gene expression variance

Notwithstanding these considerable achievements, eQTL studies focuse primarily on the contribution of genetic variants to the mean differences in gene expression between genotypes, largely ignoring the differences in gene expression variance. The reason for this is that quantitative genetics is based on the assumption that phenotypic mean difference is explained by differences in mean phenotypes among

different genotypes, while genotypic variability is the result of environmental (nongenetic) perturbations, and thus is not genetically controlled. However, recent studies have shown that the variance of phenotypic expression is also genetically controlled. For example, Yang's study [58] showed the SNP rs7202116 at the FTO gene locus is associated with variability of body mass index in the human population, and Shen's study [59] explored genetic effects on the variance heterogeneity in Arabidopsis. In addition, a recent study by Ayroles and colleagues [60] showed that several genes affect variability in handedness without affecting the mean, which indicated that different genotypes differ dramatically for phenotypic variability. Recently, a number of studies have focused on the associations between genetic variants and variances of the phenotypic trait (vQTL) [58,61,62]. To study the influence of genetic variants on the variances of phenotypic traits, robust statistical methods are required. The most commonly used methods for vQTL identification include: (1) Levene's test [63], (2) Brown-Forysthe test [64], and (3) the correlation least squares (CLS) test [62]. Both Levene's and Brown-Forysthe tests use ANOVA-based statistics, the difference between the two tests is that the Levene's test uses the mean in computing the spread within each group while Brown-Forysthe test uses the median instead of the mean and therefor overcomes the assumption of symmetric noise [64]. The CLS test first apply a liner regression test to the genotypes and traits and residuals are calculated, then a spearman rank correction test between the squared residuals and genotypes is used to detect the evidence of variance effects. However, despite the capability for vQTL detection, each method mentioned above has their own drawbacks. For Levene's and Brown-Forysthe

tests, both of them not allowing continuous and additional possibly confounding covariates. Although CLS test addresses this problem, it has the problem of easy overfitting.

When gene expression variance is considered as a heritable, quantitative trait, the variance should be genetically controlled as shown in biological systems. However, despite a few initial efforts focus on the quantification of the variance of gene expression [65-67], the influence of genetic variants on gene expression variability remains largely unknown. A recent study in our laboratory introduced the concept of expression variability QTL (evQTL) [68], which are genetic loci linked to or associated with expression variance of genotypes at another locus. To identify evQTL, we adopted a full parametric approach called the double generalized linear model (DGLM) method [69] with several advantages. For example, it accounts for the uncertainty of fitted parameters for both the mean and the variance aspects of the model, and also allows fitting of covariates [70]; it is also highly flexible, allowing for any response distribution from the exponential family [71] (such as binomial, Poisson, or gamma) to be modeled. Despite those advantages, DGLM method is computationally expensive and not suitable for genome-wide *trans*-evQTL detection for considerable costs in terms of computing time. To solve this problem, a fast scanning approach called Squared residual Value Linear Modeling (SVLM) was applied for the genome-wide trans-evQTL detection. The SVLM method consists of two steps. First, a regression analysis is applied where the trait value is adjusted for a possible SNP effect and other covariates. Second, regression analysis is applied to the squared residuals obtained from the first stage, using the SNP as the

predictor. Using genotype and gene expression data from 210 HapMap individuals, we showed that the gene expression variances, as opposed to mean, also have a strong association with genotypes, both in *cis* and *trans*. Although it is just an initial step to understand the effect of genetic factors on gene expression variances, the conclusion from this study is that gene expression variance is likely to be genetically controlled. The method we adopted in this study allows us to explore the relationship between genetic variants and gene expression variance from one tissue type and infer them to the other tissue types. It is, therefore, fair to ask several key questions to expand our understanding of the extent to which, and in what ways, genotypes influence gene expression variability. (1) our previously evQTL detection was based on the DGLM using a single data set, the results need to be further validated using additional data sets and multiple tissue types to validate whether gene expression variance is really under genetic control or evQTLs are just a statistical phenomenon, (2) it remains unclear what genetic and/or environmental conditions contribute to the creation of an evQTL, which is essential information for understanding the mechanisms that underlie the existence of evQTL, (3) In our previous study, we focus on the identification of single locus effect, which is the association between the expression of a single gene and a single locus. However, increasing evidence shows that a lot of gene expression traits are associated with multiple loci [72,73], which could also explain the variability of gene expression. In addition, epistasis has emerged as an important factor to understand the multiple loci effect [72,74] and the phenotypic variability of a population can be increased by epistasis [61,75]. Epistasis was initially defined by Bateson [76] to describe one

phenotype is determined by the interaction effects of two genes. The definition of epistasis varies a lot since its introduction. Currently, epistasis is mostly defined as a masking effect whereby the effect of a genotype on a phenotype is prevented by anther genotype [77]. An example of the epistatic effect is shown in **Table 1.1**, which are the possible outcomes of hair color in mice for two genetic loci, A (alleles A and a) and B (alleles B and b). The effect of genotype at locus A is masked by the genotype at locus B, where individuals with any copy of the B allele have a grey color. In our evQTL study, gene expression is emerged as an "intermediate" phenotype, epistasis is defined as the interaction between genetic loci that control the expression of a single gene.

Table 1.1 Example of the epistatic interaction between two genetic loci

	Genotype at locus B		
Genotype at locus A	B/B	B/b	b/b
A/A	Grey	Grey	Black
A/a	Grey	Grey	Black
a/a	Grey	Grey	White

1.6 The influence of rare genetic variants on disease and gene expression

Over the past several years, the rapid advance of NGS technology has made population level sequence or genotype data sets broadly available, and has revealed the existence of a huge store of previously unknown rare variants (MAF < 1%) in human

populations [78-81]. The 1000 genome project showed that there around 30,000 to 150,000 low-frequency and rare genetic variants per individual. Compare to common variants, rare variants are relatively new mutations and usually have a weaker correlation with other variants [82]. The impact of rare genetic variants on diseases has obtained much attention as an increasing number of disease associated variants are reported by different studies. Although NGS technology provides a great opportunity to study the contribution of rare genetic variants to diseases, detecting rare genetic variants is still a challenge due to the huge cost of sequencing a large number of individuals. To address this shortcoming, several strategies are applied to decrease the cost. For example, exome sequencing [83], this strategy based on two considerations, on one hand, exomes only count for 1%-2% of the genome, which will decrease the sequence cost significantly, on the other hand, many identified causal genetic variants for diseases are located in exome regions. Moreover, classical association tests used for the study of common genetic variants have limited statistical power when applied to the rare genetic variants study unless samples or effect sizes are very large. In light of those limitations, several statistical methods are developed, (1) burden tests [84-87], which assess the cumulative effects by summarizing rare genetic variants information in a region. All burden tests are based on the assumption that all rare variants are associated with phenotype with same effect size and direction and, therefore, is powerful for such rare variant set. However, oftentimes, it is not surprise that the influence of rare variants with distinct effect sizes and directions, (2) variance-component tests, including the sum of squared score (SSU) test [88], C-alpha test [89], and SKAT test [90]. These methods are more powerful than

burden tests for variants with different effects sizes and directions by evaluating the distribution of the aggregated score test statistics for a variant set using distinct test model [91], (3) integrative test [92], which combine the advantages of Burden tests and variance-component tests. Despite those achievements, few frameworks or statistics are available for assessing the impacts of rare genetic variants to gene expression. The only exception is the study by Li et al. [93], in which they found that rare variants were enriched in larger effect eQTLs and splicing quantitative trait loci (sQTLs) which indicated that rare variants are likely associated with gene expression. However, the method used in Li's study based on the full genome sequencing data within a family with limited power for unrelated individuals. Therefore, we need a new analytical approach for studying the possible effects of rare or private mutations on gene expression at the n=1 level.

1.7 Project rationale

Identifying the influence of genetic variants on gene expression variation is the primary focus of the field of quantitative genetics. Mot available methods are limited to identify mean differences of gene expression. However, increasing evidence shows that genetic variants may also contribute to the variances of gene expression. Moreover, the potential impact of rare genetic variants (or private SNP) on gene expression is difficult to study as most available methods are powerful for the common genetic variants study.

I attempt to answer these questions based on the recently accumulated data. Chapter II describes the study on the genetic influence on variable gene expression by using expression data from a large twin cohort. Firstly, I performed a global evQTL

mapping with three different tissues (lymphoblastoid cell lines, skin, and fat) to identify genetic loci that contribute to gene expression variance. To show the influence of genetic background on expression variability, I measured the relative difference between pairs of dizygotic twins and pairs of monozygotic twins. Moreover, I investigated the genetic interactions in the formation of evQTL through additive effects.

Chapter III is an extended discussion of how evQTLs are created. In this chapter, I demonstrate two distinct modes of action (epistasis and decanalization) that create the instances of evQTLs in humans. To validate the decanalization mode, I then measured discordant expression between monozygotic twins, as well as the level of transcriptional noise in individual clonal cell lines.

In complementary to Chapters II and III, which center on the impact of common genetic variants on gene expression, Chapter IV focuses on the impact of rare or private genetic variants on gene expression. Specifically, I used a multivariate approach to first identify outlier individuals that show markedly different gene expression from the majority of a population and then quantified the contributions of private SNPs to aberrant gene expression in these outliers.

CHAPTER II

ADDITIVE, EPISTATIC, AND ENVIRONMENTAL EFFECTS THROUGH THE LENS OF EXPRESSION VARIABILITY QTLS IN A TWIN COHORT*

2.1 Introduction

Variation and variability are central concepts in biology [94]. Although often used interchangeably in the scientific literature, the two are not synonymous. Variation refers to the differences among individuals, whereas variability refers to the potential of a population to vary [95,96]. In many cases, greater phenotypic variability (e.g., transcriptional noise) is disadvantageous [97-99] unless it gives rise to greater organismal plasticity—first at the level of an individual organism and eventually at the population level. Genetic factors resulting in more variable phenotypes become favored when they enable a population to more effectively respond to environmental changes [100-103]. Thus, understanding to what extent and in what ways genotypes influence phenotypic variability is of fundamental importance.

Much effort has been focused on identifying genetic loci such as eQTL [104-109], that affect the average value of a phenotype, while ignoring those that affect the variance of a phenotype. However, there is increasing evidence across species for genetic loci that affect the variance of phenotype [69,110-114]. Recently we introduced the concept of expression variability QTL, or evQTL [68]. By definition, an evQTL is a

* This chapter has been reprinted from: Wang G, Yang E, Brinkmeyer-Langford CL, Cai JJ* (2014) Additive, epistatic, and environmental effects through the lens of expression variability QTLs in a twin cohort. Genetics, 196:413-25, with permission from GENETICS. It is available online at http://genetics.org/content/196/2/413

genetic locus linked to or associated with genetic variation influencing the variance of gene expression in a population. To identify evQTLs, we adapted the method developed by Ronnegard and Valdar [69], based on the DGLM model [115]. The DGLM method tests for expression variances and measures the contribution of genetic variants to the expression heteroscedasticity. The DGLM method compares the fit of a full model, which takes into account the contribution of genotype to both the mean and the variance of gene expression simultaneously, and a mean model, which only takes into account the contribution of genotype to the mean, ignoring the contribution to the variance. A significant result of DGLM shows the nonrandom association between genotypes and gene expression variances. Using this method, we have conducted a genome-wide scan for evQTLs in the human genome [68].

How an evQTL is created in the first place is not clear. One possibility is that specific genetic variants disrupt the stabilizing genetic architecture that buffers stochastic variation in phenotype. As a result of such an effect of decanalization, along with the sensitizing change in the stabilizer (e.g., heat-shock protein 90), the phenotype becomes more sensitive to the external environment and varies more greatly between individuals [68,69]. Another possibility concerns the role of genetic interactions via epistatic and non-epistatic (such as additive or dominance) effects in the formation of evQTLs. It has been suggested that the variance of a quantitative trait is likely to differ based on genetic interactions [70,116]. Without extra information, however, it is extremely difficult to distinguish the contributions of genetic and nongenetic factors to variable expression of genes.

Here we investigated the roles and development of evQTLs, taking advantage of an existing dataset [117] derived from a population-based cohort of twin studies [118]. We interrogated this dataset for evQTLs, and investigated the roles of genetic and nongenetic factors in the formation of the evQTLs we identified. The twin cohort offered a unique advantage for studying the relative contributions of factors that influence expression variability. Importantly, comparing expression data of monozygotic (MZ) and dizygotic (DZ) twins allowed us to distinguish between genetic and nongenetic effects. In the following sections, we first present the descriptive statistics for expression variability in the twin cohort, subsequently describe the detection of evQTLs, and finally estimate the relative contributions of genetic and nongenetic factors to the creation of these evQTLs.

2.2 Materials & methods

2.2.1 The TwinsUK dataset

We obtained the TwinsUK dataset, including both genotype and expression data, as used in the eQTL study of [119]. Here we briefly describe the cohort and data processing performed in that study [119]. The TwinsUK cohort includes 856 female individuals of European descent recruited from the TwinsUK Adult twin registry [120,121]. Subcutaneous adipose tissue, skin tissue, and lymphoblastoid cell line (LCLs) were collected from each individual. Genotyping was performed with a combination of Illumina HumanHap300, HumanHap610Q, 1M-Duo and 1.2MDuo 1M chips. Genotypes were called with the Illuminus calling algorithm [122], and SNPs were filtered for MAF of >5%. Gene expression levels were measured in LCLs, skin, and adipose [119].

Expression profiling of the samples, each with either two or three technical replicates, was performed using Illumina Human HT-12 V3 BeadChips (Illumina). All samples were randomized before array hybridization, and replicates were hybridized on different BeadChips. Raw data were imported to Illumina BeadStudio software, and probes with less than three beads present were excluded. Log_2-transformed expression signals were normalized separately per tissue, with quantile normalization of the replicates of each individual followed by quantile normalization across all individuals [123].

In this study, we used available gene expression data for both individuals of a twin pair. All 48,804 probe sequences were mapped by BLAST to the reference genome (hg18), and probes found to map to more than one location were not used. Polymorphisms in the target mRNA sequence can greatly affect the binding affinity of microarray probe sequences, leading to false-positive and false-negative signals with any other polymorphisms in linkage disequilibrium (LD) [124]. In order to control for this, we used a comprehensive compendium of SNPs in European ancestry (CEU) of the 1,000 Genomes Project [125] to remove an additional 13,600 probes found to anneal in regions with SNPs present at an MAF of 5% or greater. Similarly, probes mapping to non-autosomal locations were excluded from further analysis. Finally, 35,078 probes were left for our analysis.

The coefficient of variation (CV) is used as a normalized measure of the dispersion of expression distribution [69,126,127]. The CV was computed as

$$CV = \frac{\sigma}{\mu},$$

where σ and μ are the standard deviation and the mean of gene expression levels, respectively. LD block plots were obtained by using HaploView [128].

2.2.2 Identification of evQTLs using the DGLM method

First we used the Fligner-Killeen (F-K) test filter to greatly reduce the number of SNPs for computationally intensive model fitting. We then adapted the DGLM method [115] to test for inequality in expression variances and measure the contribution of genetic variants to the expression heteroscedasticity. We considered the following model:

$$y_i = \mu + x_i \beta + g_i \alpha + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2 \exp(g_i \theta)),$$

where y_i indicates a gene expression trait of individual i, g_i is the genotype at the given SNP (encoded as 0, 1, or 2 for homozygous rare, heterozygous and homozygous common alleles, respectively), ε_i is the residual with variance σ^2, and θ is the corresponding vector of coefficients of genotype g_i on the residual variance. Age of subjects and the batch of data collection were modeled as covariates x_i. With this full model, both mean and variance of expression y_i were controlled by SNP genotype g_i. We coded the fitting procedure using the DGLM package in R. A snippet of R code for the DGLM analysis can be obtained from the Supporting Theory of ref [69]. We assumed that the input gene expression data were approximately normally distributed, conditional on the evQTL and covariates, and set family Gaussian in the DGLM R code to specify the error distribution and link function used. We tested for each input probe-SNP pair and obtained two P-values: $P_{dispersion}$ and P_{mean}, for the effects of genotypes on the variance and the mean of expression levels, respectively [69]. Probe-gene pairs that did not make the algorithm converge during computation were discarded. To control for the

effect of outlier expression data points, permutation tests [107] were conducted for all $P_{dispersion}$ significant pairs. Specifically, for each probe-SNP pair, we performed 10,000 permutations of expression phenotype relative to SNP genotypes. An association was considered significant if the P-value from the analysis of the observed $P_{dispersion}$ was lower than the threshold of the 0.001 tail of the distribution of the $P_{dispersion}$ from the 10,000 permutations ($P_{permutation} < 0.001$).

2.2.3 Single-cell expression and mRNA decay rate

The expression level of 96 genes was measured in 1,440 single lymphoblastoid single cells by qPCR assays in another study [129]. We used these data to compute the CV of expression of the same gene in different cells. The mRNA decay rates of 16,823 genes were estimated in 70 human LCLs [130]. We obtained the mRNA decay rate data to compute the average mRNA decay rate for each gene among these LCL samples.

2.2.4 Estimation of the fraction of isoform transcription using MISO

We obtained genotype data for 43 samples of CEU from the phase 1 release of the 1,000 Genomes Project [125]. Short sequence data produced for RNA-seq studies of the LCLs from the same 43 individuals were accessed through GEO (Gene Expression Omnibus) accession number GSE19480 [108]. The Sequence Read Archives (SRA) files were downloaded and subsequently converted into FASTQ files using the NCBI SRA toolkit program, fastq-dump (v 2.1.16). To estimate FPKM (fragments per kilobase of exon per million fragments), RNA-seq short reads were mapped to reference genome (hg19) using Tophat2 (v 2.0.1) [131]. We then used the mixture-of-isoforms (MISO) [132] isoform-centric model (which estimates expression level of whole transcripts) to

assess expression levels of different isoforms by quantifying the presence of alternatively spliced exons. Mapped data were analyzed with the default parameters using the compute-genes-psi function and summarized using the summarize-samples function.

2.2.5 Identification of interacting SNPs

We used a two-step procedure to identify SNPs that may "interact" with evSNPs. Assuming the interaction between the SNP to be identified and an evSNP is additive, we first partitioned individuals into L and S groups according to genotypes of the evSNP, which were associated with large (L) and small (S) variances of gene expression. Next we scanned genome-wide SNPs. For each SNP, we computed genotype heterozygosity among individuals in L and S groups using

$$Het_L = p^2_{AA\_L} + p^2_{Aa\_L} + p^2_{aa\_L}$$

and

$$Het_S = p^2_{AA\_S} + p^2_{Aa\_S} + p^2_{aa\_S}$$

, respectively, where P_{AA} , P_{Aa} and P_{aa} are frequencies of three possible genotypes defined by the scanned SNP. All SNPs were then ranked by $Het_L - Het_S$, and top 100 SNPs with the largest value were taken to next step. In the next step, a typical eQTL (not evQTL) analysis was conducted among individuals of the L group. For each top SNP with high genotype heterozygosity difference, a simple linear regression [107] was performed between the SNP's genotypes and gene expression. The most significant SNPs were retained after applying an arbitrary P-value cutoff = 0.0005 and were

reported as candidate interacting SNPs. To maintain sample independence, we used only one individual from each twin pair for this analysis.

2.3 Results

2.3.1 Expression and genotype data

To investigate the genetic influences underlying variable gene expression, we revisited the published expression data [117] of the MuTHER (Multiple Tissue Human Expression Resource) project [133]. In that study, gene expression was measured for LCL, adipose tissue (subcutaneous fat), and skin (tissue biopsies) using Illumina Human HT-12 V3 BeadChips. These tissues were sampled from a cohort of 856 female twins from the TwinsUK adult registry, including 154 MZ twin pairs, 232 DZ twin pairs and 84 singletons [118]. After quality control, expression data for 825 (adipose and LCL) and 705 (skin) individuals were retained [117]. For each tissue, we downloaded the processed MuTHER expression data files deposited at ArrayExpression (http://www.ebi.ac.uk/arrayexpress/) using accession E-TABM-1140. The data were the quantile-normalized $\log_2$-transformed expression signals. Quantile normalization was performed first across the replicates of a single individual and then across all individuals as described in [117]. Along with the expression data, we also obtained the genotype data of this cohort [117]. In our analysis, all available twin pairs with complete expression and genotype information were included, corresponding to 134 MZ and 195 DZ pairs with LCL profiles, 139 MZ and 188 DZ pairs with adipose profiles, and 105 MZ and 148 DZ pairs with skin profiles. Members of the TwinsUK cohort have health and lifestyle characteristics that are comparable to those of population singletons [134].

Because of this, we were able to use this cohort as a representative general population to investigate both genetic and nongenetic factors behind expression variability in this study.

2.3.2 Expression variability in the twin cohort

Here we present basic, descriptive statistics for expression data (independent of genotype information), with particular attention to disparities in gene expression among individuals. We chose to focus on the LCL data for this analysis, due to the availability of additional expression-related statistics (such as single-cell expression data and mRNA decay data).

We used the quantile-normalized $\log_2$-transformed expression data in all analysis throughout the paper unless otherwise stated. From these data, we first determined that expression values for most probes (n = 35,078) approximately fit the normal distribution: 97% of probes were with a skewness between -0.80 and 0.80 and a kurtosis of ~3.0 (**Figure 2.1A**); less than 7% of probes were rejected by Shapiro-Wilk test of normality with Bonferroni adjustment to the level of $\alpha = 0.01$. These justified the use of the Gaussian error distribution and link function in our DGLM model (Materials and Methods). Retrospectively, we showed that the profile distributions for evQTL probes are approximately normal before and after Box-Cox transformation (**Figure 2.1B**).

To measure the level of dispersion of gene expression values, we computed the CV for each probe. The CVs ranged from 0.0024 [for ILMN_1765043 (RPL38)] to 0.2115 [for ILMN_1715169 (HLA-DRB1)], with a median of 0.0154. The distributions of CVs measured in sub-cohorts are indistinguishable from one another such as when

comparing one set of MZ twins with the other set (i.e., MZ 1 vs. MZ 2) or comparing a set of MZ twins with a set of DZ twins (e.g., MZ 1 vs. DZ 1)(**Figure 2.2A**). Probe data points are located along or close to the 1-1 diagonal line in the CV-CV scatter plot for the majority of probes, regardless of the CV-CV comparison between MZ 1 and MZ 2 or between MZ 1 and DZ 1 (**Figure 2.2B**). These results indicate that the extent and overall distribution of expression variability measured between individuals across different MZ and DZ cohorts are highly similar when all genes are taken into account.

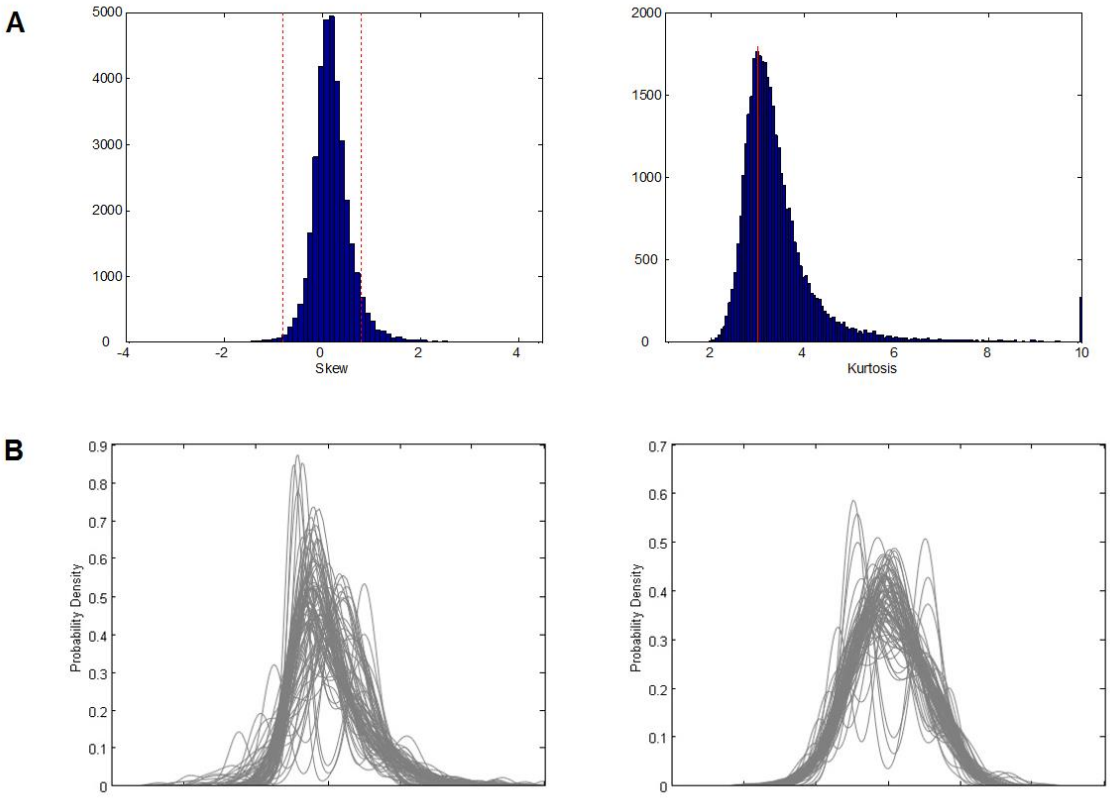


Figure 2.1 Normality of expression data measured in LCLs. (A) Distributions of skewness and kurtosis. Red dashed lines indicate -0.8 and +0.8 skewness; red solid line indicates kurtosis = 3. (B) Profile distributions of expression data for selected probes (i.e., probes involved in evQTLs). (Left) Quantile-normalized expression data; (Right) Box-Cox normalized expression data.

Next, we measured expression differences between each pair of twins. For each probe, we computed the relative mean difference (RMD) in expression between MZ twin pairs and between DZ twin pairs, separately. For a pair of MZ twin, for example, the RMD was computed using

$$RMD = \frac{\frac{1}{2} \cdot |y_{MZ1} - y_{MZ2}|}{\bar{y}},$$

where $\bar{y}$ is the arithmetic mean the levels of gene expression for that MZ twin pair (designated as y_{MZ1} and y_{MZ2}). For most probes, the median RMD of expression between DZ pairs is larger than it is between MZ pairs, as indicated by the fact that most genes are located above the 1-1 diagonal line in the scatter plot (**Figure 2.2C**). That is to say, the normalized difference in gene expression between DZ pairs (DZ 1 and DZ 2) tends to be larger than that between MZ pairs (MZ 1 and MZ 2), suggesting that genetic factors influence expression variability for most of these genes.

To determine the influence of single-cell expression variability on population-level expression variability, we computed the CVs of expression for a selection of genes, whose expression levels have been measured in single LCL cells [129]. No correlation between the single-cell CVs and the between-individual CVs measured was detected for MZ 1 (Spearman's correlation test, P = 0.21, n = 59; **Figure 2.2D**). This suggests a limited contribution of single-cell expression variability (or transcriptional noise at the single-cell level) to the variability between individuals (or transcriptional noise at the population level).

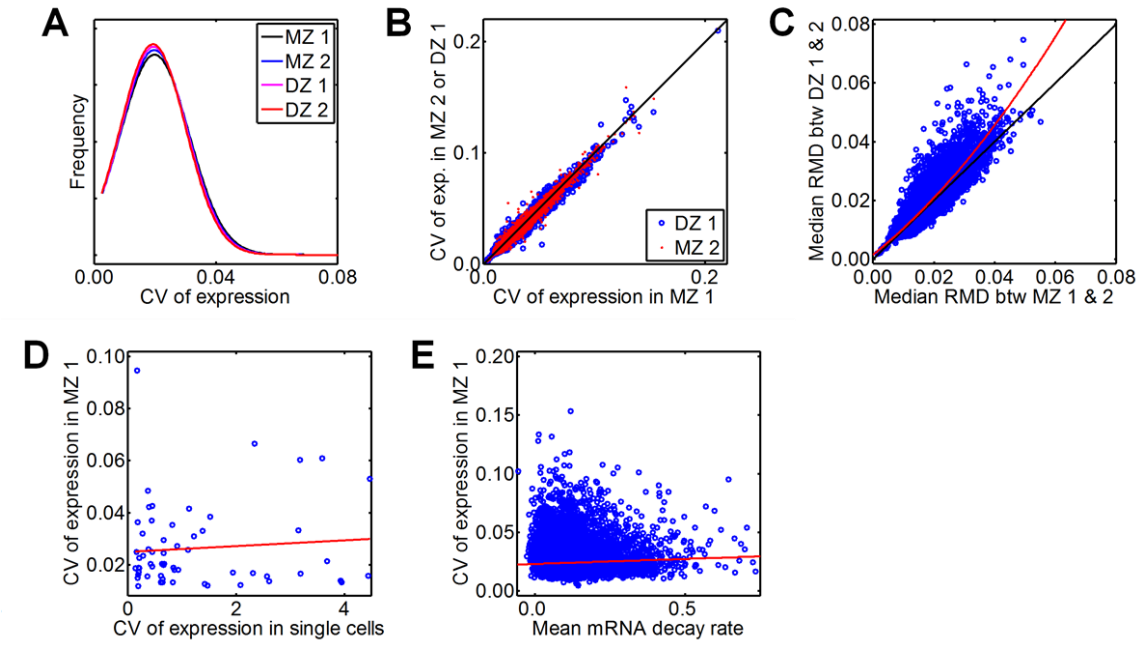


Figure 2.2 Distributions of expression variability in LCLs. (A) Distribution of CVs of gene expression (probe n = 35,078) measured in MZ and DZ twins. MZ 1 is the set of first pairs of all MZ twins and MZ 2 is the set of second pairs of all MZ twins. Similarly, DZ 1 is the set of first pairs of all DZ twins and DZ 2 is the set of second pairs of all DZ twins. (B) Scatter plot of CVs of gene expression (probe n = 35,078) in MZ 1 against those in MZ 2 (blue) or DZ 1 (red) cohorts. (C) Scatter plot of median RMD between pairs of MZ twins against median RMD between pairs of DZ twins. Each blue dot indicates a single expression probe (or a gene) and the position of the blue dot indicates the median value of RMD of expression between all MZ pairs (MZ 1 − MZ 2) on the x-axis and that between all DZ pairs (DZ 1 − DZ 2) on the y-axis. The red line is based on quadratic regression to show a more pronounced difference between MZ and DZ with greater RMD. (D) Scatter plot of CVs of gene expression (n = 59) in single cells against CVs of gene expression in MZ 1. (E) Scatter plot of mean mRNA decay rate against CVs of gene expression in the MZ 1 cohort. The red line is based on the linear regression

Finally, we hypothesized that variable gene expression may be due to different mRNA decay rates for different genes. To test this, we used the mRNA decay rate data from the study of Pai et al. [130] et al. The correlation between mean mRNA decay rate and CV of expression among genes is not specific as shown by the opposite signs of two correlation coefficients: Spearman's ρ = -0.027 (P = 0.00498) and Pearson's r = 0.044 (P = 4e-6, n = 11,083; **Figure 2.2E**). Thus, gene expression variability showed no signs of correlation with the mRNA decay rate of genes.

2.3.3 Genetic variants underlying expression variability

To systematically assess the genetic influence on expression variability, we identified genome-wide evQTLs using the method we previously established [68]. We focused on *cis*-acting evQTLs by limiting our search to those SNPs that flanked probes within 1.0 Mb on either side.

After filtering for quality control (Materials and Methods), a total of 35,078 probes were available for analysis. On average, each probe corresponded to 1,212 SNPs in the 2-Mb *cis* region (i.e., 6 SNPs per 10 kb). For each SNP-probe pair, we conducted a three-step test to determine the evQTL relationship as described previously [68]. Briefly, we first tested for the homogeneity of variances in gene expression among different genotype groups using F-K test [135]. Only those SNPs with a $P < 0.01$ [following [136]] were carried on to the next step of analysis. We then applied the DGLM method [69] to each SNP-probe pair, ultimately computing $P_{dispersion}$ for a total of 1,251,611 SNP-probe pairs. To account for multiple tests performed between these probe-SNP pairs, we used the threshold of $P_{dispersion} < 1 \times 10^{-8}$, which is roughly equivalent to Bonferroni adjusted $P < 0.01$, to assess the genome-wide significance. Finally, we conducted permutation tests for each significant SNP-probe pair to control for the influence of outlier data points on the DGLM results (Materials and Methods). The detection of evQTLs was performed independently for each of the two sets of twin data. Assignment of individual twins to each data set was purely random and did not influence the overall results in any substantial way (data not shown). Each evQTL detected with one twin data set and was then validated with the other data set to confirm

its authenticity. For all three tissues, concordance was prevalent (**Figure 2.3A**) and the cases of discordance were mostly due to outliers present in one set of twin data but not in the other set. The direction of effect (association with increased or decreased gene expression variability) was the same between the two set of twin data for all evaluated SNPs.

A total of 99, 79, and 56 genes were identified and confirmed to have at least one validated *cis*-evQTL SNP (or evSNP for short) in LCLs, fat and skin, respectively. This corresponded to 8 evQTL genes shared in all three tissues (**Figure 2.3B**). One of these shared evQTL genes, SEMA4G, is given as an example to illustrate the consistent influence of genotypes on the variance of gene expression across the three tissues (**Figure 2.3C**). All evQTLs shared across tissues showed the same directional effect, defined as on either increased or decreased the variance of gene expression. That is to say, the directionality of evQTL effects is not tissue- or cell-type specific.

Given that many evQTL genes have more than one *cis*-evSNP, we examined the structure of haplotypes of these multiple *cis*-evSNPs. We found that *cis*-evSNPs of the same gene are likely to be located within same LD block and that typically these blocks contained only few prominent haplotypes (see **Figure 2.4** for an example involving gene *ALG11*). This suggests that multiple evSNPs are likely to be linked with the same causal variant. We furthermore found that, compared with ancestral alleles, derived alleles of evSNPs are more likely to be associated with greater expression variability (Fisher's exact test: P = 0.0036, 0.022 and 0.036 for LCLs, skin and fat, respectively).

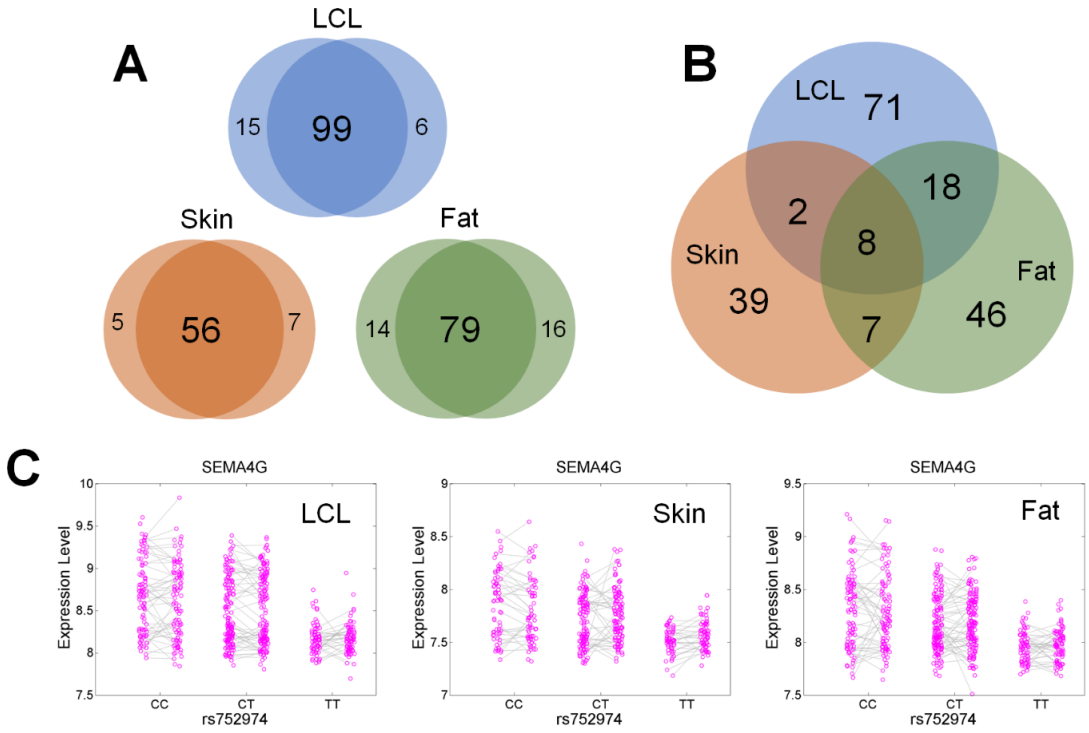


Figure 2.3 Numbers of evQTL in LCL, skin, and fat. (A) Venn diagrams of evQTL genes detected in two groups of twin sets. Each group of the twin sets is composed of one set of unrelated twin individuals. Overlapping areas of the Venn diagrams contain numbers of validated evQTL genes identified with both sets of twins. (B) Numbers correspond to evQTL genes within a subset of tissues. (C) One example of evQTL shared by all three tissues: evQTL at SEMA4G.

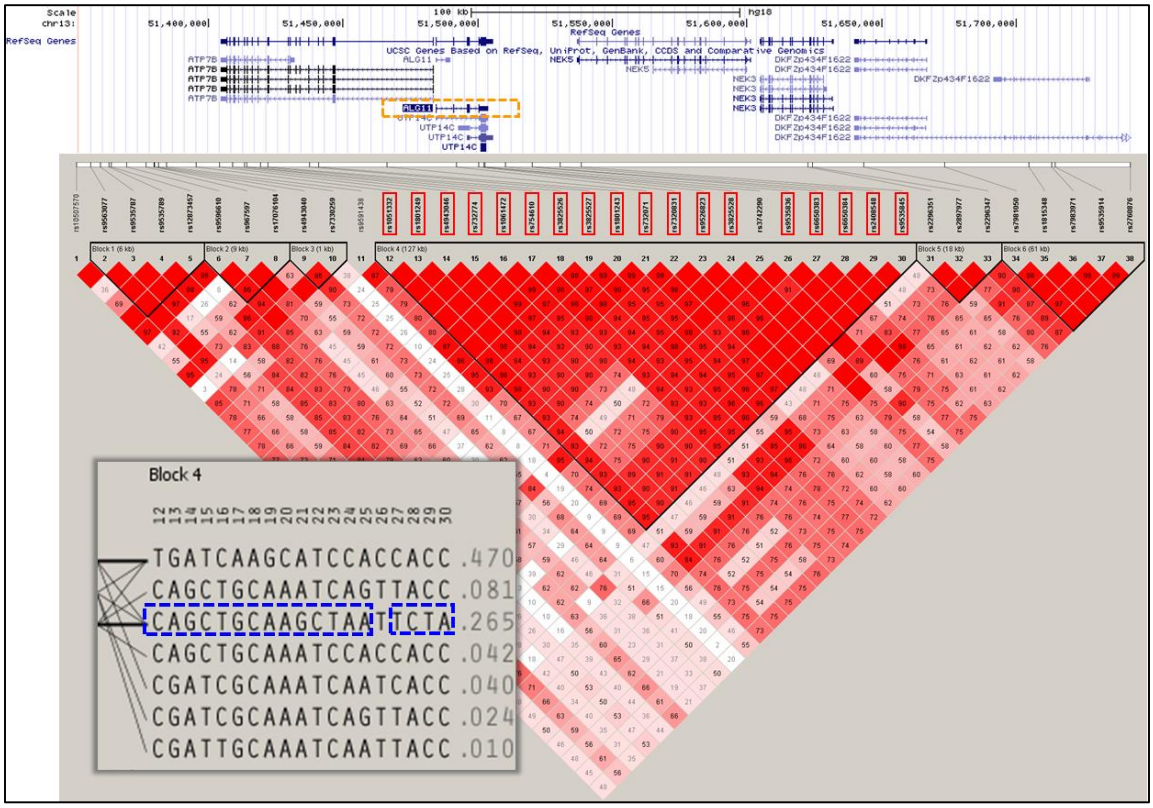


Figure 2.4 LD patterns of the genomic region surrounding evQTL at ALG11. The entire region of the analysis included 38 SNPs over a ~400 kb span. The *cis*-evSNPs are indicated with red boxes. The haplotypes in the LD block accommodating evSNPs are displayed in the insert, with corresponding haplotype frequencies. Of note, alleles of evSNPs resulting in a larger variance of gene expression are allocated in one haplotype highlighted with the blue box.

2.3.4 Dissecting the genetic and nongenetic effects of evQTLs

Twin data facilitate the dissection of the contributions of genetic and nongenetic factors to gene expression. Phenotypic variability measured between pairs of DZ twins is expected to be larger than that between pairs of MZ twins, as the phenotypic difference between DZ pairs may result from both genetic and environmental (nongenetic) effects while differences between genetically identical MZ pairs are attributable to varying environmental effects on the two twins, assuming that the environments influencing MZ and DZ twin individuals are essentially identical. **Figure 2.5** depicts the difference in

expression level of evQTL gene AXIN2 in three genotypes (GG, AG, and AA) defined by rs740026. **Figure 2.5A** and **B** illustrate genotypes at rs740026 by linking the two data points for each twin pair by a straight line: **Figure 2.5A** shows genotype similarities between MZ twins, while in **Figure 2.5B**, similarities between DZ twin pairs are shown. Note that linkers between DZ twin pairs with different genotypes at the SNP site (i.e., DZ 1 ≠ DZ 2) are not plotted. The expression difference between a pair of twins can be visually quantified by the slope of the straight line: a steeper line reflects a more dissimilar expression level between the twins. In the case of AXIN2, it is apparent that expression differences between DZ pairs tend to be larger than between MZ pairs. This is especially true for the AA genotype group, which shows a larger variance in expression between individuals.

For each evSNP and its associated genes in LCLs, we computed the RMD in gene expression between all pairs of MZ or DZ twins, as long as the genotypes of two individuals of the pair of twin were both identical to each other and homozygous at the SNP site. By definition, (one of alleles of) evSNP is associated with either larger (L) or smaller (S) variance in gene expression. Thus, the RMD values (for evSNPs and associated genes) were separated according to whether homozygous genotypes defined by evSNPs were associated with larger (L) or smaller (S) variance in gene expression. The cumulative distribution functions (CDFs) of these RMD values were plotted (**Figure 2.5C**). The curves were based on the RMD values calculated between all possible twin pairs for all evSNPs and genes, and classified into four groups: MZ-S, MZ-L, DZ-S, and DZ-L. The MZ-S and DZ-S groups included pairs whose genotypes showed a small (S)

amount of variance, while the MZ-L and DZ-L groups included pairs whose genotypes were associated with large (L) variances. In the end, the four groups, MZ-S, MZ-L, DZ-S, and DZ-L, contained 3,629, 2,548, 3,825, and 2,520 RMD values, respectively. We found that CDF curves for the large-variance groups (MZ-L and DZ-L) were shifted toward the right compared to those for small-variance groups (MZ-S and DZ-S) [Kolmogorov-Smirnov (K-S) test, all $P < 10^{-5}$]. This indicated that the distribution of RMD between twin pairs (either MZ or DZ) in the large-variance groups was significantly different from that of the small-variance groups, with a larger RMD median for the large-variance group. This difference (in RMD distribution between L and S groups) remained even when we randomly assigned the identities of MZ and DZ pairs (see insert of **Figure 2.5C**). Together, these results suggested that the increased discrepancy in gene expression between twin pairs (shown as a larger median RMD) contributed to the elevated variability in expression, which is true for both MZ and DZ twins. Because MZ twins are genetically identical, the increased RMD between MZ pairs was likely due to an increased sensitivity of gene expression to environmental factors.

More importantly, we found that there is a significant discrepancy in distribution of RMD between MZ and DZ: DZ groups tended to have larger RMD values than MZ groups. This trend applied to both L and S groups, but was more salient in the L group (all K-S test, $P < 0.01$) (**Figure 2.5C**). These results suggested that the different genetic backgrounds resulted in a larger difference in gene expression between DZ twin pairs, which is more pronounced than that observed between MZ twin pairs.

For comparison, we randomly selected the same number of genes and *cis*-SNPs and conducted the same analysis of RMD distribution. There was no difference between CDFs of RMD in these non-evQTL genes regarding either MZ or DZ twins, larger or smaller variance groups, as well as before or after shuffling of the twin identities. CDFs of all groups were more similar to each other (K-S test, all $P > 0.025$, except between MZ-S and DZ-L, $P = 2.94e\text{-}4$, **Figure 2.5D**). That is to say, the influence of genetic and/or environmental effects on variable expression was not detected at the genomic level for all genes, but was limited to evQTL regions.

Finally, we repeated the CDF analyses using the RMD values computed from the Box-Cox normalized $\log_2$-transformed expression data, as well as using the absolute difference (instead of RMD) in gene expression. In both cases, we obtained results highly similar to those obtained above (**Figure 2.6**), which supports the robustness of the results presented above.

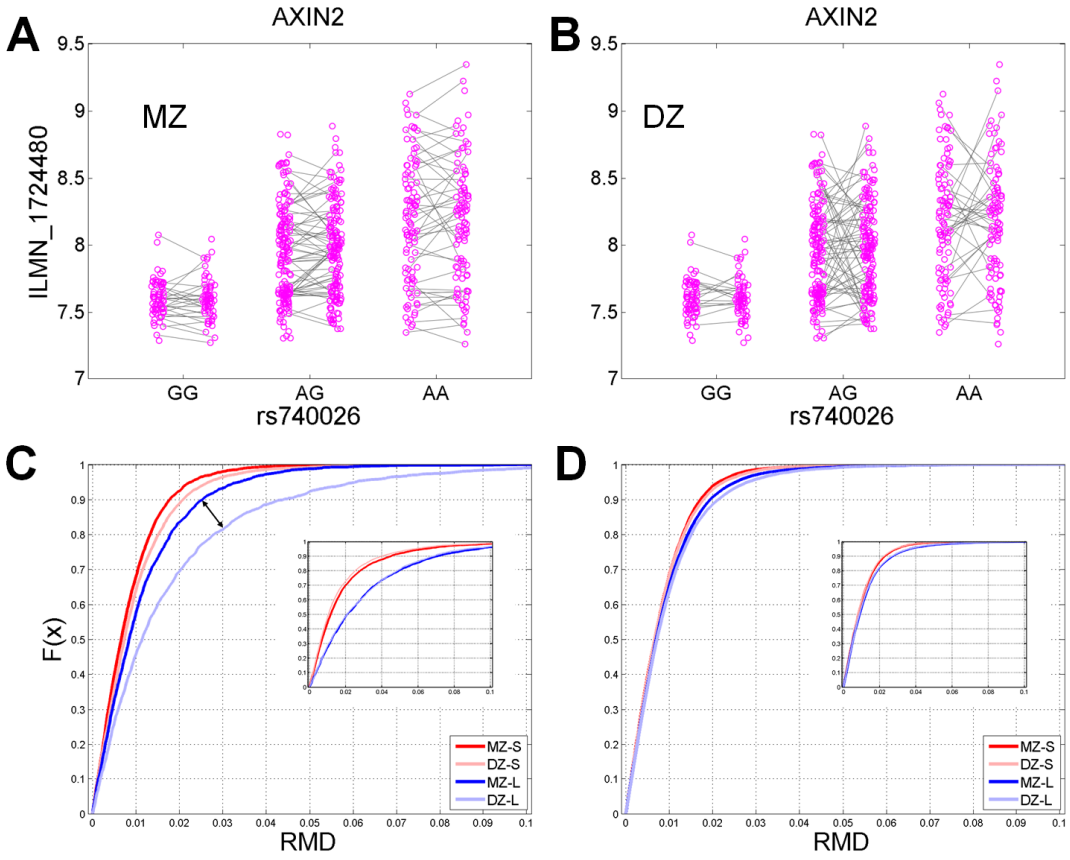


Figure 2.5 Dissection of genetic and nongenetic effects of evQTL using twins data. (A) The evQTL between AXIN2 and rs740026. The expression data points from pairs of MZ twins are linked. (B) Same as A except that DZ twins are linked. (C) CDFs of RMD between twins classified into four groups, namely MZ-S, DZ-S, MZ-L, and DZ-L (see main text for definitions). The double arrow highlights the highly significant discrepancy in RMD distribution between MZ-L and DZ-L (K–S test, $P < 0.01$). The insert shows the same CDFs of RMD recomputed after randomly shuffling identities of corresponding MZ and DZ pairs. (D) Same as C except that data are randomly sampled from non-evQTL genes.

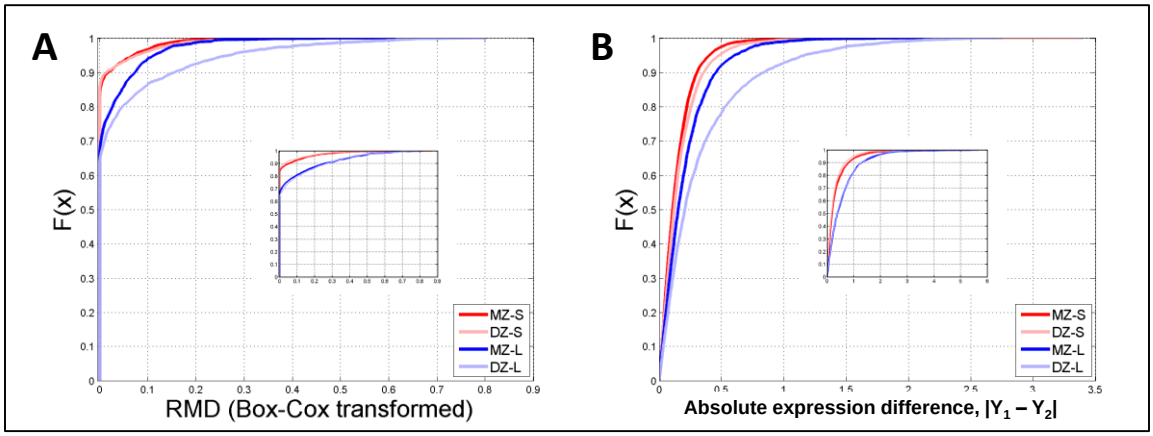


Figure 2.6 Comparison between results of the CDF analysis for the expression difference between twin pairs in evQTL genes. (A) Results obtained using RMD of Box-Cox normalized log2-transformed data between twin pairs. (B) Results obtained using the absolute difference of log2-transformed data between twin pairs.

2.3.5 Validation using RNA-seq data and SNPs of the 1,000 Genomes Project

We obtained genotype data for fully-sequenced samples of CEU from the phase 1 release of the 1,000 Genomes Project [125], along with short reads from RNA-seq experiments in LCLs for these same individuals (n = 43)[108]. After mapping the short reads, we estimated the expression level in FPKM for all genes. For the same evQTL gene-SNP pairs detected in LCLs, we plotted the relationships between genotype and expression for each. Even with as few as 43 data points, many evQTL relationships could be recognized by visual inspection. To examine expression variability in isoforms, we used MISO [132] to compute percent-spliced-in (Ψ) values for all known isoforms of each evQTL gene (Materials and Methods). We could not observe any clear pattern, perhaps due to the limited size of samples. But, in MYH11, the increased expression variability seems linked to the higher heterogeneity of isoform expression (**Figure 2.7**).

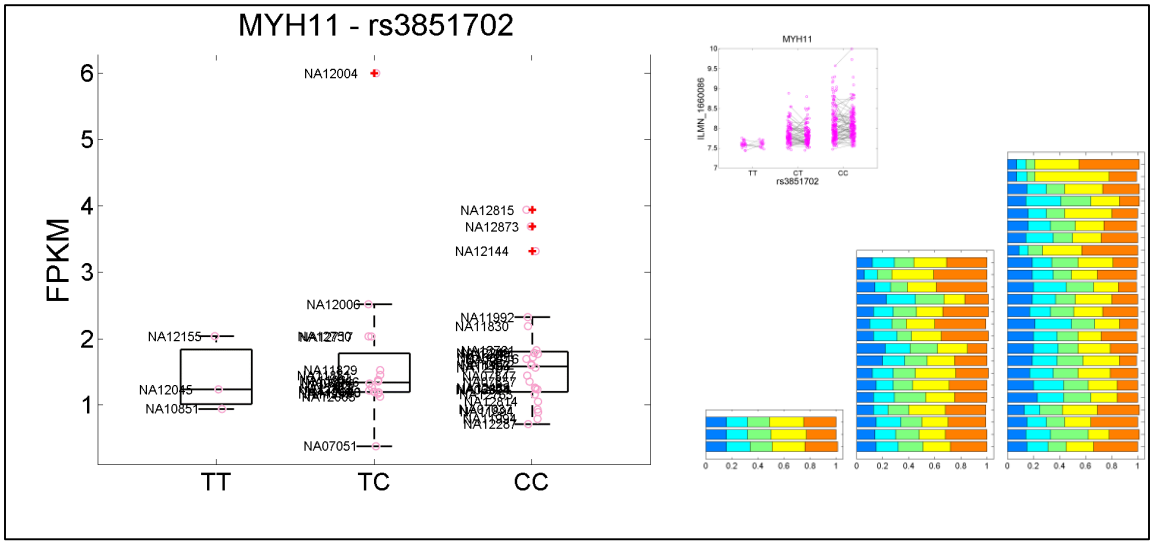


Figure 2.7 An example shows the possible association between gene expression variability and heterogeneity of isoform expression. Left panel is the evQTL relationship between MYH11 expression and genotypes of SNP rs3851702 depicted using RNA-seq data [15] and the 1,000 Genomes Project genotype data [37]. The insert is the evQTL relationship between the same evQTL gene and SNP depicted using TwinsUK data. Right panel is the fraction of each isoform expression estimated as the percent spliced in (Ψ) values using MISO [38] to the RNA-seq data.

2.3.6 Partially-linked SNPs contribute to variable gene expression

Recent theoretical work showed that the within-genotype variance of a quantitative trait varies when a non-additive genetic interaction or epistasis is present [116,137]. Alternatively, variance of a quantitative trait may result from the interaction between genetic variants additively associated with the mean of the quantitative trait. To discriminate between these alternatives, we employed a two-step procedure to identify SNPs partially associated with (or interacting) with evSNPs through an incomplete haplotype structure (Materials and Methods). In an ideal scenario (**Figure 2.8A**), the genotype heterozygosity of the partially-linked SNP is large among individuals (L-group) whose the evSNP genotype associated with larger expression variability, while, the genotype heterozygosity is small or equals zero among individual of S-group. If the

interacting SNP is associated with the mean level of gene expression, then the association between the evSNP genotype and greater expression variability is likely due to the partial association between the evSNP and the interacting SNP.

Given these considerations, we performed a genome-wide search to identify a set of candidate interacting SNPs for each evQTL SNP, and then used simple linear regression analysis to evaluate whether the potential interacting SNPs are significantly associated with gene expression among L-group individuals (Materials and Methods). For the 99 evQTL in LCLs, we identified 56 with at least one interacting SNP. Among these interacting SNPs, 54 are located within the *cis*-region of the evSNPs, with which they interact. **Figure 2.8B** presents one such relationship between evSNP rs742090 and interacting SNP rs3799378, both at BTN3A2. Individuals with CC genotype of evSNP rs742090 were further sorted by rs3799378 genotypes. Clearly, the expression level of BTN3A2 in individuals with the rs742090-CC genotype is significantly influenced by rs3799378 genotypes. The increased variability in gene expression showed in individuals with rs742090-CC genotype is caused by the heterogeneity of rs3799378 genotypes. These results suggest that local haplotype structure between SNPs contributed to the creation of evQTLs.

2.3.7 Linking evQTLs with complex disease phenotypes

Several studies have utilized eQTL data to interpret discoveries from association studies of complex traits [138-140]. Along this same vein, we identified evQTLs associated with complex traits from the catalog of published GWAS studies (http://www.genome.gov/gwastudies/). From the results of these GWA studies, we

identified 61 reported genes that are evQTL genes. In four cases, the exact same SNP was found to be both an evSNP and a marker SNP associated with risk or susceptibility of the complex trait (**Table 2.1**). Intriguingly, the "T" allele of rs8070463, associated with smaller expression variability of TBKBP1, is a reported culprit in multiple sclerosis [141] while the "C" allele for this same SNP, associated with larger expression variability, is linked with risk for ankylosing spondylitis [142].

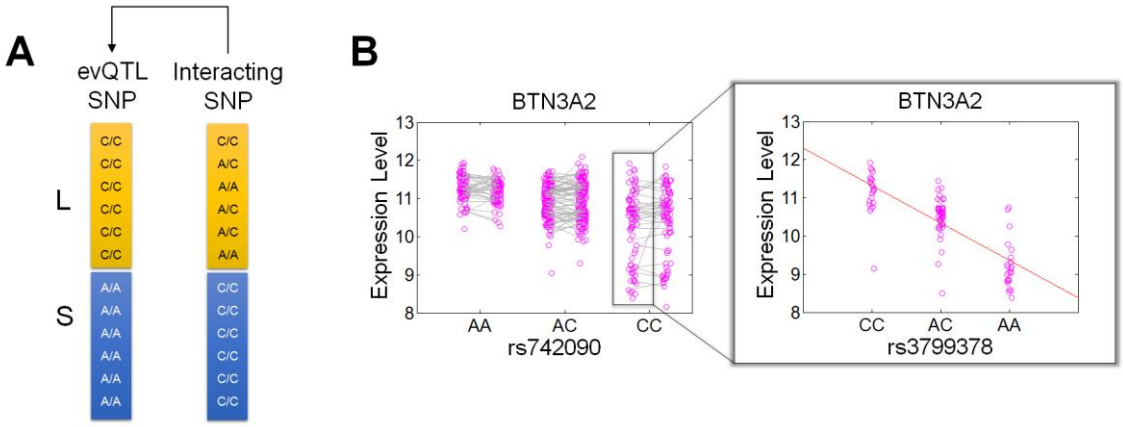


Figure 2.8 Schematic and example of an interacting SNP that helps the creation of an evQTL. (A) L indicates the group of individuals with evSNP genotype (C/C) associated with large variance in gene expression, while S indicates that with evSNP genotype (A/A) associated with small variance. The interacting SNP shows large genotype heterogeneity in the L group and small or no genotype heterogeneity in the S group. (B) Real example of evSNP rs742090 and interacting SNP rs3799378 at BTN3A2. Individuals with rs742090-CC genotype are further broken down by rs3799378 into three subgenotype groups, which are associated with different means of gene expression levels (shadowed panel).

Table 2.1 SNPs associated with gene expression variability and human complex trait. L and S indicate that individuals carrying homozygotic genotype of the risk allele have large and small variance in gene expression, respectively.

Gene (evSNP)	Tissue	GWAS complex trait	Risk allele	Reference
PAX8 (rs11123170)	LCL, Fat, Skin	Renal function related traits (BUN)	rs11123170-G^L	[143]
WDR41 (rs163030)	LCL, Fat, Skin	Caudate nucleus volume	rs163030-A^L	[144]
HCG22 (rs2517532)	LCL	Hypothyroidism	rs2517532-G^S	[145]
TBKBP1 (rs8070463)	LCL	Multiple sclerosis Ankylosing spondylitis	rs8070463-T^S rs8070463-C^L	[141] [142]

2.4 Discussion

There is empirical evidence across several species that the variance among phenotypes is genotype dependent [111,127,146,147]. Understanding genetic control of phenotypic variability has become increasingly important in evolutionary biology, human medicine, the agricultural industry and other branches of biological science [114,148]. Despite the importance, few research programs focus on genetic variants associated with trait variance, while studies of trait averages abound. Recently, a powerful statistical framework based on the DGLM model has been developed for studying the phenotypic variability of complex traits [69]. Given that gene expression is a complex trait with highly variable and heritable patterns [104,107,149], we have

previously adapted the DGLM method to investigate genetic variants controlling expression variability [68].

In this study, we further investigated the relative contribution of genetic and nongenetic (environmental) factors to expression variability and the role of these factors in the formation of evQTLs. We started by exploring basic statistics of gene expression measured in the TwinsUK cohort. For all genes, expression level dispersions were highly similar in and between both MZ and DZ twins. No correlations with expression variability were detected when compared between individuals, between single cells, or relative to the average mRNA decay rate, highlighting the marked discrepancies in variability measured at population and molecular levels. Further results showed that the discordance in expression between each pair of DZ twins was more pronounced than that between MZ twins, implying that the increased amount of genetic variation between DZ twins influences expression variability. Next, we systematically identified *cis*-acting evQTLs in three tissues of the TwinsUK cohort. Twin data greatly facilitated the validation of detected evQTLs and revealed overall robust signals that would otherwise not be appreciable in studies of non-twin design. Focusing on the detected evQTLs, we showed that the discordance in expression between DZ pairs was larger than that between MZ pairs, and further showed that the discordance in expression between MZ pairs whose genotypes were associated with large expression variability was significantly larger than that between MZ pairs whose genotypes were associated with small expression variability. It is intriguing to find that the phenotypic discordance remained even in the absence of genetic variation between MZ twins. This might be

explained by incomplete penetrance of mutations, which is frequent in isogenic model organisms in homogeneous environments [150,151]. This might also be epigenetic: for example, DNA methylation, which can be influenced by environmental factors such as diet and lifestyle, is also known to affect gene expression [152,153]. Lastly, much to our surprise, we found that more than half of evQTLs could be explained by a conceptually simple scenario in which the evSNP was occasionally associated with a nearby SNP that influenced gene expression both additively and independently. We suspect there should be many different ways of non-epistatic interaction between two or more genetic variants, such as the mode of partial association we have described here, giving rise to genotype-dependent expression variances. That is to say, the majority of phenotypic variability across individuals might be explained without invoking epistasis [154,155].

In light of our new findings, several related considerations are discussed below.

2.4.1 Methodological considerations for studying phenotypic variability

The procedure we used for identifying evQTLs [Materials and Methods, and [68]] consisted of three steps. First, the F-K test was applied to test for the heterogeneity of variances of gene expression between different genotypes and identify corresponding SNPs. Next, the DGLM method was applied to the selected SNPs. The significant results of DGLM test were then subjected to permutation tests to reduce the influence of outliers in the data. This procedure is less likely to be susceptible to issues related to multiple testing and outliers in input data, though a formal assessment of its statistical power remains to be done.

Given the flexibility of the DGLM method, we acknowledge that the results of our evQTL analysis are likely to be dependent on how the DGLM analysis was set up. For this study, we adapted the Gaussian error distribution and link function because no significant departure from normality was found in the expression data. The effect of different methods of normalization on statistical interpretation of gene expression remains subject to careful scrutiny [156-158]. For example, normalizations may perturb the covariance structure of input data or change the scale of the resulting data. Thus, the impacts of different methods of data transformation and normalization should be carefully considered in future studies involving evQTL analysis. Finally, we acknowledge that the DGLM analysis described in this paper may be influenced by the scale effect (e.g. mean-variance relationship). It is not uncommon for trait variance to change with trait mean, often causing trait skewness. If this occurs, any SNP associated with a large increase in mean expression would also be associated with an increase in variability [69] and that is why we standardized using the CV.. Analyses studying a specific phenotype and/or with a more narrowly-targeted focus than that of the broad-based study described in this paper should employ a more conservative approach in which QTL associated strictly with variance (i.e. those affecting only variability and not the mean) are identified, using the procedure proposed by Ronnegard and Valdar [69].

2.4.2 Additive vs. epistatic effect of genotypes on phenotypic variation in a population

Quantitative geneticists partition the genetic effect on phenotypic variation between individuals into additive, dominance, and epistatic components. The additive

component describes the variance associated with the independent contributions of alleles, while dominance describes the variance contributed by interactions between alleles at the same locus, and epistasis refers to the contribution of interactions between alleles at different loci. For most complex traits, quantitative genetic theory [154,159] suggests that epistasis is unlikely to contribute substantially to the between-individual variation. That is to say, most of the variation in a population will be due to the additive effects of specific allelic combinations. Yet this assertion is not without controversy. The results of empirical linkage mapping and association studies suggest that epistasis seems to explain considerable variation in complex trait characteristics within natural populations [160,161].

Our results showed that >50% of evQTLs can be explained by a partial association between haplotypes of the evQTL SNP and another SNP nearby. Our interacting SNP analysis only considered a simplistic scenario of the association. There are many other possible ways of partial associations in which SNPs interact. For example, consider the genotyped SNP "A/a" and the causative eQTL "Q/q", with only three haplotypes segregating in the population: AQ, aQ and aq (as would occur if the novel "q" allele arose on the "a" haplotype). Then the "a" SNP allele would be associated with a changed trait mean and a higher trait variance as the eQTL segregates within that genotype. If we could take all possible partial associations into account, we would anticipate that even more evQTLs could be explained by the effect of partial association, rather than epistasis. We therefore conclude that much variance in a quantitative trait may be explained by partial association between locally interacting

genetic variants, each additively associated with the trait. Our view is supported by the results of recent studies. Powell et al. [155] conducted a gene expression study using blood samples from 862 individuals from nuclear families containing MZ or DZ twin pairs, using both pedigree and genotype information. They found that the genetic architecture of gene expression is predominantly additive, with a minority of transcripts displaying non-additive effects. Hill et al. [154] evaluated the evidence from empirical studies of genetic variance components and found that additive variance typically accounts for over half and often close to 100% of the total genetic variance.

2.4.3 Detecting evQTL as a shortcut for detecting epistasis?

Detection of the variance of a quantitative trait in genetic association studies is thought to increase knowledge about the interaction between genetic variants. More specifically, detecting variability QTL (e.g., evQTL) is considered to be a shortcut for detecting genetic interactions [69,70]. So far, many methods for detecting genetic interactions are based on testing for different variances of phenotype between genotypes, with the underlying assumption that the variance of a quantitative trait is likely to differ under the influence of epistasis [69,116]. However, our new discovery that evQTLs are formed due to the partial haplotype association between SNPs refutes this assumption. As stated above, more than half (and probably much more) evQTLs could be explained by partial association between SNPs with additive effects. Both additive and epistatic effects can result in increased phenotypic variation (as schematically illustrated in **Figure 2.9**). Merely detecting the variance of a quantitative trait cannot in itself distinguish between additive and epistatic effects; thus, no specific conclusions can be

made. The relationship between partially associated SNPs, each additively associated with phenotypic variation, needs to be integrated more carefully in the study of phenotypic variability. Thus, the variance of a quantitative trait should not serve as a hallmark of genetic interaction or epistasis.

2.4.4 Phenotypic variability and implications in complex traits and diseases

High-throughput sequencing and genotyping technologies have spurred an increasing number of studies detecting genotype-phenotype relationships and mapping in complex, polygenic traits and human diseases [162]. The remarkable success of GWAS is accompanied by the issue of the "missing heritability" [163], namely the fact that the trait-associated SNPs identified through GWAS often account for only a small proportion of the observed correlations in phenotype between relatives. The reason behind this issue has been thought to be that additional genetic factors remain to be found, and that the presence of epistasis is a particular cause for concern [160,164,165]. In reality, if the effect of one locus is altered or masked by effects at another locus, power to detect the first locus is likely to be reduced, and elucidation of the joint effects at the two loci will be hindered by their interaction. Consequently, a large amount of research has been devoted to the detection and investigation of epistatic interactions; a number of methods for detecting the interaction between SNPs have been proposed [116,137,166-168], yet there has been much confusion in the literature over definitions and interpretations of epistasis [169].

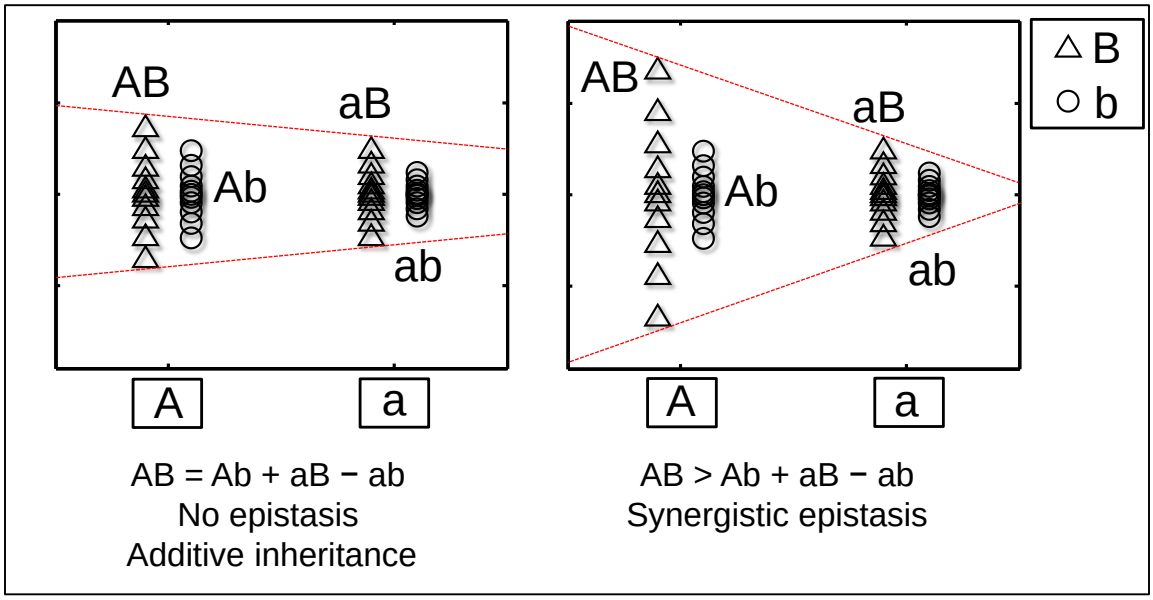


Figure 2.9 Schematic shows that both additive (left) and epistatic (right) effects create similar evQTL signals."A" and "a" are two alleles of evSNP, while "B" and "b" are alleles of interacting SNP.

This study, together with previous findings [154,155], clearly shows that a detailed investigation of local haplotype structure between SNPs at the same locus is necessary to reveal their combined influences on phenotypes of complex traits. For example, we have identified a list of evSNPs that are also associated with human complex trait (see **Table 2.1**). Further investigations on partial associations between closely linked SNPs that may influence these traits should be performed. The same should also be done for FTO whose genotype is associated with phenotypic variability of body mass index [114].

Finally, we point out that an interaction detected via statistical models is different from the biological interaction [169-171]. The lack of direct correspondence between statistical and biological interactions makes it difficult to make strong inferences concerning biological mechanisms based on interaction terms from a statistical model

[164]. Therefore, detection of statistical interaction merely provides a good starting point for a more focused investigation of the joint involvement of the relevant factors, which can perhaps be better addressed through other types of experimental data. Our findings suggest that there is a lot that can be done at the statistical level to prioritize those loci that are most likely to produce significant experimental results.

2.4.5 Conclusions

In conclusion, we used evQTLs as a statistical model system for studying phenotypic variability and dissected the genetic and nongenetic effects by using twin data. Our findings concerning evQTLs offer new insights into the relative contributions of genetic and environmental factors in the formation of evQTLs. Dissecting the genetic components underlying phenotypic variability into additive and epistatic effects allowed the dominant role of additive effect to be revealed.

CHAPTER III

EPISTASIS AND DECANALIZATION SHAPE GENE EXPRESSION VARIABILITY IN HUMANS VIA DISTINCT MODES OF ACTION

3.1 Introduction

Phenotypic variability refers to the likelihood of the phenotypic variation being observed in a population. Quantitative genetics assumes that phenotypic variation, i.e., the difference in phenotypic mean between individuals, is genetically controlled [172]. Under such an assumption, phenotypic variation is explained solely by differences in phenotypic mean among genotypes. This deterministic view, however, has come under challenge. New studies show that phenotypic variance is genetically controlled, and the variance itself is a quantitative trait [68,69,111,112,114,116,127,173-176]. Increasing evidence of genetic control over the variance calls for a paradigm shift in quantitative genetics. Understanding the mechanism of how phenotypic variance is controlled is of great importance for evolutionary biology, agriculture or animal sciences, and medicine [68,69,148,177]. In evolutionary biology, for example, variability offers an adaptive solution to environmental changes [110,178,179]. Genetic factors resulting in more variable phenotypes become favored when they enable a population to respond more effectively to environmental changes [100-103]. In medicine, disease states emerge when the relevant phenotype of affected individuals goes beyond a threshold. As such, high variability genotypes will produce a larger proportion of individuals exceeding that

threshold than will low variability genotypes, even if these genotypes have the same mean. By ignoring the effect of genotypes on phenotypic variance, an important axis of genetic variation contributing to phenotypic differences among individuals has been overlooked [147,172]. The lack of empirical studies in this regard has hindered the discovery of variance-associated mutations that modulate disease susceptibility and the phenotypic variability of other human health-related traits.

Several studies have been conducted to reveal gene expression variability, i.e., the differences in variance of gene expression between groups, in various systems [180-182]. Nevertheless, our understanding of how genetic diversity can control or influence gene expression variability remains limited. Promising new developments along this line have come from our findings in complex trait analysis of gene expression. Using variance-association mapping, we and others identified genetic loci associated with gene expression variance, called evQTLs [68,176] or v-eQTL [175]. How evQTLs effects come about is not completely known. While epistasis has been widely accepted as a mechanism introducing phenotypic variability, here we offer a more straightforward explanation, that is, evQTL variants disrupt or stabilize the genetic architecture that buffers stochastic variation in gene expression. As a result of decanalization, phenotypic expression becomes more sensitive to the external environment and varies more greatly [68,69]. We reveal evQTLs with epistasis and decanalization, two distinct modes of action on gene expression variability and lay the foundation for a new analytical framework that accounts for the genetic contribution to phenotypic variability. We anticipate that methods derived from the new framework will allow us to identify novel

causal loci, which would otherwise be missed by traditional mean-focused methods, in complex disease mapping.

3.2 Materials & methods

3.2.1 Gene expression and genotype data for evQTL analysis

The gene expression data generated by the Geuvadis project RNA-seq study [183] was downloaded from the website of EBI ArrayExpress using accession E-GEUV-1. The downloaded data matrix contained the expression values of Gencode (v12)-annotated genes measured in 462 unique LCL samples. The data were normalized by using the method of probabilistic estimation of expression residuals (PEER)[184]. From the data matrix, we extracted the expression values of autosomal protein-coding genes of 345 EUR samples, whose genotype data is available from the website of the 1,000 Genomes Project [185]. Based on the result of a principal component analysis, we excluded 19 samples whose global expression profile apparently deviated from those of the rest of samples. The final data matrix used for the evQTL analysis contained gene expression values of 15,124 protein-coding genes and 326 EUR samples. Also, we obtained genotype and expression data from a cohort of female twin pairs [117] from the TwinsUK adult twin registry [118]. The data for gene expression in LCLs of 139 pairs of MZ twins were extracted and used in this study.

3.2.2 Identification of evQTLs

Cis-evQTLs were detected using the DGLM method [115] , *trans*-evQTLs were identified using the SVLM procedure [167], these two methods are described in Chapters I and II.

3.2.3 Identification of partial eQTL SNPs that interact with evQTL SNPs

We used a two-step procedure to identify SNPs that interact with evQTLs. We first partitioned individuals into L and S groups according to whether genotypes of the evQTL SNP are associated with large (L) and small (S) variances of gene expression. Then we scanned genome-wide SNPs. For each SNP, the eQTL analysis by linear regression model was conducted among individuals of the L group. For each top SNP with high genotype heterozygosity difference, a linear regression [107] was performed on the SNP genotypes and gene expression. The most significant SNPs were retained after applying an arbitrary *P*-value = 0.0005 as cutoff and were reported as candidate interacting SNPs.

3.2.4 Estimation of gene expression noise using repeated RT-qPCR assay

LCLs were purchased from the Coriell Institute (https://catalog.coriell.org/). The cells were maintained in Roswell Park Memorial Institute Medium 1640 with 2mM L-glutamine and 15% FBS (Seradigm) at 37°C in a humidified atmosphere containing 5% CO_2 (v/v). For the time course experiment, cell lines were seeded at 1×10^6 cells per 10 cm dish and then incubated in culture medium. Cell lines were screened to ensure they were mycoplasma free by using the MycoFluor mycoplasma detection kit (Invitrogen). Cells were collected at 24, 36, 48, 60, and 72 h after growth. Total RNA was extracted using Trizol reagent (Invitrogen). RNase-free DNase (Ambion) was used to remove potential contaminating DNA from RNA samples. RNA purity and concentration were determined using Nanodrop ND-100 Spectrophotometer. The concentrations of total RNA were adjusted to 100 μg/ml. Real-time RT-PCR assays were performed using iTaq

Universal SYBR Green One-Step Kit (Bio-Rad Laboratories) with primers shown in **Table 3.1**. Template total RNA was reverse transcribed and amplified in a Bio-Rad CFX96 Real-Time PCR Detection System (Bio-Rad Laboratories) in 20-µl reaction mixtures containing 10 µl of iTaq universal SYBR Green reaction mix (2×), 0.25 µl of iScript reverse transcriptase, 2 µl of 100 nM forward and reverse primer mix, 1 µl of total RNA template, and 6.75 µl of nuclease-free water, at 50°C for 10 min, 95°C for 1 min, followed by 30 cycles of 95°C for 10 s and 58°C for 30 s. Melting curves were measured from 65°C to 95°C with 0.5°C of increment. The average expression of two housekeeping genes (*CHMP2A* and *C1orf43*) was used for normalization. The choice of using these two genes as reference was based on recent scrutiny of human genes with a constant level of expression using RNA-seq data [186].

Table 3.1 List of primers used for RT-qPCR.

evQTL Gene Targets	Primers
ATMIN	5' AATGCCCTTGTCAGTAGGAAC 3'
	5' GGCTCACCAGCAATAGGATTAG 3'
BEND4	5' GTCGAATGATCTTGGATGCCTT 3'
	5' TCCAGGAGTTTTCCTCCACAAT 3'
CHMP2A	5' CGCGAGCGACAGAAACTAGAG 3'
	5' CCCGCATCAATACAAACTTGC 3'
ZNF10	5' TCAGGACAGTTGTGCAAGTAAC 3'
	5' GGGTTTCTCTCTATGTATGCCCT 3'

3.2.5 Flow cytometric analysis of cells in different phases of the cell cycle

Cell cycle distribution was evaluated by using flow cytometry. This determination was based on the measurement of the DNA content of nuclei labeled with propidium iodide [187]. Cells were harvested at 24, 36, 48, 60, and 72 h after treatment. The cells were resuspended at a concentration of 1×10^6/ml in cold PBS. After 1ml of ice-cold 100% ethanol had been added dropwise, the cells were fixed at 4°C for at least 16 hours. The fixed cells were pelleted, resuspended in 1ml of propidium iodide (PI) staining solution (50 mg/ml propidium iodide, 100 units/ml RNase A in PBS) for at least 1 hour at room temperature and analyzed on an FACS flow cytometer (BD). By using red propidium-DNA fluorescence, 30,000 events were acquired. The percentage of cells in G0/G1, S and G2/M phases of the cell cycle was calculated using Flowjo software v10 (Tree Star).

3.3 Results

3.3.1 Widespread evQTLs in the human genome

We obtained the expression data for 15,124 protein-coding genes measured in 462 LCLs by the Geuvadis Project [183]. We also obtained genotype data at 2,885,326 polymorphic sites determined in the same cell lines by the 1,000 Genomes Project [185]. After data processing, 326 LCL samples from unrelated individuals of EUR were retained for this study (Materials and Methods). To identify evQTLs, we first applied a method based on the DGLM [115]. The method has been previously adopted by us [68,176] and others [69]. Owing to computational complexity, we restricted the use of this method to the identification of *cis*-acting evQTLs. On average ~1800 SNPs that lay

within 1-Mb radii of the transcription start site were tested per gene. Using a conservative Bonferroni correction cutoff $P = 1.75\times10^{-9}$ (= 0.05 / 28,494,473), we identified a total of 17,949 *cis*-evQTLs in 1,304 unique genes, i.e., 8.6% of all genes tested (**Figure 3.1A**). Next, to identify both *cis*- and *trans*-evQTLs genome-wide, we adopted the method based on SVLM [167,175]. It is a computationally efficient, two-stage method. The effect of variants on gene expression mean (i.e., eQLT effect) is firstly removed by regression, and the residuals are squared to give a measure of expression dispersion. Then the correlation between squared residuals and genotypes is tested. We applied SVLM to test all SNPs against all genes, without pre-filtering SNPs by their locational relationship with tested genes. Such an all-against-all strategy allowed a systematic survey of *cis*- and *trans*-evQTLs across the entire genome. We used the Benjamini-Hochberg procedure [188] to determine the *P*-value cutoff of 3×10^{-9} that gave the FDR of 0.1. At this level, we identified 505 *cis*-evQTLs in 33 unique genes, and 1,008 *trans*-evQTLs in 235 unique genes (**Figure 3.1B**). Two genes *AXIN2* and *FAM86B1* were found to have both *cis*- and *trans*-evQTLs. Applying the same FDR cutoff to detect both *cis*- and *trans*-evQTL resulted in an unbiased picture of the distribution of all evQTLs across autosomes (**Figure 3.1C**). Comparing the positions of genes and their evQTLs, we did not observe a strong enrichment of data points along the diagonal of the graph, suggesting *cis*-evQTLs not be particularly enriched compared to *trans*-evQTLs. We noticed a pronounced discrepancy in the number of *cis*-evQTLs detected using DGLM and SVLM. This discrepancy may because that SVLM and DGLM have different detecting power. Computer simulations showed that, when the

sample size was set to 300, SVLM had only half of the power of DGLM (**Fig 3.2**). Furthermore, the huge multiple testing burden associated with the application of SVLM in the all-against-all tests may also contribute to the discrepancy.

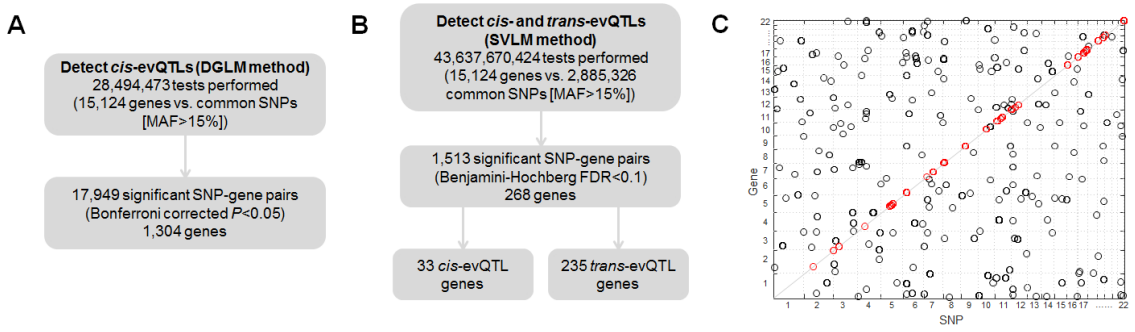


Figure 3.1 Overview of evQTL detections and the distribution of *cis*- and trans-eQTLs in autosomes. (A) Flowchart of *cis*-evQTLs identification using DGLM method. (B) Flowchart of *cis*- and *trans*-evQTL identification using SVLM method. (C) Distribution of SVLM-identified *cis*- and *trans*-evQTLs in autosomes, in which *cis*-evQTLs marked in red and t*rans*-evQTLs marked in black.

3.3.2 Epistatic interactions contribute to increasing gene expression variability

Epistasis, i.e., the interaction between loci, may increase the phenotypic variability of a population [116,137]. The evQTLs provided source materials for studying epistatic effects on gene expression variability [176]. More specifically, we sought to identify "third-party" SNPs that interact with evQTL SNPs. Such interactions result in more variable gene expression of the evQTL genes. In particular, for each evQTL SNP identified by using SVLM, we applied a two-step procedure to identify the third-party SNPs, also known as *partial eQTL SNPs* (see below). These third-party SNPs interact or are partially associated with evQTL SNPs, resulting in the increased gene expression variance [175,176]. The process of partial eQTL SNP identification is illustrated in **Fig 3.3**. Briefly, for a given evQTL (for example, the evQTL between gene

X and SNP *Y*), we extracted samples with a homozygous genotype associated with large expression variance. We called these L group samples. Accordingly, those related to small expression variance was called S group samples. Then, we conducted a genome-wide scan among the extracted L group samples to identify eQTL SNPs (e.g., SNP *Z*) that control the expression of the corresponding evQTL gene (i.e., gene *X*). The identified eQTL SNPs are called *partial* because they are detected in the sub-sampled discovery panel, and their effect on gene expression is restricted to L group samples. The evQTL SNP *Y* and its partial eQTL SNP *Z* may be co-localized proximately on the same chromosome and partially associated as we showed previously [176]. They may also be unlinked, for instance, located on different chromosomes, and interact with each other epistatically [175]. Here, we focused on the 268 evQTLs (33 *cis*- and 235 *trans*-acting ones) identified by using SVLM. In 73 out of 268 evQTL genes, we identified at least one significant interacting SNP, i.e., partial eQTL SNP with simple linear regression test $P < 10^{-8}$ in the L group samples. These results suggest that more than one-fourth of evQTLs are attributable to partial eQTL SNPs interacting with evQTL SNPs.

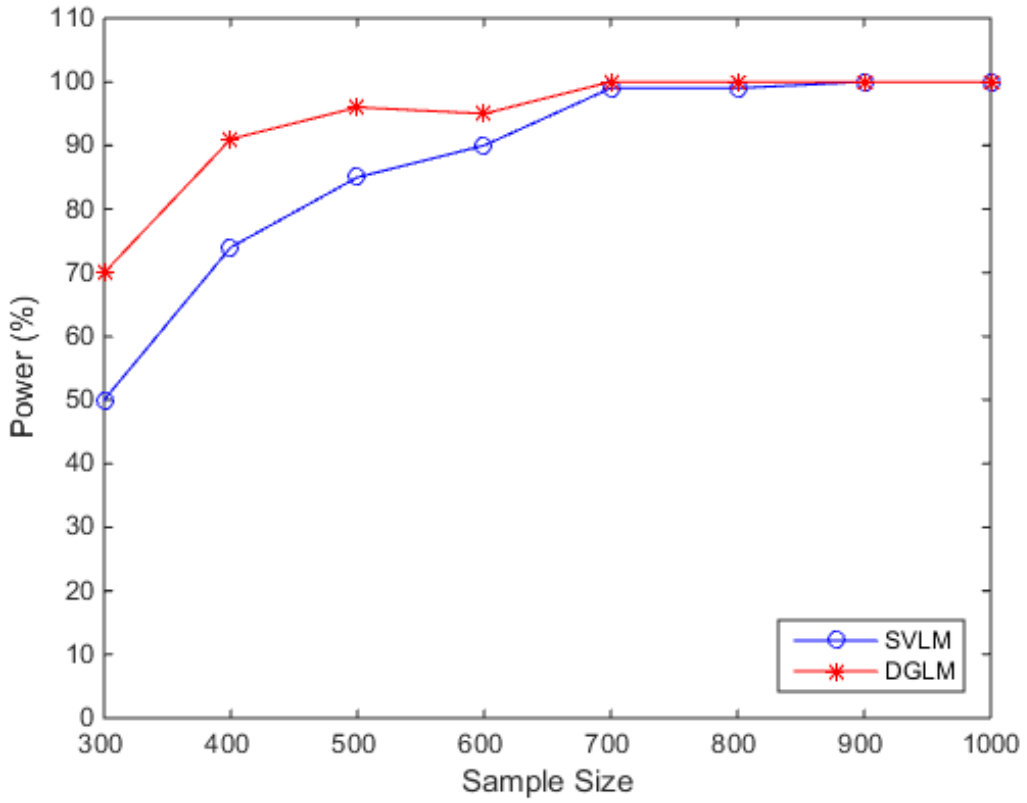


Figure 3.2 Comparison of statistical power of two evQTL detection methods: DGLM and SVLM, using computer simulations with different sample sizes. For simulations, a population of 10,000 individuals was generated, and the MAF of an evQTL SNP was set to 0.4. The genotypes of SNP were encoded to 0, 1, 2 for homozygous minor, heterozygous, and homozygous major alleles, respectively. The gene expression of each genotype was generated from a normal distribution with the same mean but different variances, 1.0, 2.0, and 4.0, respectively. Before testing a method, the population was subsampled to the designated sample size, ranging from 300 to 1,000. For each sample size, the tested method was applied to the subsamples. The whole procedure was repeated 1,000 times, and the power was computed as the ratio of the times of P-value being smaller than 5×10^{-5} (i.e., 0.05/1000).

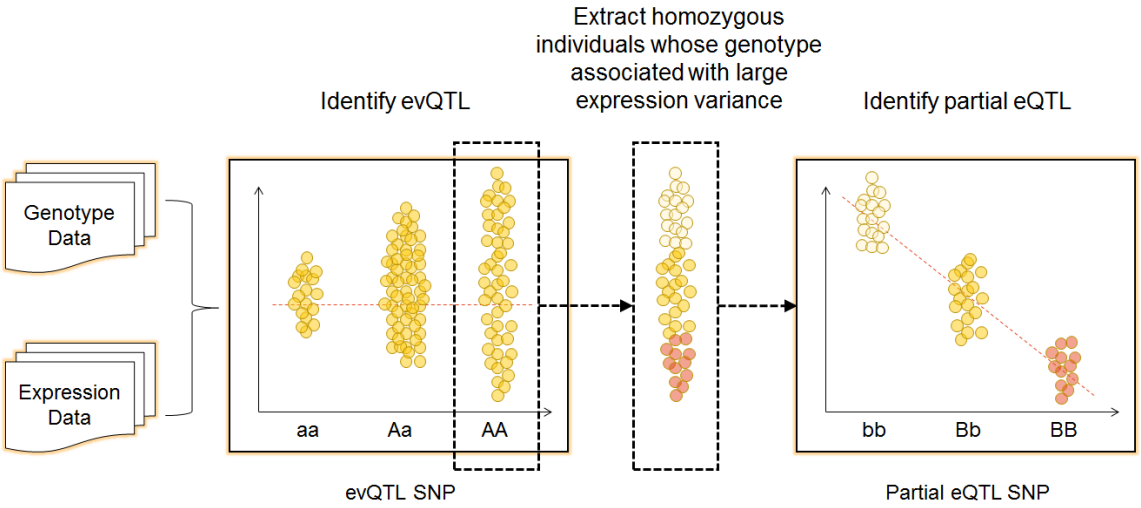


Figure 3.3 Schematic illustration of the method for identifying partial eQTLs. After the identification of evQTL, the partial eQTL method involves two steps: (1) extraction of homozygous individuals whose genotype of the evQTL variant is associated with increased expression variability, and (2) identification of the eQTL between the gene and third-party variant among extracted individuals.

3.3.3 Decanalization contributes to increasing gene expression variability without genetic interactions

Here we put forward the decanalization model to explain the formation of evQTLs. The model emphasizes the interaction between gene (or genotype) and the environment. Unlike the epistasis model that concerns the epistatic interactions or associations between variants at different loci [175,176], the decanalization model concerns a single variant that perturbs stable genetic systems through a decanalizing effect on the expression of the specific genotype. We hypothesized that some evQTL SNPs are associated with gene expression variability because one of their two alleles confers the decanalization function, causing more variable gene expression. In other words, decanalizing SNPs increase gene expression variability via the single-locus effect,

without interacting with any other SNPs. Thus, these decanalizing evQTLs have a different formation mechanism in contrast to that of epistatic evQTLs.

To show the decanalizing effect, by further controlling the diversity of samples' genetic backgrounds, we re-visited the genotype and expression data from our previous study [176]. The data were derived from LCLs of a cohort of twin pairs [117]. In the previous study, we used a single set of the twin pairs, which contains one individual from each twin pair for evQTL analysis and identified *cis*-evQTLs in 99 unique genes [176]. Here, we first classified the 99 evQTLs (between each gene and the most significant SNP) into 56 epistatic and 43 decanalizing evQTLs. The classification was based on whether or not an interacting SNP (i.e., partial eQTL SNP) could be identified using the two-step procedure described above. The idea was that if no interacting SNP can be detected for an evQTL, then the evQTL cannot be explained by the epistasis model. Thus, the evQTL is likely to be a decanalizing evQTL, explained by the decanalization model, in which increased gene expression variability is driven by the allele of evQTL SNPs with decanalizing function. Next, we extracted expression data of the 139 pairs of MZ twins. We classified MZ twin pairs whose genotypes were homozygous at evQTL SNP sites into MZ-L or MZ-S groups, according to whether their evQTL SNPs were associated with large or small variance. For all MZ twin pairs in the same group (either MZ-L or MZ-S), we quantified discordant gene expression between two individuals of the same pairs. Discordant gene expression was calculated as the RMD in gene expression, which is the difference between two individual's gene expression values normalized by the mean (Materials and Methods).

To illustrate the difference in discordant gene expression between groups, we use two example evQTLs. One is a decanalizing evQTL between *TBKBP1* and rs1912483 (**Fig 3.4A**, right), and the other is an epistatic evQTL between *PTER* and rs7913889 (**Fig 3.4B**, right). The data points of gene expression levels were grouped by the genotype. Within each genotype category, data points from the same twin pairs are displayed side-by-side. Every two individuals of the same MZ pairs are linked by a line. The slope of the lines is an indicator of discordant gene expression between twin pairs. In the decanalizing evQTL example, the slopes between MZ twins with genotypes associated with large expression variance (i.e., MZ-L group) tend to be steeper than those with small expression variance (i.e., MZ-S group)(**Fig 3.4 A**, left). In contrast, in the epistatic evQTL example, the difference in slope skewness between MZ-L and MZ-S groups is less pronounced (**Fig 3.4 B**, left). We pooled RMD values from different twin pairs together by MZ-L or MZ-S group and compared the distributions of RMD values between the two groups. For decanalizing evQTLs, the distributions of RMD values between L and S groups were significantly different ($P = 1.3\times10^{-5}$, **Fig 3.4 A**, right), with larger RMD values for L group. In contrast, for epistatic evQTLs, this difference in RMD distribution was not detected between L and S groups ($P = 0.052$, **Fig 3.4 B**, right).

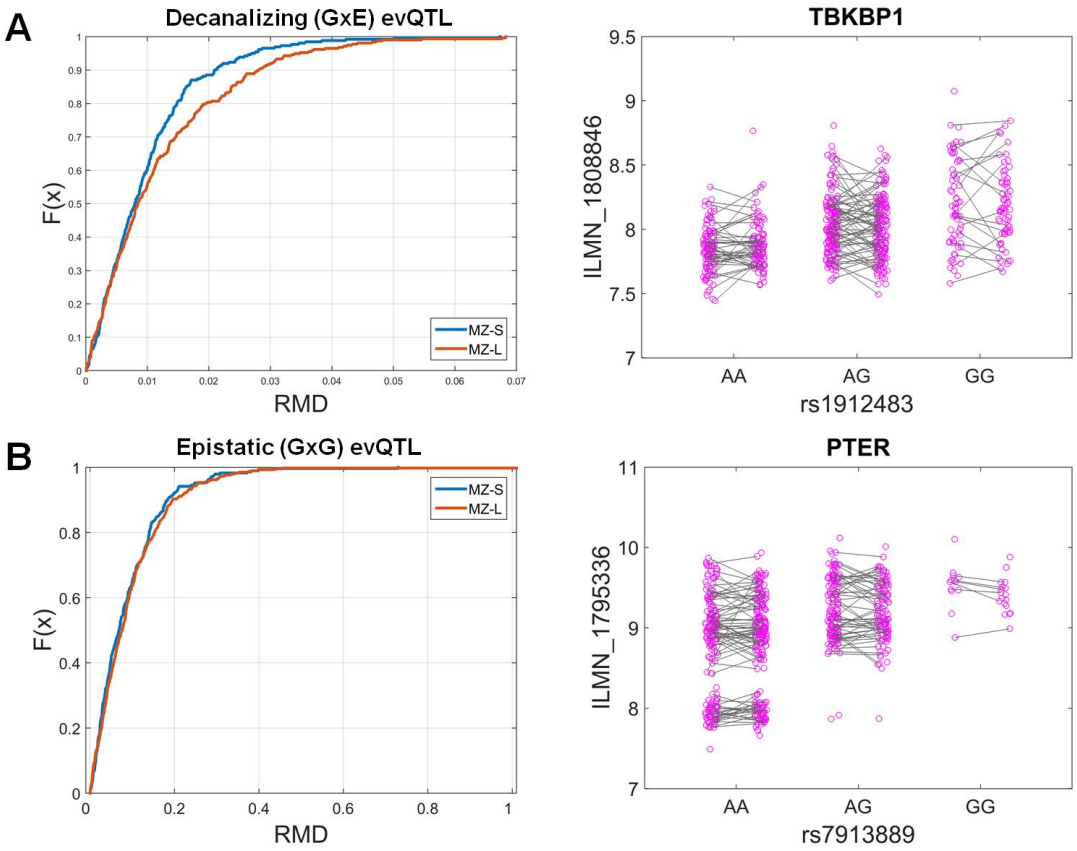


Figure 3.4 Dissection of decanalizing and epistatic effects of evQTLs using twin data.(A) An example of decanalizing evQTL, TBKBP1-rs1912483. The expression data points for each of two individuals from the same pairs of MZ twins are linked. Twin pairs are grouped as MZ-L and MZ-S based on whether the homozygous genotype at rs1912483 is associated with large or small gene expression variance. The right panel shows the CDF of normalized discordant gene expression (measured using RMD) for MZ-S and MZ-L groups. (B) Same as (A) but showing an example of epistatic evQTL, PTER-rs7913889.

3.3.4 Decanalizing evQTL SNPs are associated with gene expression noise

Our decanalization model works by the action of a single genetic variant conferring a decanalizing effect on gene expression. One of the underlying sources of gene expression variability is stochastic noise in gene expression [189]. We hypothesized that different alleles of a decanalizing evQTL SNP might be associated with different levels of expression noise of the corresponding evQTL gene. To test this hypothesis, we set out to estimate the expression noise using RT-qPCR by repeatedly

measuring gene expression level in the same cell line. If our hypothesis is correct, then the expression variance of cells from an individual with an evQTL genotype associated with larger variance should be more pronounced than the expression variance in cells from an individual with genotype with smaller variance.

We selected two decanalizing evQTLs: *ATMIN*-rs1018804 and *BEND4*-rs7659929, for testing. *ATMIN* is an essential cofactor for checkpoint kinase ATM, which *trans*duces genomic stress signals to halt cell cycle progression and promote DNA repair [190]. We picked two LCLs, HG00097 and HG00364, which have the similar *ATMIN* expression levels. Both were derived from female individuals of European descent. The difference is that HG00097's genotype CC at rs1018804 is associated with larger variance, while HG00364's genotype AA at rs1018804 is associated with smaller variance. Thus, HG00097 and HG00364 belonged to L- and S-groups, respectively. We measured the evQTL gene expression level using RT-qPCR with three technical replicates each at five different sampling time points. The same assay was repeated three times independently. Our results showed that under the same controlled experimental condition, the variance of gene expression (i.e., the variance in ΔCt values) in HG00097 was greater than HG00364. The same trend was observed from all three biological replicates (**Fig 3.5 A**). In all three replicates, the difference was statistically significant (Brown-Forsythe test, $P = 0.034$, 0.019, and 0.0096, respectively).

We repeated the experiment with two biological replicates on the same evQTL *ATMIN*-rs1018804 using a different pair of LCLs (NA12144 and NA12736 from L- and S-group, respectively) to replace HG00097 and HG00364. We obtained the similar

results showing a consistent pattern, that is, the gene expression in the cell line of L-group is more variable than that of S-group (**Fig 3.5A**). Furthermore, we repeated the experiment on a different decanalizing evQTL (*BEND4*-rs7659929) with another pair of LCLs (NA12889 and NA18858). Again, we obtained the consistent pattern that supports the correlation between gene expression variability and stochastic noise (**Fig 3.6A**).

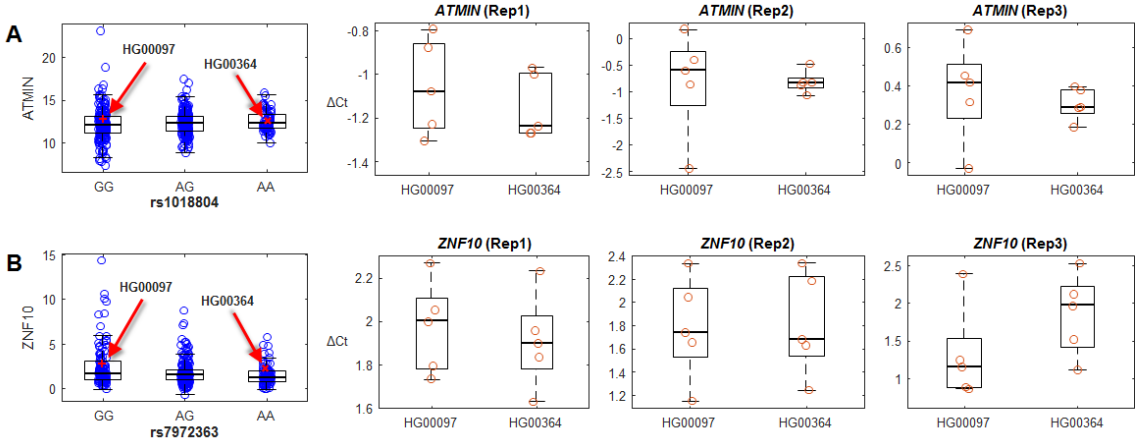


Figure 3.5 The correlation between gene expression variability and noise in the decanalizing evQTL, *ATMIN*-rs1018804, but not in the epistatic evQTL, *ZNF10*-rs7972363, in the same cell line pair (HG00097 and HG00364).(A) The leftmost panel shows the distribution of gene expression levels of *ATMIN* among three different genotypes defined by two alleles of rs1018804. Red arrows indicate the genotype and expression level of HG00097 and HG00364. Right panels show the results of three biological replicates of repeated RT-qPCR analysis for *ATMIN* at five different time points ranging from 12 to 60 h after incubation. At each time point of each biological replicate, three technical replicates were performed to obtain ΔCt values. Red circles indicate the average ΔCt values. The difference in variance of ΔCt between two cell lines was tested using F-test for equal variances (P = 0.429, 0.012, and 0.049, respectively, for the three replicates). (B) Same as (A) but showing the results of evQTL *ZNF10*-rs7972363. P values of F-test for the three replicates are 0.981, 0.195, and 0.066, respectively.

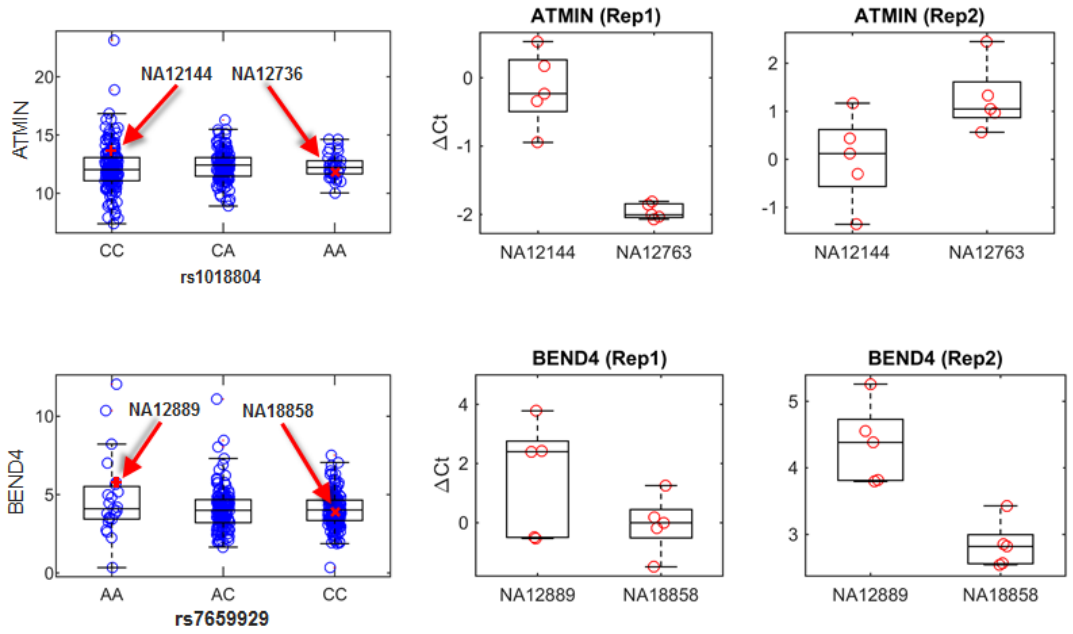


Figure 3.6 The correlation between gene expression variability and noise in the GxE evQTLs, *ATMIN*-rs1018804 in the cell line pair NA12144 and NA12763, BEND4-rs7659929 in the cell line pair NA12889 and NA18858.(A) The leftmost panel shows the distribution of gene expression levels of *ATMIN* among three different genotypes defined by two alleles of rs1018804. The genotype and expression level of NA12144 and NA12736 are indicated by red arrows. Right panels show the results of two biological replicates of repeated RT-qPCR analysis for *ATMIN* at five different time points ranging from 12 to 60 h after incubation. At each time point of each biological replicate, three technical replicates are performed to obtain ΔCt values, and the average is presented by the red circle. The difference in variance of ΔCt between two cell lines was tested using F-test for equal variances (P = 0.037 and 0.6148, respectively, for the two replicates). (B) Same as (A) but showing the results of evQTL *BEND4*-rs7659929 using cell line pair NA12889 and NA18858. P-values of F-test for the two replicates are 0.217 and 0.334, respectively.

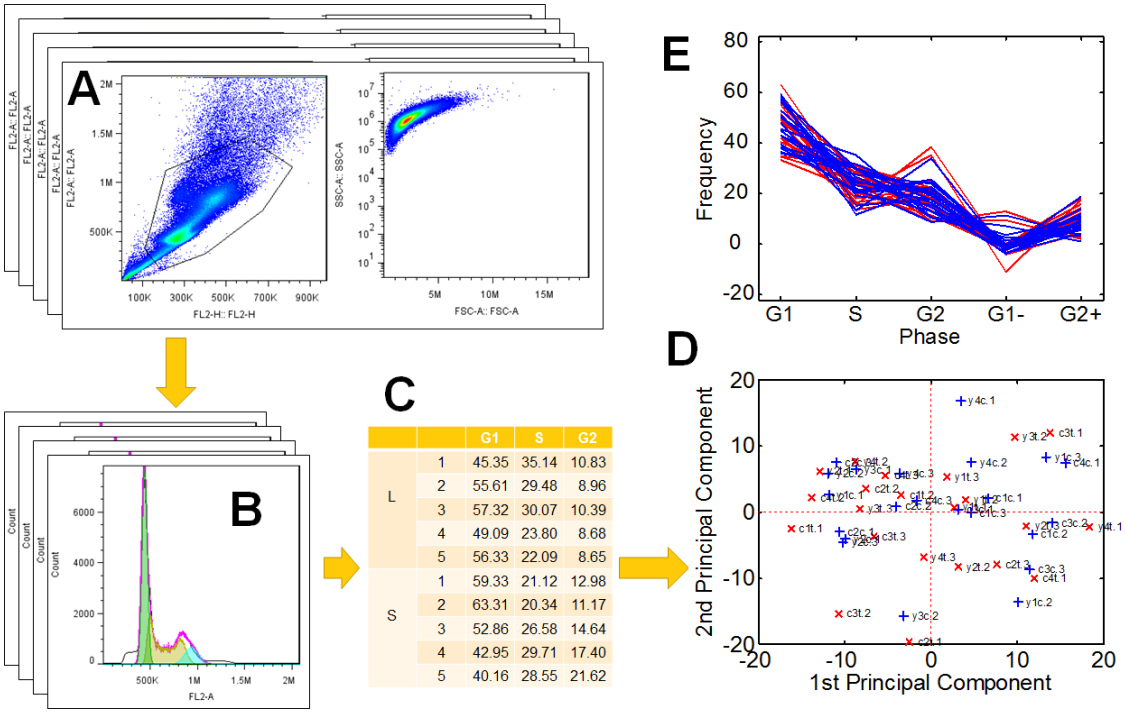
We hypothesized that the correlation between gene expression variability and noise exists exclusively in decanalizing evQTLs. We did not expect such a correlation would be recapitulated in epistatic evQTLs. This is because the two kinds of evQTLs work through different modes of action. To test this, we repeated the same RT-qPCR experiment with an epistatic evQTL *ZNF10*-rs7972363 using the same cell lines HG00097 and HG00364 (**Fig 3.5B**). The genotype AA of HG00097 at rs7972363 is associated with larger variance while the genotype GG of HG00364 is associated with smaller variance. As an epistatic evQTL, the interacting SNP rs1567910, which interacts

with rs7972363 and helps the creation of the evQTL, has been identified by using the two-step partial eQTL detection method. Samples with AA genotype at rs7972363 can be further broken down by rs1567910 into three subgenotype groups associated with different levels of gene expression mean. Consistent with our expectation, the gene expression variance in ΔCt values was similar between HG00097 and HG00364 (**Fig 3.5B,** Brown-Forsythe test, P = 0.96, 0.83, and 0.73, for the three replicates, respectively). Together, our results suggest that the level of gene expression noise—the random fluctuation of gene expression—is associated with decanalizing evQTL, but not epistatic, SNPs.

3.3.5 Differences in cell cycle status and alternative splicing do not account for the decanalizing function conferred by decanalizing evQTL SNPs

Finally, we controlled for two additional confounding factors that might account for the increased gene expression variability associated with evQTLs. The first is the cell cycle status of cell lines. At the same sampling time, cell lines may differ in the percentage or number of cells in different cell cycle phases. Could the difference in cell cycle status explain the difference in gene expression variability or noise between cell lines? To test this, we performed the cell cycle analysis by flow cytometry with HG00097 and HG00364 at 36 h after incubation (Materials and Methods). The results showed no difference in the percentage of cells in G0/G1, S and G2/M phases between the two cell lines (**Fig 3.7**). The second confounding factor we considered is the alternative splicing pattern. Different splicing patterns between cell lines might result in different gene-level expression measurements. We used the Integrative Genomics

Viewer [191] to visualize the alternatively spliced mRNA of *ATMIN* and compared the pattern of splicing between HG00097 and HG00364, as well as that of *BEND4* between NA12889 and NA18858. In either case, we observed no difference in splicing patterns (**Fig 3.8**).



		G1	S	G2
L	1	45.35	35.14	10.83
	2	55.61	29.48	8.96
	3	57.32	30.07	10.39
	4	49.09	23.80	8.68
	5	56.33	22.09	8.65
S	1	59.33	21.12	12.98
	2	63.31	20.34	11.17
	3	52.86	26.58	14.64
	4	42.95	29.71	17.40
	5	40.16	28.55	21.62

Figure 3.7 Cell cycle analysis to determine the relative abundance of cells in different phases. (A) Representative flow cytometric dot plots. (B) Representative histograms obtained using TUNEL assay. (C) Relative frequencies of cells in G1, S, and G2 phases. (D) Principal component analysis (PCA) of cell cycle profiles. (E) Relative frequencies of cells in different phases of HG00097 (red) and HG00364 (blue).

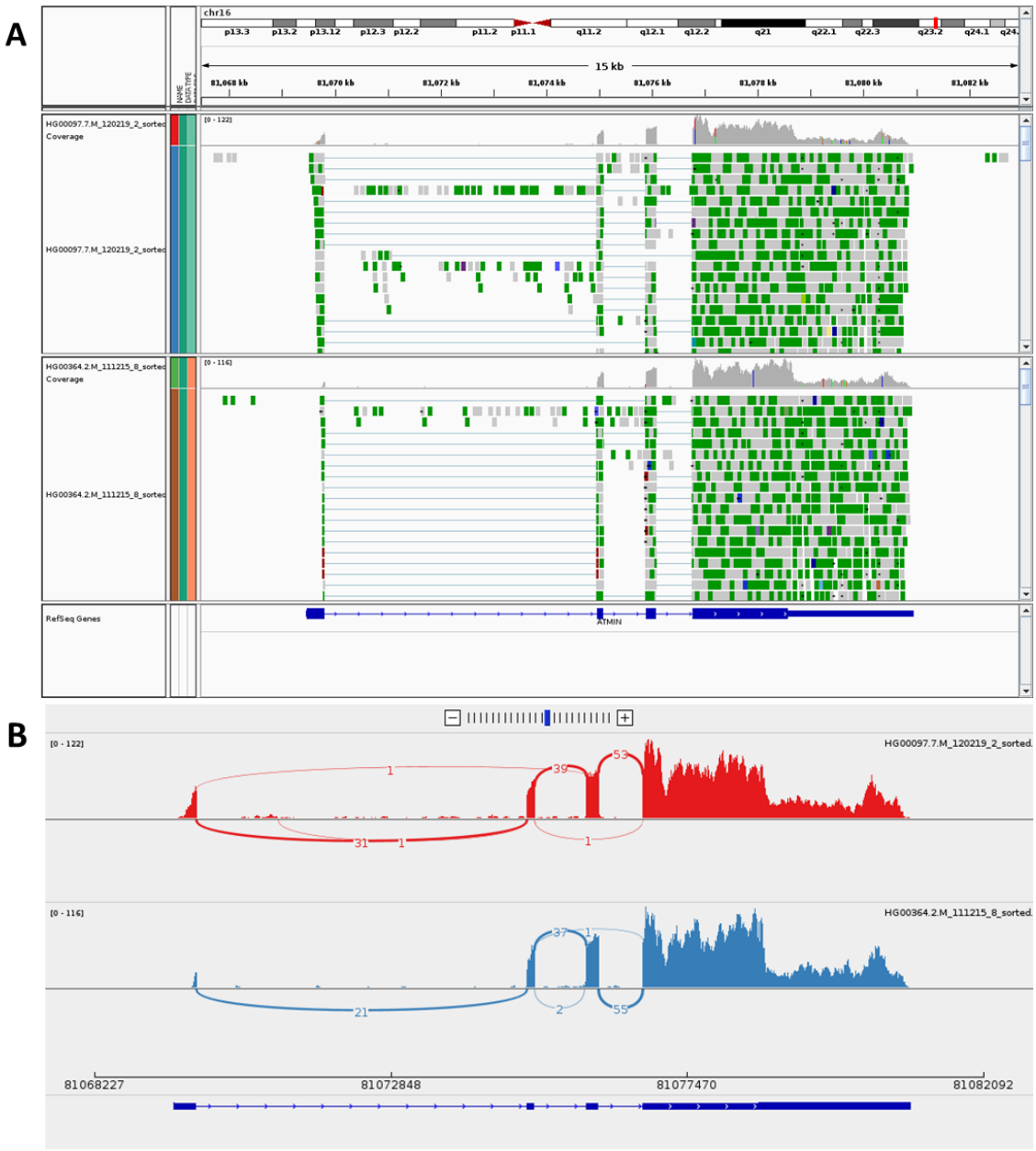
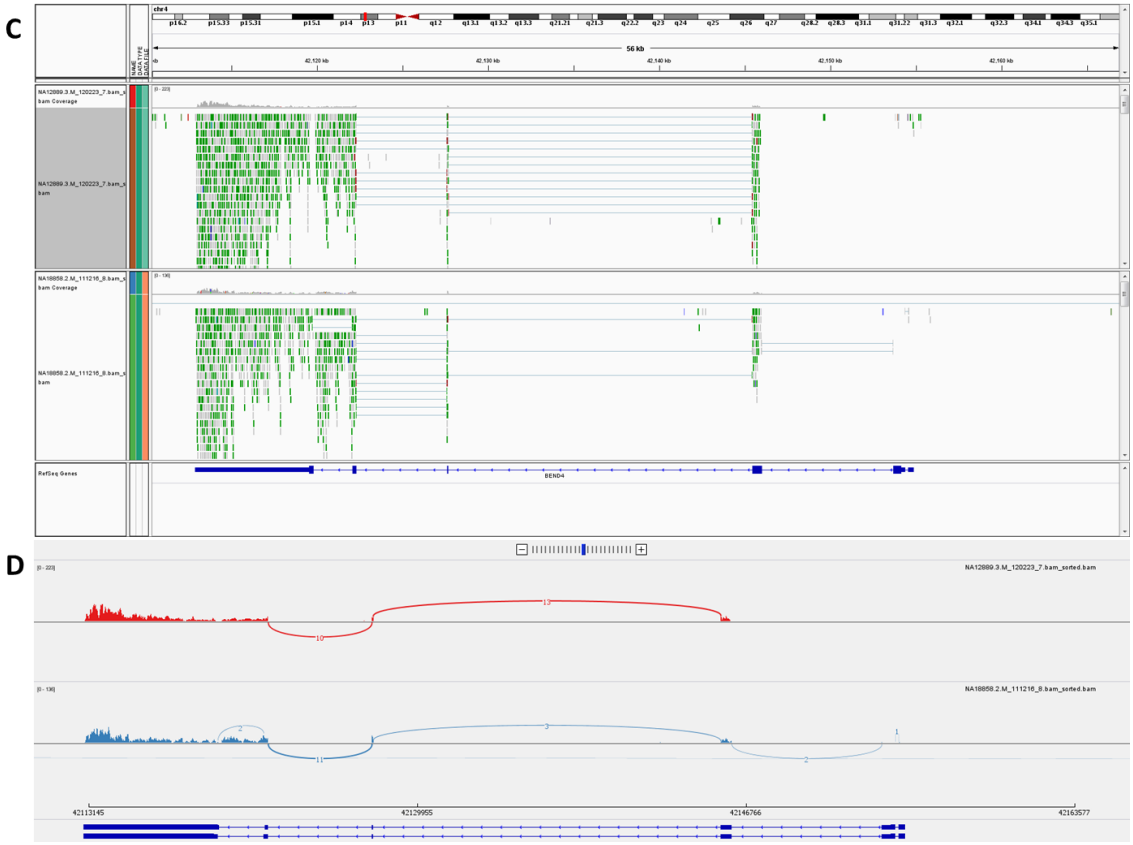


Figure 3.8 IGV view of RNA-seq read alignments and sashimi plot of mRNA splicing patterns of evQTL genes in different cell lines. (A) IGV view of RNA-seq read alignment of *ATMIN* in HG00097 and HG00364. (B) Sashimi plots of *ATMIN* mRNA in HG00097 and HG00364. (C) IGV view of RNA-seq read alignment of *BEND4* in NA12889 and NA18858. (D) Sashimi plots of *BEND4* mRNA in NA12889 and NA18858.

Figure 3.8 continued



3.4 Discussion

Variability, which refers to the potential of a population to vary, is a central concept in biology [94] Emerging experimental and statistical techniques have allowed the variability of various phenotypes to be rigorously analyzed [177]. Focusing on variability QTLs of gene expression, we found that evQTLs are abundant and widespread across the human genome [68,176]. In the light of evQTLs, the present study reveals two distinct modes of action: epistasis and decanalization, through which common genetic variation controls or influences gene expression variability. The

epistasis model concerns two or more variants (at separate loci?), which interact in a non-additive fashion [175,192] or link to each other through incomplete LD [176,193]. Consistent with this model, a number of methods for identifying epistasis have been proposed, based on detecting increased phenotypic variability [116,137,166]. The decanalization model is simpler and more direct, concerning single variants that work alone to destabilize phenotypic expression and pushing a proportion of individuals away from the robust optimum.

Dissecting the GxG and GxE effects, respectively underlying the epistatic and decanalizing modes of action, in the context of variability QTLs is technically challenging. Here we have taken advantage of the identical genetic background of MZ twins and showed that different genotypes are associated with varying degrees of responsiveness to environmental perturbation. We also detected an unexpected link between population-level gene expression variability and stochastic gene expression noise as measured in single individuals. This suggests that variable gene expression in each sample may be synthesized and aggregated together and eventually contribute to the gene expression variability of the population as a whole. In other words, the same underlying force destabilizing gene expression might be proposed as a unified explanation for gene expression variability at different scales (i.e., from the population level to the individual level). At the other end of the spectrum, we werte able to examine cell-to-cell variability in gene expression, thanks to the rapid development of single-cell based technologies [194]. For example, the genetic control of variability in burst size and

frequency of single-cell *trans*cription may not be too different from that of the population and individual levels.

We were unable to describe the precise mechanism of a decanalizing function conferred by evQTL variants. However, we were able to utilize bioinformatics analysis to provide a rationale to substantiate the link between the variants, the possible genetic mechanisms, and the phenotype, i.e., expression variability of the corresponding gene. By synthesizing different sources of information, we attempted to build working models for evQTLs explaining how evQTL variants can influence gene expression variability. Here we use GxE evQTL *ATMIN*-rs1018804 as an example to illustrate one of the tentative models. The intronic rs1018804, which lies 43-bp downstream from the nearest exon-intron boundary, may play a role in regulating the splicing of *WDR24* mRNA. *WDR24*,encodes WD repeat-containing protein 24, a key component of Rag-interacting complex essential for the activation of mTORC1 [195]. The WDR24 protein is predicted [196] to interact with Hsp70 and DNAJ proteins [197]. The latter two interact with the dynein light chain DYNLL1 [198]. Finally, *DYNLL1* and *ATMIN* form a feedback regulation loop—one of few known examples of negative auto-regulation of gene expression where a gene product directly inhibits the main *trans*criptional activator while bound at its own promoter [199]. Taken together, the working model can be represented as rs1018804 → WDR24 → Hsp70/DNAJ → *DYNLL1* ↔ *ATMIN*. This working model offers a workable blueprint for functional dissection of all components involved. This information flow model provides potential insights into the mechanisms of evQTL variants influencing gene expression variability.

We anticipate that our results have implications for studying human diseases in which regulatory variation plays critical roles [200]. Decanalization effect has been proposed to influence brain development and contribute to the risk of psychological disorders like schizophrenia [201,202] and other complex diseases [148]. Increased gene expression variability was found to be associated with the aging in a mouse model [98], and the aggressiveness of lymphocytic leukemia [203]. Understanding how genetic variation contributes to increasing gene expression variability or variability of other phenotypic traits will facilitate the identification of causal variants. This is especially true when gene expression heterogeneity characterizes the disease under consideration. Indeed, many human diseases are characterized by etiological and phenotypic heterogeneity, echoing the so-called "Anna Karenina principle," that is, each sick person is sick in his or her own way. If even a small fraction of the increased phenotypic variability among patients is due to the variability-controlling mutations (such as evQTL variants), understanding how these mutations influence the variability is of importance. Better understanding of variation control may bring us closer to causal mutations underlying an individual's predisposition to disease. This strategy, if combined with other methods for estimating the impact of rare mutations, such as aberrant gene expression analysis for private mutations [204], would be provide increased power for personalized medicine. Furthermore, we suggest that variability-controlling mutations are potential targets for genomic editing or drug development. Drug targeting these mutations might bring the dysregulated and dysfunctional gene expression in patients back to normal levels of gene exprssion and health.

CHAPTER IV

ABERRANT GENE EXPRESSION IN HUMANS[†]

4.1 Introduction

The advent of high-throughput sequencing and genotyping technologies enables a comprehensive characterization of the genomic and *trans*criptomic landscapes of each individual. Deciphering the massive body of data associated with samples from many individuals presents a major challenge [205,206]. Over the last few years, eQTL analyses have provided in-depth insights into the effect of genetic variation on the regulation of gene expression [105,108,183,207]. More recently, research has also focused on the contribution of genetic variation to variance of gene expression [68,175,176].

The analytical frameworks adopted by most eQTL studies have historically been based on population-level test statistics, which are powerful for establishing associations between commonly occurring genetic variations and gene expression. However, few frameworks or statistics are available for assessing the impacts of rare genetic variants on gene expression (except, for example, [93]). The problem is exacerbated by the fact that individual gene expression is a function of both genetic and nongenetic (such as epigenetic and environmental) factors, as well as their combined action. Our failure to

[†] This chapter has been reprinted from: Zeng Y*, Wang G*, Yang E, Ji G, Brinkmeyer-Langford CL, Cai JJ (2015) Aberrant gene expression in humans. PLoS Genetics, 11(1):e1004942. doi:10.1371/journal.pgen.1004942, with permission from PLoS Genetics. It is available online at http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004942

detect the effects of rare variants with large effects in biological samples, along with the inherent difficulty in dissecting the complex factors influencing gene expression, will hinder efforts to define and prioritize relevant variants and impede the development of improved personalized diagnostic and therapeutic options.

Here, we envision an alternative approach based on the theory of multivariate outliers to address these technical challenges. More specifically, we measure how any two individuals differ in their expression profiles and quantify these differences with respect to a set of genes between individuals. Based on the expression differences, we detect outlier individuals whose expression profiles are so divergent from those of others in the population that the divergence cannot be explained by random sampling variation alone. Many methods of outlier detection have been developed. The most commonly used of these methods, such as those based on the estimation of the location and scatter of the data points or the quantiles of the data, are more applicable to univariate than multivariate settings. In practice, however, phenotypic traits are associated with changes in multiple genes in biological pathways and molecular networks, more often than single gene alterations. Reliably identifying outliers in such a multivariate setting is a challenging problem—unlike the simpler case of univariate outlier detection, simple graphical diagnostic tools like the boxplot often lack statistical power when the analysis of more than one dimension is attempted [208].

To this end, we adapted a multivariate outlier method that allows simultaneous evaluation of expression data with respect to many dimensions derived from multiple genes. With this method, even though there is no natural ordering of multivariate data on

which “extremeness” of an observation can be ascertained, outliers showing markedly different data profiles can be detected. Using a framework based on this approach, we specifically address the following research questions: Are there any differences between the functional properties of genes tending to (or tending not to) be aberrantly expressed? Is aberrant expression population-specific? What are the roles of genetic and nongenetic factors in aberrant expression? Do common or rare genetic variants contribute to aberrant expression? Our overall results clearly demonstrate that outliers, while often considered as error or noise, do carry important biologically-relevant information. Thus, the careful characterization of the genetic bases underlying the markedly different expression profiles of outlier samples is both worthwhile and necessary. Accurate description of inter-individual expression differences requires the incorporation of the effects of both common and rare regulatory genetic variants.

4.2 Materials & methods

4.2.1 Geuvadis RNA-seq data

We used gene expression data produced by the Geuvadis project RNA-seq study [183] as described in chapter III. We excluded individuals whose genotype information was unavailable in the 1000 Genome Project Phase 1, resulting in a total of 402 remaining samples (326 EUR and 76 AFR).

4.2.2 Annotated gene sets

Gene sets were downloaded from Molecular Signatures Database (MSigDB) v4.0 [209]. The MSigDB gene sets had been divided into seven groups: C1 - positional gene sets (n = 326), C2 - curated gene set (n = 4,722), C3 - motif gene sets (n = 836), C4 -

computational gene sets (*n* = 858), C5 - GO gene sets (*n* = 1,454), C6 - oncogenic signatures (*n* = 189), and C7 - immunologic signatures (*n* = 1,910). The annotated gene sets of the NHGRI GWAS Catalog [11] were obtained from http://www.genome.gov/gwastudies (accessed April 2014).

4.2.3 Robust *Mahalanobis distance* (MD) calculation

To calculate MD, the correlation between the expression profiles of individuals was captured by the *inter-individual expression covariance*, Cov_{ab}. For expression E between any two individuals a and b, Cov_{ab} is computed as:

$$Cov_{ab} = \frac{\sum_{k=1}^{m}(E_{ak} - \mu_a)(E_{bk} - \mu_b)}{m-1},$$

where m is the number of genes in the gene set under study, and μ_a and μ_b are the mean gene expression values for individuals a and b, respectively. Given all pair-wise comparisons of individuals we obtained the inter-individual covariance matrix Cov. We employed the minimum covariance determinant (MCD) estimator [210] to compute a robust version of Cov, as implemented in the Matlab toolbox LIBRA [211]. We then computed the MD for each individual as

$$MD_i = \sqrt{(E_{i\cdot} - \vec{\mu})^T Cov^{-1}(E_{i\cdot} - \vec{\mu})},$$

where $\vec{\mu}$ is m length vector of the per-gene mean values across all individuals. The statistic

$$SSMD = \sum MD_i^2$$

was calculated for each set. To approximate the empirical null distributions for SSMD we applied resampling for gene sets with different numbers of genes, ranging

from 2 to 150. For a given number of genes *m*, we randomly sampled *m* genes from the full expression matrix without replacement, and then computed SSMD for the resampled gene set. The procedure was repeated 1,000 times for all gene sets. More permutations were performed for significant gene sets until the desired Bonferroni correction level P = 0.01 was either achieved or rejected. The resampling process breaks correlation structure between genes, hence providing a background distribution of expected random distribution of SSMD. We compared the SSMD in the observed gene set to equally-sized sets drawn at random from all assayed genes.

A chi-square plot was constructed as the I ranked MD value against the values of $\chi^2(p,m)$, where $p = (i\text{-}0.5)/I$ and m is the number of genes in the gene set. The right panel of Figure 1 is the chi-square plot that supports the multivariate outliers identified [212]. A chi-square plot draws the empirical distribution function of the square of the MD against the χ^2 distribution with degree of freedom equal to *m*. A break in the tail of the χ^2 distribution is an indicator for outliers [213], given that the square of the MD is approximately distributed as a χ^2 distribution [212,214].

4.2.4 Power analysis for SSMD test

To evaluate the sensitivity of SSMD as a statistic for detecting L-SSMD gene sets, power analyses were conducted. One selected L-SSMD gene set, POTTI_ETOPOSIDE_SENSITIVITY, was used as the test set The impacts of sample size (n) and the size of gene set (m) were considered. The selected L-SSMD gene set contained 37 genes, that is, m = 37, while the sample size n = 326. The original expression data matrix was subsampled by lowering either n or m. For each subsampled

n or *m* value, 100 random replicates of the expression data matrix were constructed. The SSMD was computed for each subsampled replicate and the significance of the observed SSMD was assessed by permutation tests, as described above for detecting L-SSMD gene sets. The more sensitive is SSMD to *n* or *m*, the less would subsampled replicates remain significant as an L-SSMD.

4.2.5 Discordant expression, heritability, and single-cell gene expression

To compute the discordant expression of genes between twin pairs, twinsUK gene expression data from the study of [117] were acquired. The discordant expression, i.e., the expression differences between each pair of twins, was measured as described previously [176]. Briefly, for each gene the RMD in expression between MZ twin pairs and between DZ twin pairs was computed. For a pair of MZ twins, *i*, for example, the RMD was computed using

$$RMD_i = \frac{\left| y_i^{MZ1} - y_i^{MZ2} \right|}{2\bar{y}_i},$$

where $\bar{y}_i$ is the arithmetic mean of the levels of gene expression for that MZ twin pair (designated as y_i^{MZ1} and y_i^{MZ2}). For each gene, the data from all MZ or DZ twin pairs were pooled to compute the mean RMD per gene, $\frac{1}{n}\sum RMD_i$, where *n* is the number of twin pairs. The computed mean RMD per gene was normalized by the value computed in the same way but with the expression data reconstructed by randomly assigning the identities of twin pairs. The values of narrow-sense heritability (h^2) of gene expression were obtained from the study of [215]. Different estimates of h^2 were also

obtained from the studies of [117] and [216]. Single-cell gene expression levels measured in 42 LCLs were acquired from the study of [217].

4.2.6 Effect size of common eSNPs

The absolute value of the slope coefficient (|β|) of the linear regression model was used as the measure of the effect size of each eSNP. Gene expression levels across individuals were normalized using Z-score to make the values of β uncorrelated with total gene expression levels. The sign of β was ignored because it is only relative against the genotypes of each eSNP, which were denoted by 0 for homozygous major alleles, 1 for heterozygous alleles, and 2 for homozygous minor alleles. Instead, an eSNP's effect direction was determined by whether the eSNP causes gene expression to shift away from or towards the mean gene expression for the majority of individuals in the populations. In this sense, the notation of genotypes (0, 1, 2) provided information of effect direction for eSNPs. If an individual's eSNP genotype is 0, then the effect of the eSNP is to maintain the same expression level for the eSNP-regulated gene between outlier individuals and the majority of individuals in the population; on the other hand, if the eSNP's genotype is 1 or 2, then the effect of the eSNP is to either increase or decrease (depending on the sign of the slope) the expression of the gene by one or two times of |β| than that of genotype 0. Therefore, the effect size was weighted by the genotype: $\boldsymbol{\beta} = |\beta| * \text{genotype}$. The genotype-scaled effect size was used in the comparison of the combined eSNP effects between outlier and non-outlier individuals.

4.2.7 Density of private SNPs in regulatory regions of L-SSMD genes

Both heterozygous and homozygous private SNPs, with an allele frequency of $1/(2N)$ and $1/N$, respectively, for each individual (where N is the number of individuals), were counted. The *cis*-regions of tested genes were split into seven subclasses of regulatory regions, according to the combined chromatin state segmentation of the ENCODE GM12878 sample [218]. The density of private SNPs in each subclass of the regions was assessed for enrichment significance by comparing the observed density with that of randomly generated control regions. To provide comprehensive controls, four different means were used to construct control regions: (1) randomly selected non-outlier individuals to replace outlier individuals, (2) randomly selected genomic regions located 10 Mb away from L-SSMD genes, (3) randomly selected shuffled L-SSMD genes in the same amount of original gene set, and (4) shuffled S-SSMD genes in the same amount of original gene sets.

4.3 Results

4.3.1 Study overview

The main results of our study comprise three parts. The first part concerns the identification of sets of functionally related genes whose expression discrepancies among individuals are significantly greater (or smaller) than those of random gene sets. The second part concerns the identification of outlier individuals whose expression profiles with respect to gene sets are significantly divergent from those of others in the population. The third part concerns evidence that private SNPs contribute to aberrant expression in outlier individuals.

Data analysis in the first two parts relied on a metric of statistical distance that can quantify the dissimilarities between individuals in the expression levels of gene sets, rather than single genes. For this purpose, we adapted MD, a multivariate metric that can be used to measure the dissimilarity between two vectors [219]. Key features of MD are illustrated in **Figure 4.1**, which shows a hypothetical example of MD, compared to simple Euclidean distance. Here, the expression levels of two genes are correlated and the Euclidean distance is not an appropriate measure of distance between data points (or individuals). MD, on the other hand, accounts for the correlation through estimating the covariance matrix from the observations, making MD a more appropriate distance statistic. With a given gene set (e.g., the two genes of the hypothetical example), we can calculate MD_i for N individuals under consideration (i = 1 to N). Each MD_i is the multivariate distance from the individual i to the population mean, with the correlation between expression profiles of individuals captured by the inter-individual expression covariance. In **Figure 4.1A**, the top three data points with largest MD_i are labeled with 1, 2, and 3, while the Euclidean distances from these data points to the population mean are not the largest. With MD_i of each individual, we can calculate the SSMD. SSMD summarizes the overall distribution of MD_i across individuals for the gene set. The squaring operation puts more weight on larger MD_i values of outlier individuals. Gene sets with larger SSMD are more likely to contain genes that are aberrantly expressed by outlier individuals. Thus, comparing SSMD values of gene sets, we can identify sets of genes that tend to (or tend not to be) aberrantly expressed (i.e., Part 1 of the main results).

The outlier individuals can be identified with ordered MD_i. To do so, we used the tool for multivariate outlier recognition, chi-square plot [212]. As seen in **Figure 4.1B**, the three data points with the largest MD_i are recognized as outliers. These data points, as shown in **Figure 4.1A**, are the most remote observations with the largest MD_i to the population mean. None of the three data points would be identified as outliers by using Euclidean distance. More important, none of them would be identified as outliers if we used any univariate approach, when the two genes are considered separately, the expression levels of either gene in the three individuals are in the "normal" range. Finally the purpose of identifying outlier individuals is to study the genetic basis of aberrant expression of genes in outliers. That is to say, once the outlier individuals are identified, the genetic variation associated with outlier individuals can be further analyzed to see what kinds of genetic variation contribute to aberrant expression (i.e., Part 2 of the main results).

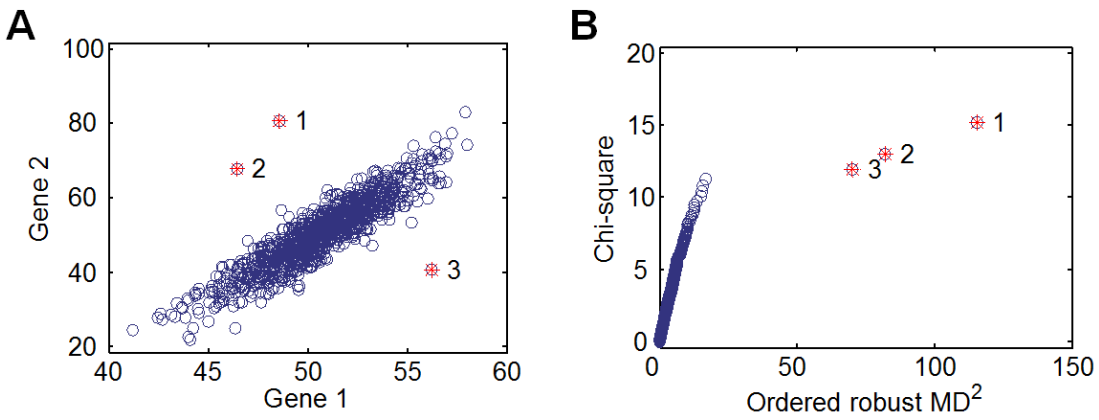


Figure 4.1 MD-based multivariate outlier detection. (**A**) Scatter plot for the expression levels of two hypothetical genes. Three outliers indicated with red stars have the largest MD values to the population mean. (**B**) The chi-square plot showing the relative position and order of the three outlier data points, compared to those of non-outlier data points.

4.3.2 Gene sets (L-SSMD) that tend to be aberrantly expressed

We started by identifying gene sets that are more likely to be aberrantly expressed. We obtained the expression data matrix of 10,231 protein-coding genes in 326 LCLs from individuals of European descent (EUR) from the Geuvadis project RNA-seq study [183]. We used SSMD to measure the total deviation of expression profiles from all individuals to the population mean for gene sets. We computed SSMD for all gene sets with fewer than 150 expressed genes in the MSigDB [209] and the GWAS catalog [11].

We identified 31 MSigDB gene sets whose SSMD values were significantly larger than those of control gene sets that contain the same number of genes randomly selected from all expressed genes (Bonferroni corrected $P < 0.01$, permutation test) (**Table 4.1**). These 31 gene sets contain 1,855 distinct genes that are more likely to be aberrantly expressed in defined outlier individuals. We named these gene sets and genes large SSMD (L-SSMD) gene sets and genes. **Figure 4.2** shows one of L-SSMD gene sets, *G-protein coupled receptor activity*, which contains 94 genes. In addition, eight GWAS catalog gene sets showed relatively large SSMD ($P < 0.001$, permutation test), though not significant following Bonferroni correction. These sets included genes implicated in adverse responses to chemotherapy, conduct disorder, fasting insulin-related traits, metabolite levels, obesity, retinal vascular caliber, temperament, or thyroid hormone levels.

4.3.3 Outlier individuals in L-SSMD gene sets

To identify outlier individuals, we applied chi-square plot to examine MD values of all individuals with respect to each of the 31 L-SSMD gene sets. We identified 17 distinct outliers in total, 11 of which were found in more than one gene set, and almost all gene sets had more than one outlier. **Figure 4.2** shows that three outliers were detected in the L-SSMD gene set, G-protein coupled receptor activity, using chi-square plot. The distributions of outliers in 31 gene sets from 17 individuals are given in **Figure 4.3**.

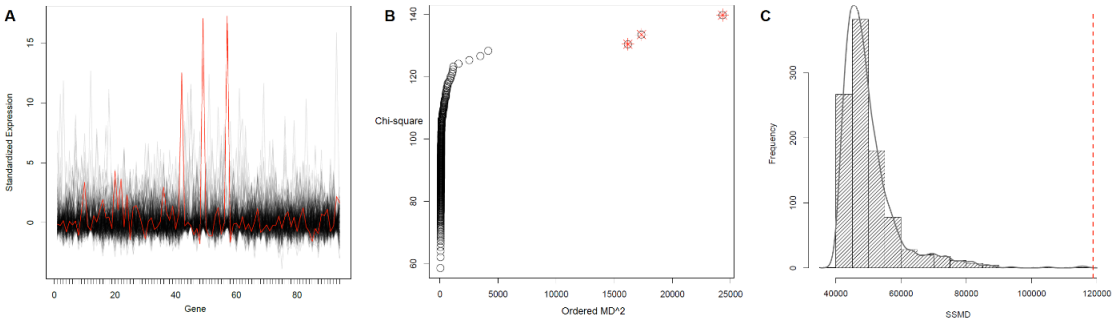


Figure 4.2 Gene expression profiles and outlier detection in the gene set, G-protein coupled receptor activity. (**A**) The expression profiles of 326 EUR samples for 94 genes in the gene set. The expression profile of the outlier individual with the largest SSMD is outlined in red. (**B**) The chi-square plot showing three outliers, as highlighted with the star symbol. (**C**) The null distribution of SSMD established from 1,000 permutations of 94 randomly selected genes. The red vertical line indicates the observed value of SSMD computed for the original gene set.

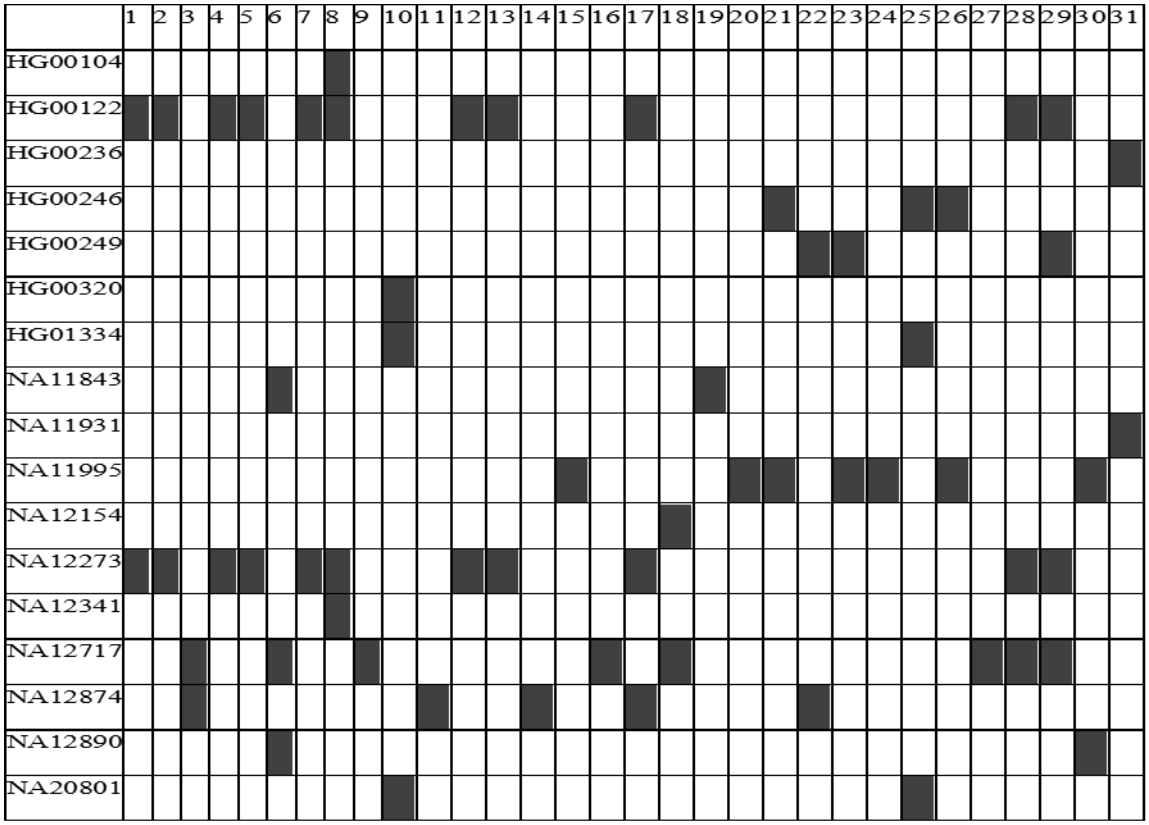


Figure 4.3 Distribution of outliers in corresponding gene sets. The 63 outliers (involving 17 distinct individuals) with respect to the 31 L-SSMD gene sets, detected by using chi-square plot, are highlighted with shaded box. The indexes of L-SSMD gene sets are given and their names are given in Table 4.1 of the main text.

4.3.4 Gene sets (S-SSMD) that tend not to be aberrantly expressed

Fourteen gene sets with significantly smaller SSMD (S-SSMD) were identified (Bonferroni corrected $P < 0.01$, **Table 4.2**). The S-SSMD genes ($n = 534$) in the 14 S-SSMD gene sets are involved in homologous recombination repair of replication-independent double-strand breaks, catalysis of the *trans*fer of a phosphate group to a carbohydrate substrate molecule, or cell cycle control. GWAS gene sets implicated in alcohol dependence and metabolic syndrome showed significantly smaller SSMD than random gene set.

4.3.5 Validation of L- and S-SSMD gene sets

We evaluated the power of SSMD as a statistic describing the propensity of a gene set for aberrant expression. We considered the influences of the sample size (*n*) and the size of gene set (*m*). In cases where the SSMD are insensitive to *n* or *m*, the power would be maintained when *n* or *m* changes. However, we found that the power dropped substantially when *n* dropped from 326 to 300 or when *m* dropped from 37 to 31, suggesting that SSMD is sensitive to both *n* and *m* (**Figure 4.4 A, B**). This might be due to only a small number of genes in the gene set tested being aberrantly expressed in a few individuals and the power analyses for *m* and *n* being based on the sub-sampling of genes and individual samples (Materials and Methods).

Table 4.1 Gene sets that tend to be aberrantly expressed in LCLs from individuals of European descent. The names of gene sets and MSigDB subclasses are given. Number of genes (#) shows the number of genes included in SSMD computation / the number of genes in the original gene set.

Gene set		# of genes
C2: curated gene sets (Chemical and genetic perturbations, Reactome gene sets)		
1. AIGNER_ZEB1_TARGETS	Genes up-regulated in MDA-MB-231 cells (breast cancer) after knockdown of ZEB1 [GeneID=6935] by RNAi	28 / 35
2. CAFFAREL_RESPONSE_TO_THC_8HR_3_UP	Genes up-regulated in EVSA-T cells (breast cancer) treated with 3 micromolar THC (delta-9-tetrahydrocannabinol) [PubChem=6610319] for 8 h.	5 / 5
3. GAUSSMANN_MLL_AF4_FUSION_TARGETS_E_UP	Up-regulated genes from the set E (Fig. 5a): specific signature shared by cells expressing either MLL-AF4 [GeneID=4297;4299] or AF4-MLL fusion proteins alone, and those expressing both fusion proteins.	76 / 97
4. HOFMANN_MYELODYSPLASTIC_SYNDROM_RISK _UP	Genes up-regulated in bone marrow hematopoietic stem cells (HSC, CD34+ [GeneID=947]) from patients with high risk of myelodysplastic syndrom (MDS) compared to the low risk patients.	19 / 24

Table 4.1 continued

Gene set		# of genes
5. IWANAGA_CARCINOGENESIS_BY_KRAS_UP	Cluster 3: genes up-regulated in lung tissue samples from mice with tumor-bearing genotypes (activated KRAS [GeneID=3845] alone or together with inactivated PTEN [GeneID=5728]).	141 / 170
6. LEIN_CHOROID_PLEXUS_MARKERS	Genes enriched in choroid plexus cells in the brain identified through correlation-based searches seeded with the choroid plexus cell-type specific gene expression patterns.	79 /103
7. LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN	Genes down-regulated between two breast carcinoma subtypes: metaplastic (MCB) and ductal (DCB).	77 / 114
8. LIU_PROSTATE_CANCER_UP	Genes up-regulated in prostate cancer samples.	79 / 96
9. MASRI_RESISTANCE_TO_TAMOXIFEN_AND_AROMATASE_INHIBITORS_UP	Genes up-regulated in derivatives of MCF-7aro cells (breast cancer) that developed resistance to tamoxifen [PubChem=5376] or inhibitors of aromatase (CYP19A1) [GeneID=1588].	11 / 20
10. MIKKELSEN_MEF_ICP_WITH_H3K27ME3	Genes with intermediate-CpG-density promoters (ICP) bearing the tri-methylation mark at H3K27 (H3K27me3) in MEF cells (embryonic fibroblasts).	115 / 206

Table 4.1 continued

Gene set		# of genes
11. PEPPER_CHRONIC_LYMPHOCYTIC_LEUKEMIA_DN	Genes down-regulated in CD38+ [GeneID=952] CLL (chronic lymphocytic leukemia) cells.	11 / 21
12. POTTI_ETOPOSIDE_SENSITIVITY	Genes predicting sensitivity to etoposide [PubChem=36462].	37 / 43
13. QI_PLASMACYTOMA_DN	Down-regulated genes that best disciminate plasmablastic plasmacytoma from plasmacytic plasmacytoma tumors.	85 / 100
14. REACTOME_CGMP_EFFECTS	Genes involved in cGMP effects	15 / 19
15. REACTOME_LIGAND_GATED_ION_CHANNEL_*TRANS*PORT	Genes involved in Ligand-gated ion channel *trans*port	6 / 21
16. VANHARANTA_UTERINE_FIBROID_UP	Genes up-regulated in uterine fibroids vs normal myometrium samples.	39 / 45
17. WU_CELL_MIGRATION	Genes associated with migration rate of 40 human bladder cancer cells.	143 / 184
C3: motif gene sets (microRNA targets)		
18. TCCAGAG,MIR-518C	Targets of MicroRNA TCCAGAG,MIR-518C	132 / 148
C4: computational gene sets (cancer modules, cancer gene neighborhoods)		

Table 4.1 continued

Gene set		# of genes
19. MODULE_122	Genes in the cancer module 122	111 / 141
20. MODULE_215	Genes in the cancer module 215	3 / 15
21. MODULE_274	Genes in the cancer module 274	44 / 82
22. MORF_BCL2L11	Neighborhood of BCL2L11	123 / 188
23. MORF_MYL3	Neighborhood of MYL3	44 / 71
C5: GO gene sets (GO biological process, GO molecular function)		
24. EXTRACELLULAR_LIGAND_GATED_ION_CHANNEL_ACTIVITY	Genes annotated by the GO term GO:0005230. Catalysis of the *trans*membrane *trans*fer of an ion by a channel that opens when a specific extracellular ligand has been bound by the channel complex or one of its constituent parts.	14 / 22
25. G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY	Genes annotated by the GO term GO:0004930. A receptor that binds an extracellular ligand and *trans*mits the signal to a heterotrimeric G-protein complex. These receptors are characteristically seven-*trans*membrane receptors and are made up of hetero- or homodimers.	94 / 191

Table 4.1 continued

Gene set		# of genes
26. *TRANS*MISSION_OF_NERVE_IMPULSE	Genes annotated by the GO term GO:0019226. The sequential electrochemical polarization and depolarization that travels across the membrane of a nerve cell (neuron) in response to stimulation.	108 / 189
C6: oncogenic signatures		
27. MEL18_DN.V1_DN	Genes down-regulated in DAOY cells (medulloblastoma) upon knockdown of PCGF2 [GeneID=7703] gene by RNAi.	104 / 148
C7: immunologic signatures		
28. GSE19825_NAIVE_VS_DAY3_EFF_CD8_TCELL_UP	Genes up-regulated in comparison of naive CD8 T cells versus effector CD8 T cells.	128 / 200
29. GSE19825_NAIVE_VS_IL2RALOW_DAY3_EFF_CD8_ TCELL_UP	Genes up-regulated in comparison of naive CD8 T cells versus effector CD8 IL2RA [GeneID=3559] low T cells at.	133 / 200
30. GSE3982_NKCELL_VS_TH2_UP	Genes up-regulated in comparison of NK cells versus Th2 cells.	136 / 200

Table 4.1 continued

Gene set		# of genes
31. GSE8515_CTRL_VS_IL6_4H_STIM_MAC_DN	Genes down-regulated in comparison of untreated macrophages versus those treated with IL6 [GeneID=3569].	144 / 200

Table 4.2 Gene sets that tend not to be aberrantly expressed in LCLs from individuals of European descent.

Gene set		# of genes
C2: curated gene sets (Canonical pathways, KEGG, Reactome, BioCarta, chemical and genetic perturbations)		
1. PID_ATM_PATHWAY	ATM pathway	34 / 34
2. KEGG_HOMOLOGOUS_RECOMBINATION	Homologous recombination	28 / 28
3. MORI_PRE_BI_LYMPHOCYTE_DN	Down-regulated genes in the B lymphocyte developmental signature, based on expression profiling of lymphomas from the Emu-myc *trans*genic mice: the Pre-BI stage.	73 / 77
4. XU_RESPONSE_TO_TRETINOIN_UP	Genes up-regulated in NB4 cells (acute promyelocytic leukemia, APL) by tretinoinalone.	14 / 16

Table 4.2 continued

Gene set		# of genes
5. FLECHNER_PBL_KIDNEY_*TRANS*PLANT_OK_VS_DONOR_DN	Genes downregulated in peripheral blood lymphocytes (PBL) from patients with well functioning kidneys more than 1-year post *trans*plant compared to those from normal living kidney donors	40 / 41
6. GARGALOVIC_RESPONSE_TO_OXIDIZED_PHOSPHOLIPIDS_LIGHTYELLOW_UP	Genes from the lightyellow module which are up-regulated in HAEC cells (primary aortic endothelium) after exposure to the oxidized 1-palmitoyl-2-arachidonyl-sn-3-glycerophosphorylcholine (oxPAPC).	11 / 11
7. REACTOME_HOMOLOGOUS_RECOMBINATION_REPAIR_OF_REPLICATION_INDEPENDENT_DOUBLE_STRAND_BREAKS	Genes involved in Homologous recombination repair of replication-independent double-strand breaks	16 / 17
8. REACTOME_G1_PHASE	Genes involved in G1 Phase.	35 / 38
9. BIOCARTA_ATRBRCA_PATHWAY	Role of BRCA1, BRCA2 and ATR in Cancer Susceptibility .	21 / 21
C4: computational gene sets (cancer modules, cancer gene neighborhoods)		
10. MODULE_87	Genes in the cancer module 87	44 / 44

Table 4.2 continued

Gene set		# of genes
11. MORF_PRKAR1A	Neighborhood of PRKAR1A protein kinase, cAMP-dependent, regulatory, type I, alpha (tissue specific extinguisher 1) in the MORF expression compendium	139 / 142
12. MORF_REV3L	Neighborhood of REV3L	55 / 57
13. GNF2_DDX5	Neighborhood of DDX5 DEAD (Asp-Glu-Ala-Asp) box polypeptide 5 in the GNF2 expression compendium	62 / 63
C5: GO gene sets (GO molecular function)		
14. CARBOHYDRATE_KINASE_ACTIVITY	Genes annotated by the GO term GO:0019200. Catalysis of the *trans*fer of a phosphate group, usually from ATP, to a carbohydrate substrate molecule.	15 / 15

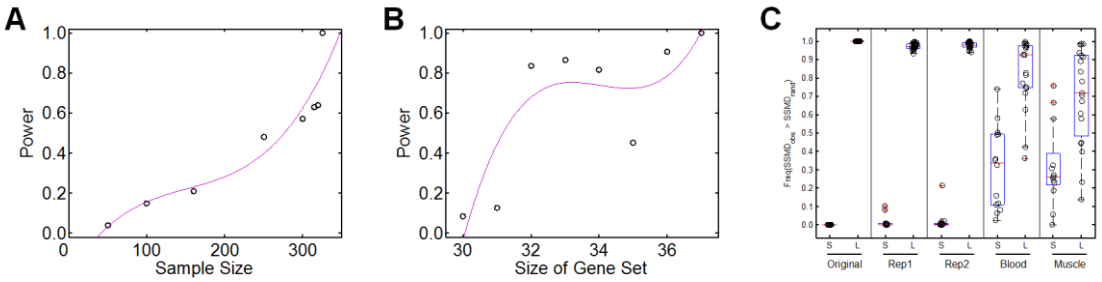


Figure 4.4 Power of SSMD test and validation of significant L- and S-SSMD gene sets. (**A**) The change of power as a function of sample size. (**B**) The change of power as a function of the size of a gene set. (**C**) Validation of significant L- and S-SSMD gene sets using different expression data. Original: Geuvadis LCL expression data normalized using PEER (i.e., data used for the main results); Rep1: first set of replication of Geuvadis LCL expression data without PEER normalization; Rep2: second set of replication of Geuvadis LCL expression data without PEER normalization; Whole blood: GTEx whole blood expression data; and Muscle: GTEx muscle expression data. The boxplot shows the frequency of observed SSMD is greater than the control SSMD of 1,000 random replicates.

Owning to the observed sensitivities, it was necessary to validate our results for the L- and S-SSMD gene sets, which were obtained using the Geuvadis LCL expression data [183]. This was accomplished by taking into consideration three factors: (1) robustness against the influence of data normalization methods, (2) replicability against technical variability, and (3) reproducibility against independent expression data of different tissues.

The "original" Geuvadis expression data we used to identify L- and S-SSMD gene sets had been normalized by using the algorithm of PEER [220,221]. We first showed that the PEER normalization algorithm did not change our results. To do so, we downloaded the "raw" Geuvadis expression data quantified in reads per kilobase per million (RPKM) without PEER normalization. Two replicate sets of raw RPKM data were available for most of the Geuvadis samples. We used each set independently to test

the significance of SSMD for L- and S-SSMD gene sets against random control sets. The procedure was similar to what we used for establishing the original L- and S-SSMD gene sets. Briefly, for each L- or S-SSMD gene set, we tested whether the SSMD computed with raw RPKM data tended to be larger or smaller than that of random gene sets. The observed SSMD was compared against SSMD values computed from 1,000 replicates of randomly selected genes and the significance was evaluated by examining how many times the observed SSMD was larger or smaller than random SSMD. As expected, with the original (PEER normalized) expression data, all 31 L-SSMD gene sets had a larger SSMD than sets of randomly selected genes, while all 14 S-SSMD gene sets had a smaller SSMD. The same patterns were recovered with the raw RPKM expression data (**Figure 4.4C**). These results indicated that our results for L- or S-SSMD gene sets were robust against the normalization methods used and inherent technical variability in the measurements.

We used independent gene expression data from additional tissues to validate our results. Data from whole blood and muscle (in 156 and 138 samples, respectively) from the pilot study of the Genotype-Tissue Expression project (GTEx) [222] were used to re-compute SSMD and to conducted the same validation tests that were performed for the LCL data. With the GTEx data, the frequency of observed SSMD greater than random SSMD was significantly higher for L-SSMD gene sets than S-SSMD gene sets (K-S test, P = 1.02e-5 and 9.9e-4, for whole blood and muscle, respectively, **Figure 4.4C**). These results suggest that gene sets tending to have larger observed SSMD in LCL were more likely to have larger SSMD in the other two tested tissues, and vice versa. The

consistency in the direction of SSMD patterns among disparate tissue types validates the biological significance of L- and S-SSMD gene sets.

4.3.6 Differences in aberrant expression between Europeans and Africans

Next we examined which gene sets show strong population-specific SSMD. For a given gene set, we first computed MD_i with the gene expression data for all 402 samples of both European (EUR, n = 326) and African (AFR, n = 76) ancestries. We then use these MD_i to compute $SSMD_{EUR}$ and $SSMD_{AFR}$ for EUR and AFR samples, respectively, and calculated the difference in SSMD between them: $diffSSMD_{EUR\text{-}AFR} = SSMD_{EUR}\text{-}SSMD_{AFR}$. To assess the significance, we computed $diffSSMD_{rand}$ by randomly assigning samples without regard to their identities of original populations. For each gene set, we computed 1,000 permutations of $diffSSMD_{rand}$ to obtain the null distribution of expected $diffSSMD_{EUR\text{-}AFR}$. We compared the value of $diffSSMD_{EUR\text{-}AFR}$ with the null distribution to obtain its significance.

We used two random sets of genes (n = 20 and 40) to show that the values of *diffSSMD* were proportional to gene set size and changed linearly with the ratio by which the total samples were partitioned into two sub-groups (**Figure 4.5A**). In this test, we ignored the EUR and AFR ancestries of samples. We randomly shuffled the 402 samples, partitioned them to two sub-groups with different ratios (such as, 201/201 or 326/76), and computed the *diffSSMD* between the two sub-groups. We repeated this 1,000 times per ratio to obtain null distributions of *diffSSMD*. We found that, regardless of gene set size, when samples were partitioned into groups of equal size (i.e., 201/201), the average *diffSSMD* was close to zero. When samples were partitioned unequally, the

average value of *diffSSMD* increased with the degree of inequality in a linear manner. When the ratio of partition was fixed (e.g., 326/76, the actual sample ratio of EUR and AFR), the average *diffSSMD* reflected the size of the gene set (e.g., twice as large for the 40-gene set as the 20-gene set). When both the ratio of partition and the gene set was fixed, as we did in the real test for each gene set, the values of null *diffSSMD* fluctuated only due to the random assignment of samples into the two sub-groups. Similarly, in our significance test for $diffSSMD_{EUR\text{-}AFR}$, both the gene set size and the ratio of partition (=326/76) were fixed, and the null distribution of *diffSSMD*, $diffSSMD_{rand}$, was constructed from 1,000 random repeats of the partition of shuffled samples. An observed $diffSSMD_{EUR\text{-}AFR}$ was considered to be significant when it was greater or smaller than all values of $diffSSMD_{rand}$.

In total, 231 gene sets showed significantly smaller $diffSSMD_{EUR\text{-}AFR}$ than $diffSSMD_{rand}$ in our analysis. For these gene sets, the differences between $SSMD_{EUR}$ and $SSMD_{AFR}$ were relatively smaller than those differences calculated when EUR and AFR individuals were randomly assigned. This was likely caused by the relatively large $SSMD_{AFR}$ in real data. In other words, AFR samples were more likely to produce disproportionally larger SSMD than EUR samples.

In contrast, only four gene sets showed the opposite pattern—that is $diffSSMD_{EUR\text{-}AFR}$ significantly larger than $diffSSMD_{rand}$. Genes in these four sets included: (1) genes involved in the process preventing the degeneration of the photoreceptor (a specialized cell type that is sensitive to light), (2) genes down-regulated in prostate tumor (a tumor with distinct signatures differentiate between African-American and

European-American patients [223]), (3) genes associated with malignant fibrous histiocytoma tumors, and (4) genes up-regulated in colon tissue upon the knockout of *MBD2*, a methyl-CpG binding protein that mediates the methylation signal.

Finally, a power analysis for $diffSSMD_{\text{EUR-AFR}}$ was conducted using the first gene set among the four with significantly larger $diffSSMD_{\text{EUR-AFR}}$. The result suggested that the difference in sample size between EUR and AFR had little impact on the sensitivity of asserting that the tested gene set was significant. As shown in **Figure 4.5B**, when the EUR were subsampled from 326 to 76 (the sample size of AFR), the power of *diffSSMD* only slightly decreased.

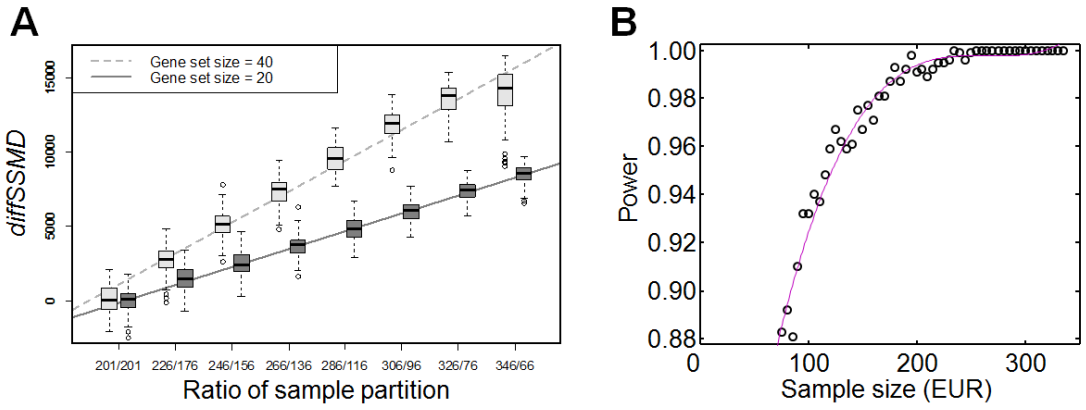


Figure 4.5 Change of *diffSSMD* as a function of the ratio between partitioned samples and the power of *diffSSMD* test under varying sample size. (**A**) The change of *diffSSMD* as a function of the size ratio of partitioned samples. The results with respect to two gene sets of size 20 and 40 are shown. For each ratio of partition, the distribution of $diffSSMD_{rand}$ was constructed from 100 randomly shuffled samples. (**B**) The change of the power of the *diffSSMD* test between EUR and AFR populations for the population-specific effect as a function of the size of EUR samples. The red line is fitted by using polynomial regression with the cubic model.

4.3.7 Genetic and nongenetic factors contributing to aberrant expression

To evaluate the contributions of genetic or nongenetic factors in causing aberrant expression, we utilized three statistical metrics to characterize L- and S-SSMD genes

and compared the properties of the two groups of genes (Materials and Methods). The three metrics were: (1) discordant gene expression, measured as the RMD in gene expression, between twin pairs, considering both MZ and DZ twins [176]; (2) the narrow-sense heritability (h^2) of gene expression [215]; and (3) the CV of single-cell gene expression [217].

Discordant expression between twin pairs in L-SSMD genes is greater than that in S-SSMD genes in both types of twins (P = 2.8e-15 between MZ pairs and 3.0e-34 between DZ pairs; K-S test, **Figure 4.6A**), but the difference is more pronounced for L-SSMD genes than for S-SSMD genes (P = 5.6e-23 and 5.4e-6 for L- and S-SSMD genes, respectively). The more pronounced discordant expression between MZ pairs for L-SSMD genes, compared to S-SSMD genes, is likely due to the effect of environmental factors. L-SSMD genes may have increased sensitivity to environmental factors. On the other hand, regardless of L- or S-SSMD genes, the discordant expression is always greater between DZ pairs than between MZ pairs. This suggests that genetic diversity increases the level of discordance in gene expression.

Expression levels of L-SSMD genes tend to have a smaller h^2 than S-SSMD genes (P = 3.6e-5, K-S test, **Figure 4.6B**). Similar results were obtained with different h^2 estimates (e.g., those using data from another twin cohort [117] and those using data from unrelated individuals [216]). Furthermore, L-SSMD genes showed greater expression variability at the single-cell level than S-SSMD (P = 7.7e-21, K-S test, **Figure 4.6C**). Forty genes were found to be shared between L-SSMD and S-SSMD

groups. Excluding these overlapping genes did not qualitatively change any of the results described above.

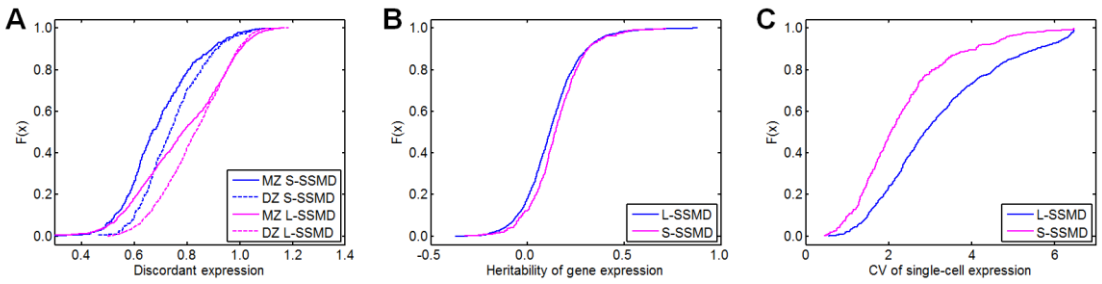


Figure 4.6 Differences in expression discordance, heritability, and variability between L- and S-SSMD genes. (**A**) Normalized mean discordant expression (measure as the relative mean difference, RMD) per gene. (**B**) Heritability of gene expression. (**C**) CV of single-cell expression.

4.3.8 Common regulatory variation is not responsible for aberrant expression

To evaluate the contribution of eQTLs to aberrant expression, we obtained 419,983 *cis*-acting eQTL SNPs (eSNPs) associated with 13,703 genes from a previous study [183]. We found that 20.3% of L-SSMD genes and 19.3% of S-SSMD genes have *cis*-eSNP(s). That is to say, there is no difference in *cis*-eSNP occurrence between L- and S-SSMD genes ($P = 0.67$, Fisher's exact test). Due to the prevalence of eSNPs, this result was not unexpected.

Next we examined whether outlier individuals are more likely to have an eQTL genotype that might explain their outlier status. In particular, we calculated the genotype-scaled effect size ($\boldsymbol{\beta} = |\beta|$*genotype, where genotype = [0, 1, 2], to take into account of the direction of the effect) for all *cis*-eSNPs of associated genes in L-SSMD gene sets for outlier individuals. Multiple eSNPs in the same genes were treated independently and the values of genotype-scaled effect sizes were pooled together as

$\boldsymbol{\beta}_{\text{outlier}}$. We did the same calculation for the same sets of genes for all non-outlier individuals to obtaine $\boldsymbol{\beta}_{\text{non-outlier}}$.

We hypothesized that if *cis*-eSNPs cause the outlier's gene expression level to deviate away from the population mean, then the genotype-scaled effect size of these eSNPs in outlier individuals should be less likely to be zero and more likely to be larger than that of non-outlier individuals. However, we found that 45.3% of $\boldsymbol{\beta}_{\text{outlier}}$ (n = 24,649, pooling from 63 outlier-gene pairs, i.e., pairs of outlier individual and gene in corresponding gene sets) and 46.2% of $\boldsymbol{\beta}_{\text{non-outlier}}$ (n = 3,329,296, pooling from 309 outlier-gene pairs) were zeros, showing that there was no difference between the two fractions (P = 0.086, χ^2 test). Considering that this result might be affected by LD between eSNPs that is not accounted for by our model, we performed the analysis again using only the most significant eSNP per gene. With this single-eSNP setting, we found that 9.49% of $\boldsymbol{\beta}_{\text{outlier}}$ (n = 875, pooling from 63 outlier-gene pairs) and 10.58% of $\boldsymbol{\beta}_{\text{non-outlier}}$ (n = 118,965, pooling from 309 outlier-gene pairs) were zeros. Again, there was no difference between the two fractions (P = 0.3448, χ^2 test). Furthermore, using only the most significant *cis*-eSNP per gene, we found that the distribution of nonzero $\boldsymbol{\beta}_{\text{outlier}}$ was highly similar to that of nonzero $\boldsymbol{\beta}_{\text{non-outlier}}$ (K-S test, P = 0.67, **Figure 4.7**).

These results suggest that eSNPs, as commonly-occurring regulatory genetic variants, may not be responsible for aberrant expression of genes under their regulation.

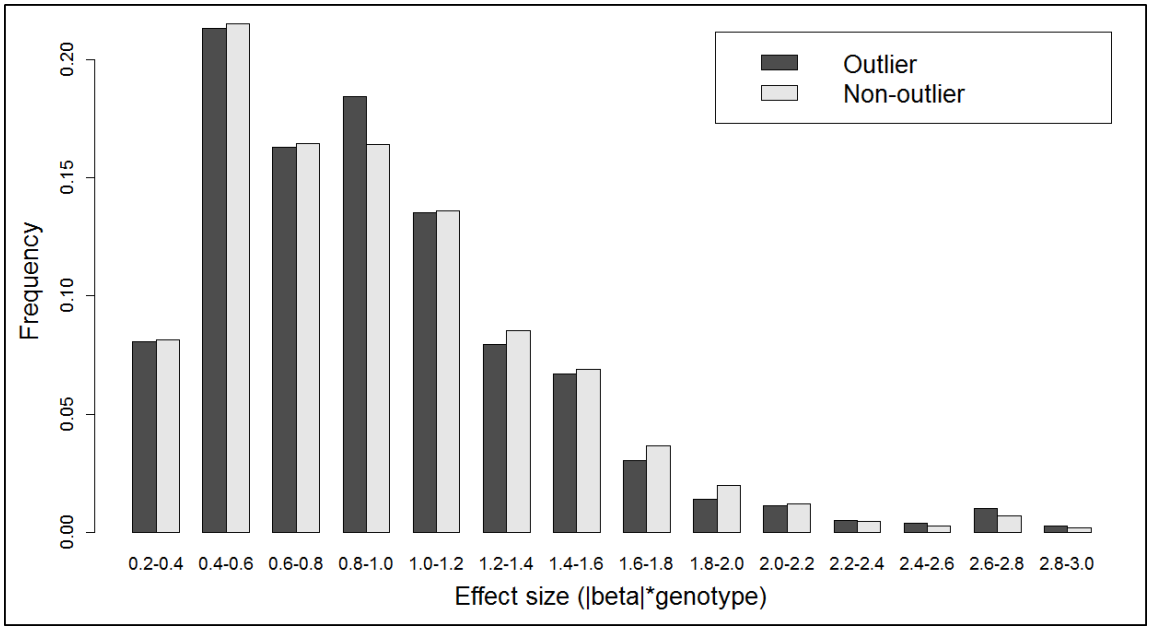


Figure 4.7 Distributions of nonzero effect size **β** of *cis*-eSNPs of L-SSMD genes in outlier and non-outlier individuals. The effect size **β** is genotype-weighted (i.e., **β** =|β|*genotype, where genotype = [0, 1, 2]).

4.3.9 Private variants may be responsible for aberrant expression

We next considered whether private SNPs might be responsible for aberrant expression by testing to see if private SNPs are enriched in regulatory regions of L-SSMD genes in outlier individuals. Private SNP density was calculated by pooling SNPs that were found uniquely in each outlier individual within the predefined 1Mb *cis*-regulatory regions of L-SSMD genes. Based on ENCODE annotations [218], regulatory regions were divided into seven subclasses, namely, E (predicted enhancer), TSS (predicted promoter region including TSS), T (predicted *trans*cribed region), PF (predicted promoter flanking region), CTCF (CTCF-enriched element), R (predicted repressed or low-activity region), and WE (predicted weak enhancer or open chromatin *cis*-regulatory element).

We found that the density of private SNPs in E regions of L-SSMD genes in outlier individuals was significantly higher than that in the same E regions in non-outlier individuals ($P < 0.001$, one-tailed *t* test). The density was also significantly higher than that derived from three additional control settings, including the reconstructed E regions from locations 10 Mb away from genes and randomly selected L-SSMD or S-SSMD genes (Materials and Methods). In summary, we randomly selected individuals or genes in four different manners to construct the control scenario, from which the private SNP density was calculated and compared with the observed density. The most salient finding was that for the E regions, the observed density of private SNPs in L-SSMD genes was significantly higher than any of the controls (**Table 4.3**). In addition, we also found that, for TSS, the density is significantly higher than in three controls ($P < 0.001$, one-tailed *t* test). These results are consistent with the findings of a previous study, which also focused on the effects of rare variant on causing outlier expression [224]. The rest of the region classes showed less significant enrichment or similar levels of density (**Table 4.3**). For illustrative purpose, two private SNPs, rs189458147 and rs117086221, located in E region of *PMAIP1* and TSS region of *NEIL1* are depicted (**Figure 4.8**).

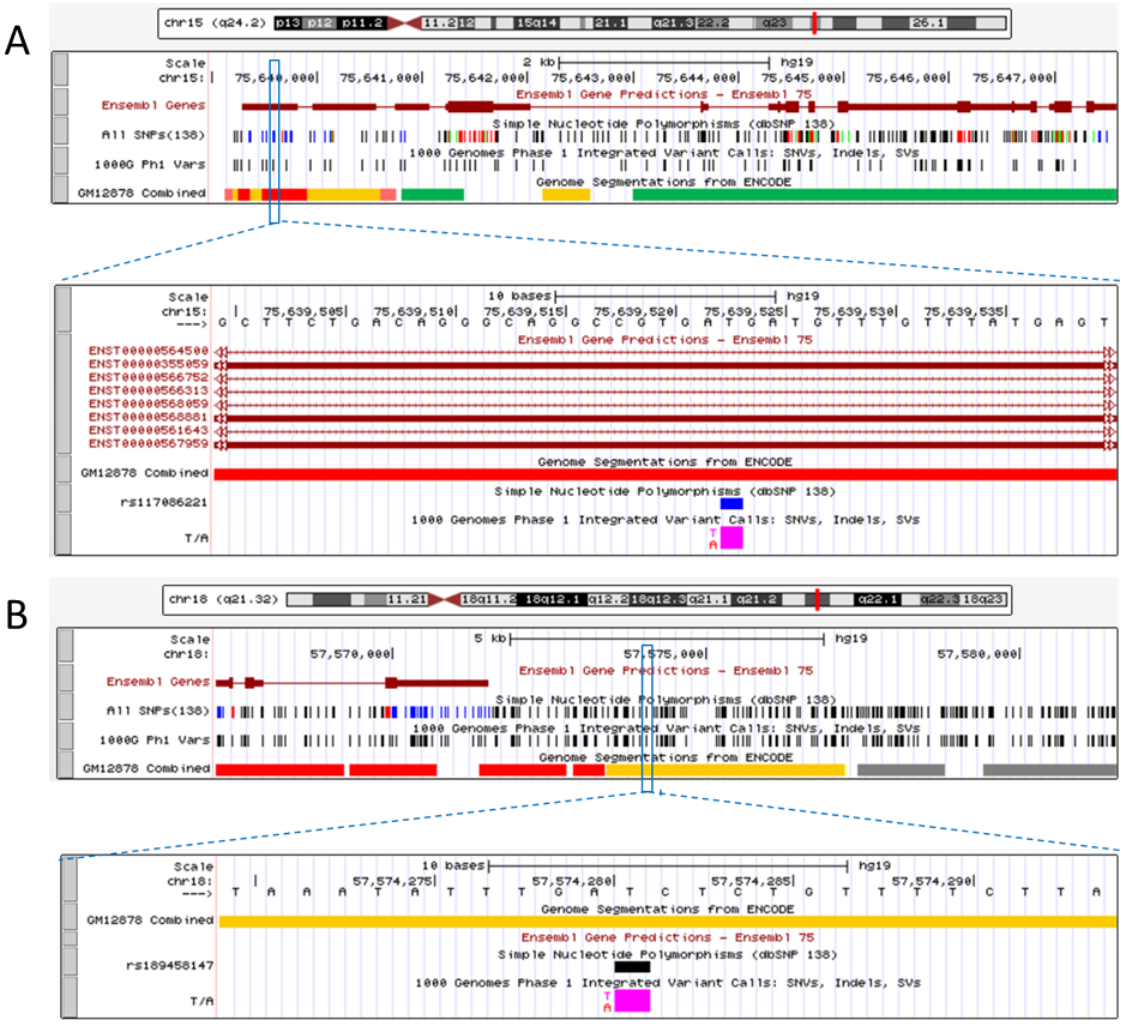


Figure 4.8 Private SNPs located in ENCODE E (predicted enhancer) and TSS (predicted *trans*cribed region) regions of corresponding L-SSMD genes. (A) Rs117086221 is located in the TSS region of gene *NEIL1* in the individual NA12154. (B) Rs189458147 locates in the potential E region of gene *PMAIP1* in the individual HG00122.

Table 4.3 Density of private SNPs in ENCODE regulatory regions of L-SSMD genes. The symbol * indicates the SNP density in the corresponding control regions is significantly lower than that in the test regions of the outlier. The significance is assessed by one-tailed t test at the level of P = 0.001. Control 1: randomly selected non-outlier individuals to replace outlier individuals. Control 2: randomly selected genomic region located 10 Mb away from L-SSMD genes. Control 3: randomly shuffled L-SSMD genes in equal number to the original gene set. Control 4: randomly shuffled S-SSMD genes in equal number to the original gene set.

		Density of private SNP (per million bp)				
Abbreviation	Description	Oberserved (#/Mb)	Control 1	Control 2	Control 3	Control 4
E	Predicted enhancer	2.07 (308/149)	1.54*	1.41*	1.76*	1.73*
TSS	Predicted promoter region including *trans*cription start site	1.91 (408/214)	1.51*	1.23*	1.45*	1.82
CTCF	CTCF enriched element	1.89 (213/113)	1.71*	1.34*	1.56*	1.79
T	Predicted *trans*cribed region	2.00 (4184/2092)	1.79*	1.52*	1.83	1.92
PF	Predicted promoter flanking region	1.79 (94/53)	1.46*	1.33*	1.69	1.96
R	Predicted repressed or low activity region	1.88 (10152/5400)	1.72*	1.45*	1.68*	1.79
WE	Predicted weak enhancer or open chromatin *cis* regulatory element	1.93 (102/53)	1.59*	1.63*	2.15	2.00
UNCL	Unclassified region	1.64 (833/508)	1.41*	0.84*	1.53	1.60

4.4 Discussion

We have used MD as a measure of distance between two points in the space defined by two or more correlated variables to quantify the deviation of individuals' gene-set expression to the population mean. This quantity allowed us to identify expression outliers. The sum of squares this quantity across individuals (i.e., SSMD) allowed us to assess how likely a gene set is to be aberrantly expressed in outlier individuals. As expected, genes involved in fundamental molecular functions and metabolic pathways are unlikely to be aberrantly expressed, showing a small SSMD. In contrast, genes in the gene sets with large SSMD tend to be involved in regulation of cellular processes and modulation of signal *trans*duction (see **Table 4.1** Notably, three gene sets with large SSMD have GO distinctive definitions: (1) extracellular ligand gated ion channel activity, (2) G-protein coupled receptor activity, and (3) *trans*mission of nerve impulse. G-protein coupled receptors constitute a large protein family of receptors that sense molecules outside the cell and activate inside signal *trans*duction pathways, implicated in various human diseases and development processes [225-227].

Widespread genetic regulatory variants have been uncovered by eQTL analyses. Most eQTLs are detected on the basis of significant linear regression between genotype and gene expression level. The inherent limitation of this method is that only commonly-occurring regulatory genetic variants will be discovered. Our analysis of *cis*-acting eQTLs in gene sets suggests that the observed patterns of expression are unlikely to be related to commonly-occurring regulatory genetic variation. Our finding that eQTLs are less likely to be responsible for aberrant expression of genes under their regulation

underscores the technical limitation of the eQTL method in dealing with gene expression regulation in outliers.

Instead, we discovered that private SNPs are likely to be responsible for a large proportion of gene expression anomalies. Our results suggest that private SNPs are significantly enriched in enhancer and promoter regions of aberrantly-expressed genes. This is in agreement with the findings of [224], in which Montgomery and colleagues reported evidence of rare SNPs underlying large changes in gene expression by calculating whether individuals with outlier array expression values were enriched for rare genetic variants. They used Z-scores as a measurement of the magnitudes of deviations from the mean of the sample. They found that individuals with gene expression Z-scores ≥ 2 have an excess of rare variants within 100 kb of the *trans*cription start site of anomalously expressed genes. The signal was found to be statistically significant for rare variants located in highly conserved sites [224]. Taken together, results from [195} and our present study suggest that rare or private SNPs contribute to large changes in gene expression. Awareness of this effect is important as it means that a rare genetic variant, even if observed in an individual genome, could potentially be regulating the expression of the phenotype to an extreme extent relative to the population mean. This makes sense from a population genetics standpoint because the recent explosion of human population size has created an abundance of rare genetic variants [80]. These variants, segregating in small groups of people or single individuals, have not been subject to the test of natural selection, and thus can potentially have stronger functional consequences. They may underlie aberrant gene expression and may

also underlie susceptibility to complex diseases. Therefore, the individual bearing private SNPs causing aberrant gene expression might be an interesting model of phenotypes relevant to the function of the aberrantly expressed gene. Otherwise, on the population level, the variants may bear little relevance to disease susceptibility phenotypes.

Intrinsic properties of gene sets are defined not only by descriptive functions of genes they include but also several measurable genetic metrics. Combined use of these metrics has demonstrated the contribution of both genetic and environmental factors to aberrant expression. First, twin data facilitated the dissection of the contributions of genetic and nongenetic factors. The discordance in gene expression is expected to be larger between pairs of DZ twins than between pairs of MZ twins, as the phenotypic difference between DZ pairs may result from both genetic and environmental effects. We indeed observed the difference between MZ and DZ in discordant expression as expected, and to the same extent for genes tending to and tending not to be aberrantly expressed. This result suggests that genetic diversity increases overall expression variability. More importantly, we found that the discordant expression in MZ pairs for genes tending to be aberrantly expressed is greater than that for genes that tend not to be aberrantly expressed. This result suggests that under the same genetic background, aberrantly expressed genes are more likely to be sensitive to the change of environmental factors than non-aberrantly expressed genes. Second, heritability is a dimensionless measure of the weight of genetic factors in explaining the phenotypic variation among individuals [228-230]. We showed that genes with small SSMD have a

higher narrow-sense heritability (h^2) of gene expression than genes with large SSMD. Third, we detected that genes tending to be aberrantly expressed have a higher expression variability at the single-cell level than genes tending not to be aberrantly expressed. This result suggests that intrinsic single-cell expression contributes to aberrant expression.

In summary, we leveraged the 1,000 genomes RNA-seq data to identify aberrant gene expression in humans, and described a multivariate framework for detecting aberrantly expressed gene sets and outlier individuals, offering a new way of measuring inter-individual variation in gene expression. This novel perspective on how to measure differences in gene expression between individual human subjects may provide important clues to the mechanisms of human adaptation, and may also be helpful for the growing field of personalized medicine.

CHAPTER V

SUMMARY AND CONCLUSION

The impact of genetic variants on gene expression variation and variance are two major topics in modern biological inquiry. Historically, most research has focused on the contribution of common allelic variants to gene expression variation (eQTL) and ignored the sources of allelic expression variance (evQTL). In addition, existing methods for eQTL studies are designed for assessing the effects of common variants, but not rare variants. On one hand, increasing evidence had demonstrated that gene expression variance, as a heritable and quantitative trait, is also under genetic control; thus understanding the relationship between genetic variants and gene expression variance has a wide range of application in evolutionary biology, medical genetics, and agriculture selection programs. On the other hand, the advent of NGS technologies has uncovered a large number of rare SNPs in human and other populations, which has created a critical need to develop a new methodologies for assessing the biological significance of these rare variants. In this dissertation, I presented three chapters that describe the major work I accomplished during my Ph.D. research to investigate the effect of common genetic variants on gene expression variance and interrogate the impact of rare variants on gene expression.

In Chapters II and III, I described a systematic exploration of genome-wide association between common genetic variants and gene expression variability using

genotype and expression datasets from TwinsUK study and Geuvadis project study, respectively. Our results showed a significant association between variances of gene expression and specific genotypes, and highlight the importance of accounting for widespread variance-controlling variants (evQTLs) in the human populations genome-wide association analyses of humans and other species. In addition, we considered and contrasted two distinct effects that may generate evQTLs: interactions between genetic variants (GxG), and interactions between genotype and environment (GxE). We explicitly searched for signs of GxG that explain the formation of evQTLs, but found few such cases. Instead, our experiment results showed that for an individual homozygous for an evQTL SNP allele associated with larger variance, the stochastic noise of the evQTL gene's expression (i.e., the random fluctuation of gene expression within replicated measures) is more pronounced than that in another individual with the evQTL SNP allele associated with smaller variance. This striking finding links gene expression variance estimated between individuals with that estimated between replicated measures, suggesting a consistent action of decanalization driven by GxE evQTL at the two different levels. Despite the progress of the evQTL study, evQTL remains largely a statistically described, population-level phenomenon, with insufficient experimental support to suggest molecular mechanisms. There is a critical need to validate the variance-controlling function of evQTLs and determine the precise molecular processes through which evQTL SNPs affect the magnitude of gene expression variance. In the absence of such knowledge, the real impact of genetic variants on changing phenotypic variances will remain elusive, which will eventually

hinder the development of effective association mapping for pinpointing causal variants of complex traits.

In Chapter IV, we developed a multivariate method called aberrant gene expression analysis to study the effects of rare variants on gene expression by measuring levels of multigene expression dispersion. This method quantifies the dissimilarity in multigene expression patterns between individuals using Mahalanobis distance, which is an appropriate measure due to the consideration of the covariance between expression levels of multiple genes. Our results showed that, rare genetic variants of outlier individuals are enriched in the regulatory elements (enhancers, promoter regions elements) of corresponding aberrantly expressed genes, which suggests that rare variants may play a specific role in gene expression regulation.

REFERENCES

1. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-251.

2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

3. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.

4. International HapMap C (2005) A haplotype map of the human genome. Nature 437: 1299-1320.

5. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-U853.

6. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

7. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 363: 166-176.

8. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273: 1516-1517.

9. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385-389.

10. Chandra A, Mitry D, Wright A, Campbell H, Charteris DG (2014) Genome-wide association studies: applications and insights gained in Ophthalmology. Eye 28: 1066-1079.

11. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 42: D1001-1006.

12. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316: 889-894.

13. Feingold EA, Good PJ, Guyer MS, Kamholz S, Liefer L, et al. (2004) The ENCODE (ENCyclopedia of DNA elements) Project. Science 306: 636-640.

14. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, et al. (2010) The NIH roadmap epigenomics mapping consortium. Nature Biotechnology 28: 1045-1048.

15. Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. Current Opinion in Genetics & Development 19: 212-219.

16. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? American Journal of Human Genetics 69: 124-137.

17. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, et al. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature Genetics 43: 1066-U1050.

18. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, et al. (2012) A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nature Genetics 44: 1326-1329.

19. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, et al. (2012) A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488: 96-99.

20. Kong SW, Sahin M, Collins CD, Wertz MH, Campbell MG, et al. (2014) Divergent dysregulation of gene expression in murine models of fragile X syndrome and tuberous sclerosis. Molecular Autism 5: 16.

21. Luthi-Carter R, Hanson SA, Strand AD, Bergstrom DA, Chun WJ, et al. (2002) Dysregulation of gene expression in the R6/2 model of polyglutamine disease: parallel changes in muscle and brain. Human Molecular Genetics 11: 1911-1926.

22. Day DA, Tuite MF (1998) Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. Journal of Endocrinology 157: 361-371.

23. Pain VM (1996) Initiation of protein synthesis in eukaryotic cells. European Journal of Biochemistry 236: 747-771.

24. Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nature Reviews Genetics 9: 102-114.

25. Bremer K, Moyes CD (2014) mRNA degradation: an underestimated factor in steady-state transcript levels of cytochrome c oxidase subunits? Journal of Experimental Biology 217: 2212-2220.

26. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-1349.

27. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5: 621-628.

28. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502.

29. Nica AC, Dermitzakis ET (2013) Expression quantitative trait loci: present and future. Philosophical Transactions of the Royal Society B-Biological Sciences 368.

30. Haldane JBS (1932) The time of action of genes, and its bearing on some evolutionary problems. American Naturalist 66: 5-24.

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

32. Damerval C, Maurice A, Josse JM, de Vienne D (1994) Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics 137: 289-301.

33. Abraham I, Doane WW (1978) Genetic regulation of tissue-specific expression of amylase structural genes in Drosophila melanogaster. Proc Natl Acad Sci U S A 75: 4446-4450.

34. Powell JR (1979) Population genetics of Drosophila amylase. II. Geographic patterns in D. pseudoobscura. Genetics 92: 613-622.

35. Trevino V, Falciani F, Barrera-Saldana HA (2007) DNA microarrays: a powerful genomic tool for biomedical and clinical research. Mol Med 13: 527-541.

36. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends in Genetics 17: 388-391.

37. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296: 752-755.

38. West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, et al. (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics 175: 1441-1450.

39. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, et al. (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nature Genetics 35: 57-64.

40. Li Y, Lvarez OAA, Gutteling EW, Tijsterman M, Fu JJ, et al. (2006) Mapping determinants of gene expression plasticity by genetical genomics in *C-elegans*. Plos Genetics 2: 2155-2161.

41. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422: 297-302.

42. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, et al. (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nature Genetics 39: 226-231.

43. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, et al. (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315: 848-853.

44. Bahcall OG (2015) Huamn genetics GTEx pilot quantifies eQTL variation across tissues and individuals. Nature Reviews Genetics 16: 375-375.

45. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2015) A global reference for human genetic variation. Nature 526: 68-74.

46. Westra HJ, Arends D, Esko T, Peters MJ, Schurmann C, et al. (2015) Cell Specific eQTL Analysis without Sorting Cells. Plos Genetics 11: e1005223.

47. Schadt EE, Molony C, Chudin E, Hao K, Yang X, et al. (2008) Mapping the genetic architecture of gene expression in human liver. Plos Biology 6: 1020-1032.

48. Vinuela A, Snoek LB, Riksen JAG, Kammenga JE (2010) Genome-wide gene expression regulation as a function of genotype and age in *C-elegans*. Genome Research 20: 929-937.

49. Francesconi M, Lehner B (2014) The effects of genetic variation on gene expression dynamics during development. Nature 505: 208-211.

50. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6: 95-108.

51. Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. Nature Reviews Genetics 16: 197-212.

52. Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12: 111-139.

53. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-9445.

54. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci U S A 102: 1572-1577.

55. Broman KW, Wu H, Sen S, Churchill GA (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19: 889-890.

56. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81: 559-575.

57. Flutre T, Wen XQ, Pritchard J, Stephens M (2013) A Statistical framework for joint eQTL analysis in multiple tissues. Plos Genetics 9: e1003486.

58. Yang J, Loos RJF, Powell JE, Medland SE, Speliotes EK, et al. (2012) FTO genotype is associated with phenotypic variability of body mass index. Nature 490: 267-272.

59. Shen X, Pettersson M, Ronnegard L, Carlborg O (2012) Inheritance beyond plain heritability: variance-controlling genes in arabidopsis thaliana. Plos Genetics 8: e1002839.

60. Ayroles JF, Buchanan SM, O'Leary C, Skutt-Kakaria K, Grenier JK, et al. (2015) Behavioral idiosyncrasy reveals genetic control of phenotypic variability. Proc Natl Acad Sci U S A 112: 6706-6711.

61. Pare G, Cook NR, Ridker PM, Chasman DI (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the women's genome health study. Plos Genetics 6: e1000981.

62. Brown AA, Buil A, Vinuela A, Lappalainen T, Zheng HF, et al. (2014) Genetic interactions affecting human gene expression identified by variance association mapping. Elife 3.

63. Schultz BB (1985) Levene's test for relative variation. Syst Biol 34 (4): 449-456.

64. Forsythe MBBaAB (1974) The small sample behavior of some statistics which test the equality of several means. Technometrics 16: 129-132.

65. Ho JWK, Stefani M, dos Remedios CG, Charleston MA (2008) Differential variability analysis of gene expression and its application to human diseases. Bioinformatics 24: I390-I398.

66. Mar JC, Matigian NA, Mackay-Sim A, Mellick GD, Sue CM, et al. (2011) Variance of gene expression identifies altered network constraints in neurological disease. Plos Genetics 7: e1002207.

67. Xu ZY, Wei W, Gagneur J, Clauder-Munster S, Smolik M, et al. (2011) Antisense expression increases gene expression variability and locus interdependency. Molecular Systems Biology 7:468.

68. Hulse AM, Cai JJ (2013) Genetic variants contribute to gene expression variability in humans. Genetics 193: 95-108.

69. Ronnegard L, Valdar W (2011) Detecting major genetic loci controlling phenotypic variability in experimental crosses. Genetics 188: 435-447.

70. Ronnegard L, Valdar W (2012) Recent developments in statistical methods for detecting genetic loci affecting phenotypic variability. BMC Genet 13: 63.

71. Smyth GK (2002) An efficient algorithm for REML in heteroscedastic regression. Journal of Computational and Graphical Statistics 11: 836-847.

72. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436: 701-703.

73. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, et al. (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. Nat Genet 35: 57-64.

74. Becker J, Wendland JR, Haenisch B, Nothen MM, Schumacher J (2012) A systematic eQTL study of cis-trans epistasis in 210 HapMap individuals. European Journal of Human Genetics 20: 97-101.

75. Struchalin MV, Dehghan A, Witteman JCM, van Duijn C, Aulchenko YS (2010) Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. BMC Genetics 11:92.

76. Bateson W (1903) Mendel's principles of heredity in mice. Nature 68: 33-34.

77. Norton B, Pearson ES (1976) Note on the background to, and refereeing of, Fisher,R.A. 1918 paper on the correlation between relatives on the supposition of Mendelian inheritance. Notes and records of the Royal Society of London 31: 151-162.

78. Nelson MR, Wegmann D, Ehm MG, Kessner D, St Jean P, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science 337: 100-104.

79. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64-69.

80. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740-743.

81. Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

82. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics 11: 415-425.

83. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, et al. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics 12: 745-755.

84. Li BS, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. American Journal of Human Genetics 83: 311-321.

85. Li BS, Leal SM (2009) Discovery of rare variants via sequencing: implications for the design of complex trait association studies. Plos Genetics 5: e1000481.

86. Morgenthaler S, Thilly WG (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis 615: 28-56.

87. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. American Journal of Human Genetics 86: 832-838.

88. Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genetic Epidemiology 33: 497-507.

89. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, et al. (2011) Testing for an unusual distribution of rare variants. Plos Genetics 7: e1001322.

90. Wu MC, Lee S, Cai TX, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics 89: 82-93.

91. Lee S, Abecasis GR, Boehnke M, Lin XH (2014) Rare-variant association analysis: study designs and statistical tests. American Journal of Human Genetics 95: 5-23.

92. Derkach A, Lawless JF, Sun L (2013) Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genetic Epidemiology 37: 110-121.

93. Li X, Battle A, Karczewski KJ, Zappala Z, Knowles DA, et al. (2014) Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. Am J Hum Genet 95: 245-256.

94. Hallgrímsson, B., and B. K. Hall (2005) Variation. Academic Press, Amsterdam.

95. Wagner G (1995) Adaptation and the modular design of organisms. Advances in Artificial Life. pp. 315-328.

96. Wagner GP, Altenberg L (1996) Complex adaptations and the evolution of evolvability. Evolution 51: 967-976.

97. Wang Z, Zhang J (2011) Impact of gene expression noise on organismal fitness and the efficacy of natural selection. Proc Natl Acad Sci U S A 108: E67-76.

98. Bahar R, Hartmann CH, Rodriguez KA, Denny AD, Busuttil RA, et al. (2006) Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature 441: 1011-1014.

99. Kemkemer R, Schrank S, Vogel W, Gruler H, Kaufmann D (2002) Increased noise as an effect of haploinsufficiency of the tumor-suppressor gene neurofibromatosis type 1 in vitro. Proc Natl Acad Sci U S A 99: 13783-13788.

100. Zhang Z, Qian W, Zhang J (2009) Positive selection for elevated gene expression noise in yeast. Mol Syst Biol 5: 299.

101. Acar M, Mettetal JT, van Oudenaarden A (2008) Stochastic switching as a survival strategy in fluctuating environments. Nat Genet 40: 471-475.

102. Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. Nat Rev Genet 6: 451-464.

103. Hill WG, Zhang XS (2004) Effects on phenotypic variability of directional selection arising through genetic differences in residual variability. Genet Res 83: 121-132.

104. Montgomery SB, Dermitzakis ET (2011) From expression QTLs to personalized transcriptomics. Nat Rev Genet 12: 277-282.

105. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217-1224.

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

107. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78.

108. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, et al. (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464: 773-777.

109. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, et al. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464: 768-772.

110. Queitsch C, Sangster TA, Lindquist S (2002) Hsp90 as a capacitor of phenotypic variation. Nature 417: 618-624.

111. Jimenez-Gomez JM, Corwin JA, Joseph B, Maloof JN, Kliebenstein DJ (2011) Genomic analysis of QTLs and genes altering natural variation in stochastic noise. PLoS Genet 7: e1002295.

112. Shen X, Pettersson M, Ronnegard L, Carlborg O (2012) Inheritance beyond plain heritability: variance-controlling genes in *Arabidopsis thaliana*. PLoS Genet 8: e1002839.

113. Perry GML, Nehrke KW, Bushinsky DA, Reid R, Lewandowski KL, et al. (2012) Sex modifies genetic effects on residual variance in urinary calcium excretion in rat (*Rattus norvegicus*). Genetics 191: 1003-U1604.

114. Yang J, Loos RJ, Powell JE, Medland SE, Speliotes EK, et al. (2012) FTO genotype is associated with phenotypic variability of body mass index. Nature 490: 267-272.

115. Verbyla AP, Smyth GK (1998) Double generalized linear models: approximate residual maximum likelihood and diagnostics. University of Adelaide: Department of Statistics. 1-15 p.

116. Pare G, Cook NR, Ridker PM, Chasman DI (2010) On the use of variance per genotype as a tool to identify quantitative trait interaction effects: a report from the Women's Genome Health Study. PLoS Genet 6: e1000981.

117. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, et al. (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet 44: 1084-1089.

118. Moayyeri A, Hammond CJ, Hart DJ, Spector TD (2013) The UK adult twin registry (TwinsUK resource). Twin Res Hum Genet 16: 144-149.

119. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, et al. (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nature Genetics 44: 1084-1089.

120. Moayyeri A, Hammond CJ, Hart DJ, Spector TD (2013) The UK adult twin registry (TwinsUK resource). Twin Research and Human Genetics 16: 144-149.

121. Spector TD, Williams FMK (2006) The UK adult twin registry (TwinsUK). Twin Research and Human Genetics 9: 899-906.

122. Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, et al. (2007) A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics 23: 2741-2746.

123. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The Architecture of gene regulatory variation across multiple human tissues: The MuTHER Study. Plos Genetics 7.

124. Ramasamy A, Trabzuni D, Gibbs JR, Dillman A, Hernandez DG, et al. (2013) Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies. Nucleic Acids Res 41: e88.

125. The-1000-Genomes-Project-Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

126. Maheshri N, O'Shea EK (2007) Living with noisy genes: how cells function reliably with inherent variability in gene expression. Annu Rev Biophys Biomol Struct 36: 413-434.

127. Ansel J, Bottin H, Rodriguez-Beltran C, Damon C, Nagarajan M, et al. (2008) Cell-to-cell stochastic variation in gene expression is a complex genetic trait. PLoS Genet 4: e1000049.

128. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263-265.

129. Livak KJ, Wills QF, Tipping AJ, Datta K, Mittal R, et al. (2013) Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. Methods 59: 71-79.

130. Pai AA, Cain CE, Mizrahi-Man O, De Leon S, Lewellen N, et al. (2012) The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. PLoS Genet 8: e1003000.

131. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7: 562-578.

132. Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7: 1009-1015.

133. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, et al. (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet 7: e1002003.

134. Andrew T, Hart DJ, Snieder H, de Lange M, Spector TD, et al. (2001) Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. Twin Res 4: 464-477.

135. Fligner MA, Killeen TJ (1976) Distribution-Free 2-Sample Tests for Scale. Journal of the American Statistical Association 71: 210-213.

136. Fraser HB, Schadt EE (2010) The quantitative genetics of phenotypic robustness. PLoS One 5: e8635.

137. Struchalin MV, Dehghan A, Witteman JC, van Duijn C, Aulchenko YS (2010) Variance heterogeneity analysis for detection of potentially interacting genetic loci: method and its limitations. BMC Genet 11: 92.

138. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet 6: e1000888.

139. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, et al. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet 6: e1000895.

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

141. Patsopoulos NA, de Bakker PIW, Working BPMG, Evaluating SCS, Antagonist SCC, et al. (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Annals of Neurology 70: 897-912.

142. Evans DM, Spencer CCA, Pointon JJ, Su Z, Harvey D, et al. (2011) Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. Nature Genetics 43: 761-U767.

143. Okada Y, Sim X, Go MJ, Wu JY, Gu DF, et al. (2012) Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. Nature Genetics 44: 904-909.

144. Stein JL, Hibar DP, Madsen SK, Khamis M, McMahon KL, et al. (2011) Discovery and replication of dopamine-related gene effects on caudate volume in young and elderly populations (N=1198) using genome-wide search. Molecular Psychiatry 16: 927-937.

145. Eriksson N, Tung JY, Kiefer AK, Hinds DA, Francke U, et al. (2012) Novel associations for hypothyroidism include known autoimmune risk loci. Plos One 7: e34442.

146. Wolc A, White IM, Avendano S, Hill WG (2009) Genetic variability in residual variation of body weight and conformation scores in broiler chickens. Poult Sci 88: 1156-1161.

147. Hill WG, Mulder HA (2010) Genetic analysis of environmental variation. Genet Res (Camb) 92: 381-395.

148. Gibson G (2009) Decanalization and the origin of complex disease. Nat Rev Genet 10: 134-140.

149. Williams RB, Chan EK, Cowley MJ, Little PF (2007) The influence of genetic variation on gene expression. Genome Res 17: 1707-1716.

150. Horvitz HR, Sulston JE (1980) Isolation and genetic characterization of cell-lineage mutants of the nematode *Caenorhabditis elegans*. Genetics 96: 435-454.

151. Gartner K (1990) A third component causing random variability beside environment and genotype. A reason for the limited success of a 30 year long effort to standardize laboratory animals. Lab Anim 24: 71-77.

152. Baranzini SE, Mudge J, van Velkinburgh JC, Khankhanian P, Khrebtukova I, et al. (2010) Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. Nature 464: 1351-1356.

153. Badano JL, Katsanis N (2002) Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet 3: 779-789.

154. Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4: e1000008.

155. Powell JE, Henders AK, McRae AF, Kim J, Hemani G, et al. (2013) Congruence of additive and non-additive effects on gene expression estimated from pedigree and SNP data. PLoS Genet 9: e1003502.

156. Qin S, Kim J, Arafat D, Gibson G (2012) Effect of normalization on statistical and biological interpretation of gene expression profiles. Front Genet 3: 160.

157. Geiler-Samerotte K, Bauer C, Li S, Ziv N, Gresham D, et al. (2013) The details in the distributions: why and how to study phenotypic variability. Curr Opin Biotechnol 24: 752-759.

158. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, et al. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24: 127-135.

159. Crow JF (2010) On epistasis: why it is unimportant in polygenic directional selection. Philos Trans R Soc Lond B Biol Sci 365: 1241-1244.

160. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? Nat Rev Genet 5: 618-625.

161. Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109: 1193-1198.

162. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

163. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747-753.

164. Ueki M, Cordell HJ (2012) Improved statistics for genome-wide interaction analysis. PLoS Genet 8: e1002625.

165. Moore JH, Williams SM (2009) Epistasis and its implications for personal genetics. Am J Hum Genet 85: 309-320.

166. Daye ZJ, Chen J, Li H (2012) High-dimensional heteroscedastic regression with an application to eQTL data analysis. Biometrics 68: 316-326.

167. Struchalin MV, Amin N, Eilers PH, van Duijn CM, Aulchenko YS (2012) An R package "VariABEL" for genome-wide searching of potentially interacting loci by testing genotypic variance heterogeneity. BMC Genet 13: 4.

168. Shang J, Zhang J, Sun Y, Liu D, Ye D, et al. (2011) Performance analysis of novel methods for detecting epistasis. BMC Bioinformatics 12: 475.

169. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11: 2463-2468.

170. Phillips PC (1998) The language of gene interaction. Genetics 149: 1167-1171.

171. Wang X, Elston RC, Zhu X (2010) The meaning of interaction. Hum Hered 70: 269-277.

172. Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sunderland, Mass.: Sinauer. xvi, 980 p. p.

173. Metzger BP, Yuan DC, Gruber JD, Duveau F, Wittkopp PJ (2015) Selection on noise constrains variation in a eukaryotic promoter. Nature.

174. Ayroles JF, Buchanan SM, O'Leary C, Skutt-Kakaria K, Grenier JK, et al. (2015) Behavioral idiosyncrasy reveals genetic control of phenotypic variability. Proc Natl Acad Sci U S A 112: 6706-6711.

175. Brown AA, Buil A, Vinuela A, Lappalainen T, Zheng HF, et al. (2014) Genetic interactions affecting human gene expression identified by variance association mapping. Elife 3: e01381.

176. Wang G, Yang E, Brinkmeyer-Langford CL, Cai JJ (2014) Additive, epistatic, and environmental effects through the lens of expression variability QTL in a twin cohort. Genetics 196: 413-425.

177. Geiler-Samerotte KA, Bauer CR, Li S, Ziv N, Gresham D, et al. (2013) The details in the distributions: why and how to study phenotypic variability. Curr Opin Biotechnol 24: 752-759.

178. Gibson G, Wagner G (2000) Canalization in evolutionary genetics: a stabilizing theory? Bioessays 22: 372-380.

179. Wolf L, Silander OK, van Nimwegen E (2015) Expression noise facilitates the evolution of gene regulation. Elife 4.

180. Mar JC, Matigian NA, Mackay-Sim A, Mellick GD, Sue CM, et al. (2011) Variance of gene expression identifies altered network constraints in neurological disease. PLoS Genet 7: e1002207.

181. Ho JW, Stefani M, dos Remedios CG, Charleston MA (2008) Differential variability analysis of gene expression and its application to human diseases. Bioinformatics 24: i390-398.

182. Campbell MG, Kohane IS, Kong SW (2013) Pathway-based outlier method reveals heterogeneous genomic structure of autism in blood transcriptome. BMC Med Genomics 6: 34.

183. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501: 506-511.

184. Stegle O, Parts L, Durbin R, Winn J (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol 6: e1000770.

185. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56-65.

186. Eisenberg E, Levanon EY (2013) Human housekeeping genes, revisited. Trends in Genetics 29: 569-574.

187. Vindelov LL, Christensen IJ (1990) A review of techniques and results obtained in one laboratory by an integrated system of methods designed for routine clinical flow cytometric DNA analysis. Cytometry 11: 753-770.

188. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57: 289-300.

189. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science 297: 1183-1186.

190. Kanu N, Behrens A (2008) ATMINistrating ATM signalling: regulation of ATM by ATMIN. Cell Cycle 7: 3483-3486.

191. Thorvaldsdottir H, Robinson JT, Mesirov JP (2013) Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14: 178-192.

192. Hemani G, Shakhbazov K, Westra HJ, Esko T, Henders AK, et al. (2014) Detection and replication of epistasis influencing transcription in humans. Nature 508: 249-253.

193. Wood AR, Tuke MA, Nalls MA, Hernandez DG, Bandinelli S, et al. (2014) Another explanation for apparent epistasis. Nature 514: E3-5.

194. Dey SS, Foley JE, Limsirichai P, Schaffer DV, Arkin AP (2015) Orthogonal control of expression mean and variance by epigenetic features at different genomic loci. Mol Syst Biol 11: 806.

195. Bar-Peled L, Chantranupong L, Cherniack AD, Chen WW, Ottina KA, et al. (2013) A tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. Science 340: 1100-1106.

196. Wiles AM, Doderer M, Ruan J, Gu TT, Ravi D, et al. (2010) Building and analyzing protein interactome networks by cross-species comparisons. BMC Syst Biol 4: 36.

197. Kampinga HH (2015) Molecular biology: It takes two to untangle. Nature 524: 169-170.

198. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957-968.

199. Jurado S, Conlan LA, Baker EK, Ng JL, Tenis N, et al. (2012) ATM substrate Chk2-interacting Zn2+ finger (ASCIZ) Is a bi-functional transcriptional activator and feedback sensor in the regulation of dynein light chain (DYNLL1) expression. J Biol Chem 287: 3156-3164.

200. Albert FW, Kruglyak L (2015) The role of regulatory variation in complex traits and disease. Nat Rev Genet 16: 197-212.

201. McGrath JJ, Hannan AJ, Gibson G (2011) Decanalization, brain development and risk of schizophrenia. Transl Psychiatry 1: e14.

202. Burrows EL, Hannan AJ (2013) Decanalization mediating gene-environment interactions in schizophrenia and other psychiatric disorders with neurodevelopmental etiology. Front Behav Neurosci 7: 157.

203. Ecker S, Pancaldi V, Rico D, Valencia A (2015) Higher gene expression variability in the more aggressive subtype of chronic lymphocytic leukemia. Genome Med 7: 8.

204. Zeng Y, Wang G, Yang E, Ji G, Brinkmeyer-Langford CL, et al. (2015) Aberrant gene expression in humans. PLoS Genet 11: e1004942.

205. Kilpinen H, Barrett JC (2013) How next-generation sequencing is transforming complex disease genetics. Trends Genet 29: 23-30.

206. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11: 415-425.

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

208. Zani S, Riani M, Corbellini A (1998) Robust bivariate boxplots and multiple outlier detection. Computational Statistics & Data Analysis 28: 257-270.

209. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545-15550.

210. Rousseeuw PJ (1984) Least Median of Squares Regression. Journal of the American Statistical Association 79: 871-880.

211. Verboven S, Hubert M (2005) LIBRA: a MATLAB library for robust analysis. Chemometrics and Intelligent Laboratory Systems 75: 127-136.

212. Garrett RG (1989) The Chi-square Plot - a tool for multivariate outlier recognition. Journal of Geochemical Exploration 32: 319-341.

213. Rousseeuw PJ, Vanzomeren BC (1990) Unmasking multivariate outliers and leverage Points. Journal of the American Statistical Association 85: 633-639.

214. Filzmoser P, Maronna R, Werner M (2008) Outlier identification in high dimensions. Computational Statistics & Data Analysis 52: 1694-1711.

215. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, et al. (2014) Heritability and genomics of gene expression in peripheral blood. Nat Genet 46, 430–437.

216. Yang S, Liu Y, Jiang N, Chen J, Leach L, et al. (2014) Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals. BMC Genomics 15: 13.

217. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, et al. (2014) From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res 24: 496-510.

218. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, et al. (2013) Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Research 41: 827-841.

219. P. C. Mahalanobis. (1936) On the generalised distance in statistics; Journ. Asiat. Soc. Bengal, 26: 49-55.

220. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, et al. (2012) Patterns of cis regulatory variation in diverse human populations. PLoS Genet 8: e1002639.

221. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3: 1724-1735.

222. GTEx_Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45: 580-585.

223. Wallace TA, Prueitt RL, Yi M, Howe TM, Gillespie JW, et al. (2008) Tumor immunobiological differences in prostate cancer between African-American and European-American men. Cancer Res 68: 927-936.

224. Montgomery SB, Lappalainen T, Gutierrez-Arcelus M, Dermitzakis ET (2011) Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet 7: e1002144.

225. Jayasinghe BS, Volz DC (2012) Aberrant ligand-induced activation of G protein-coupled estrogen receptor 1 (GPER) results in developmental malformations during vertebrate embryogenesis. Toxicol Sci 125: 262-273.

226. Lacroix A, Bourdeau I, Lampron A, Mazzuco TL, Tremblay J, et al. (2010) Aberrant G-protein coupled receptor expression in relation to adrenocortical overfunction. Clin Endocrinol (Oxf) 73: 1-15.

227. Spiegel AM, Weinstein LS (2004) Inherited diseases involving g proteins and g protein-coupled receptors. Annu Rev Med 55: 27-39.

228. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era - concepts and misconceptions. Nature Reviews Genetics 9: 255-266.

229. Price AL, Helgason A, Thorleifsson G, McCarroll SA, Kong A, et al. (2011) Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet 7: e1001317.

230. Zaitlen N, Kraft P (2012) Heritability in the genome-wide association era. Hum Genet 131: 1655-1664.