**Title**

Identifying Dominant Genetic Associations with Gene Expression in the Human Genome

**Author**

GU, JING

**Publication Date**

2017

UNIVERSITY OF CALIFORNIA, SAN DIEGO

# Identifying Dominant Genetic Associations with Gene Expression in the Human Genome

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Chemistry

by

Jing Gu

Committee in charge:

Wei Wang, Chair
Jelena Bradic
Graham McVicker
Brian Zid

2017

The Thesis of Jing Gu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2017

# DEDICATION

This Thesis is dedicated to my parents and my grandparents for their continued support and love throughout my life.

# EPIGRAPH

The Lord is at hand; do not be anxious about anything, but in everything by prayer and supplication with thanksgiving let your requests be made known to God. And the peace of God, which surpasses all understanding, will guard your hearts and your minds in Christ Jesus.

PHILIPPIANS 4:6-7

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENT

First, I would like to thank my MS thesis advisor, Dr. Graham McVicker, for giving me this great opportunity to work and learn in his lab at Salk Institute. I feel so grateful that Dr. McVicker was willing to personally train me, even though I had minimal background in bioinformatics or quantitative genetics. His passion in science and teaching always encourages me to keep learning and pursuing science. Without his guidance and encouragement, I could not imagine being able to finish this thesis.

I would like to thank my committee chair and thesis advisor, Dr. Wei Wang, for his continued support throughout my graduate study. Dr. Wang provided me helpful advice on both my thesis projects and future careers.

I would like to thank my other committee members - Dr. Brain Zid and Dr. Jelena Bradic for joining my committee. They are highly respectful for their dedication in research and great support to students.

I would like to thank two postdocs in the McVicker Lab, Dr. Hsiuyi Chen and Dr. Arko Sen, for their great suggestions and guidance on my project. I was always encouraged and inspired by Hsiuyi for her enthusiasm in discussing about projects or science in general or anything fun with other people. I truly enjoyed working and learning from her. I also would like to express my admiration to Arko for his expertise in programming and broad knowledge in biology.

I would like to thank other colleagues in the McVicker Lab, including Patrick Fiaux (PhD. student at UCSD), Ishika Iluthra (an exchange student), and Sélène Tyndale (our lab manager) for their support in science and life. Patrick always gave me useful advice on programming and suggestions on my project. Sélène impressed me with her passion in teaching science and doing research. I liked discussing problems with her and asking for her advice. Ishika was so supportive that she encouraged me when I felt down with research. Because of their company and support, I enjoyed my time working at McVicker's lab.

I would like to thank my undergraduate advisor Dr. Thomas Hermann and mentor Dr. Mark Boerneke. They guided me to have the first experience of doing research. I was encouraged to continue pursuing science and exploring my interests.

I would like to thank my parents and grandparents for their self-less love and encouragement. They have always been open-minded and let me choose my own careers.

Lastly, I would like to thank God for his blessings throughout my life and everything he brought to my life.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from dbGaP with version GTEx.v6.p1.c1 on March, 2017.

# ABSTRACT OF THE THESIS

Identifying Dominant Genetic Associations with Gene Expression in the Human Genomes

by

Jing Gu

Master of Science in Chemistry

University of California, San Diego, 2017

Professor Wei Wang, Chair

When mapping expression quantitative trait loci, a linear additive genetic model is mostly commonly used to investigate how genetic variants influence transcript levels. This model assumes that the phenotype of heterozygotes is halfway between that of the low-homozygous and high-homozygous genotypes and may miss non-additive relationships, such as those caused by dominant alleles. Here we examine RNA-Seq data to identify dominant genetic associations with gene expression in the human genome. We applied a multiple linear regression model on genotypes and RNA-Seq data from Genotype-Tissue Expression project. With stringent permutations, we discovered that on average, 0.19% of all genes tested (including non-coding RNAs and pseudogenes) show evidence for dominant genetic associations across ten different tissues. Most dominant effect sizes are positive, implying that the phenotypes of heterozygotes tend to have similar gene expression levels to high-expression homozygotes. In 8 out of the 10 tissues we examined, we found that genes encoding major histocompatibility complex (MHC) proteins are enriched for dominant effects.

# I. Introduction

## A. SNPs and quantitative traits

Despite of the diverse human traits, human genomes in average share 99.9% similarity. The variations in a single nucleotide of DNA sequences among individuals are called single nucleotide variants (SNVs), and common SNVs with a minor allele frequency of at least 1% are defined as single nucleotide polymorphisms (SNPs). While most SNPs do not affect human traits, a subset of SNPs, such as those located in regulatory regions, contribute to disease susceptibility (Emilsson et al. 2008; Nica et al. 2010).

Many human traits are complex and polygenic, which require large sample sizes to identify significant genetic associations. With the advent of array and sequencing technologies, the whole human genome can be sequenced in a much faster and economic way, which makes it possible to examine variants genome-wide to identify associations with traits of interest.

Genome-wide association studies (GWAS) have successfully discovered common genetic variation that affects traits. However, the vast majority of GWAS hits are outside of genes and are in non-coding regions of the genome (Hindorff et al. 2009). It remains challenging to recognize the precise target genes regulated by the variants identified by GWAS and the tissues where regulations occur. Gene expression quantitative trait loci give a way to link genetic variation to the genes the variants regulate and may help with interpretation of GWAS hits.

### B. Gene Expression Quantitative Trait Loci (eQTLs)

Gene expression can be treated as a quantitative trait, and it is therefore possible to identify SNPs that are associated with its gene expression levels. A SNP that is associated with the expression of a gene is known as a gene expression quantitative trait locus (eQTL). In the first eQTL study Brem et al. crossed two strains of yeast to characterize the SNPs that were linked to transcript levels quantified via microarray (Brem et al. 2002). To test if a locus is an eQTL, both the genotypes at the locus as well as the transcript levels for the gene of interest are required. When mapping eQTLs, the underlying assumption is that two alleles are expressed independently and their expression values are combined linearly to predict the expression values of the heterozygotes (Figure 1.1).

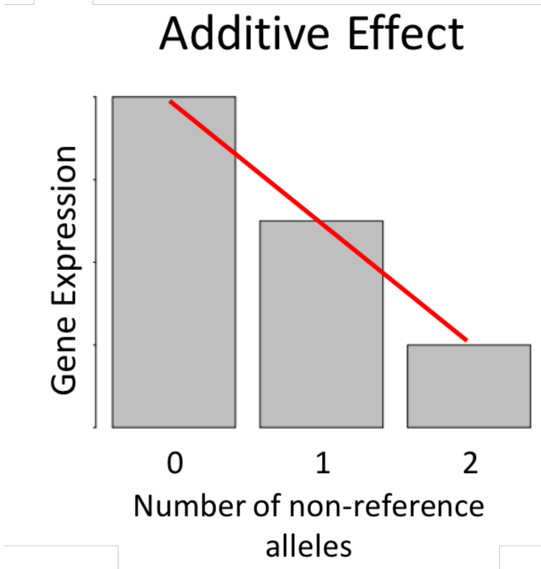


**Figure 1.1 An example of an eQTL with a linear additive association.** Schematic representation of an additive association between the genotype of a SNP and gene expression levels.

A simple linear regression model can be applied to test for a linear association between genotype of SNPs and transcript abundance. A SNP is said to be associated with the expression of a gene when the genotype effect size in the linear model is significantly different from zero. Transcript levels for the gene of interest can be tested for association with SNPs that are local, distal, or across the chromosome.

### C. Dominant genetic association

A dominant association between genotypes and gene expression means that the gene expression levels of heterozygotes are significantly different from the mean expression value of reference and non-reference homozygotes (Figure 1.2).

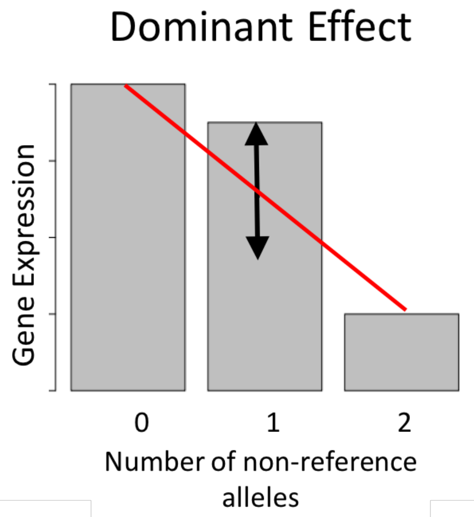
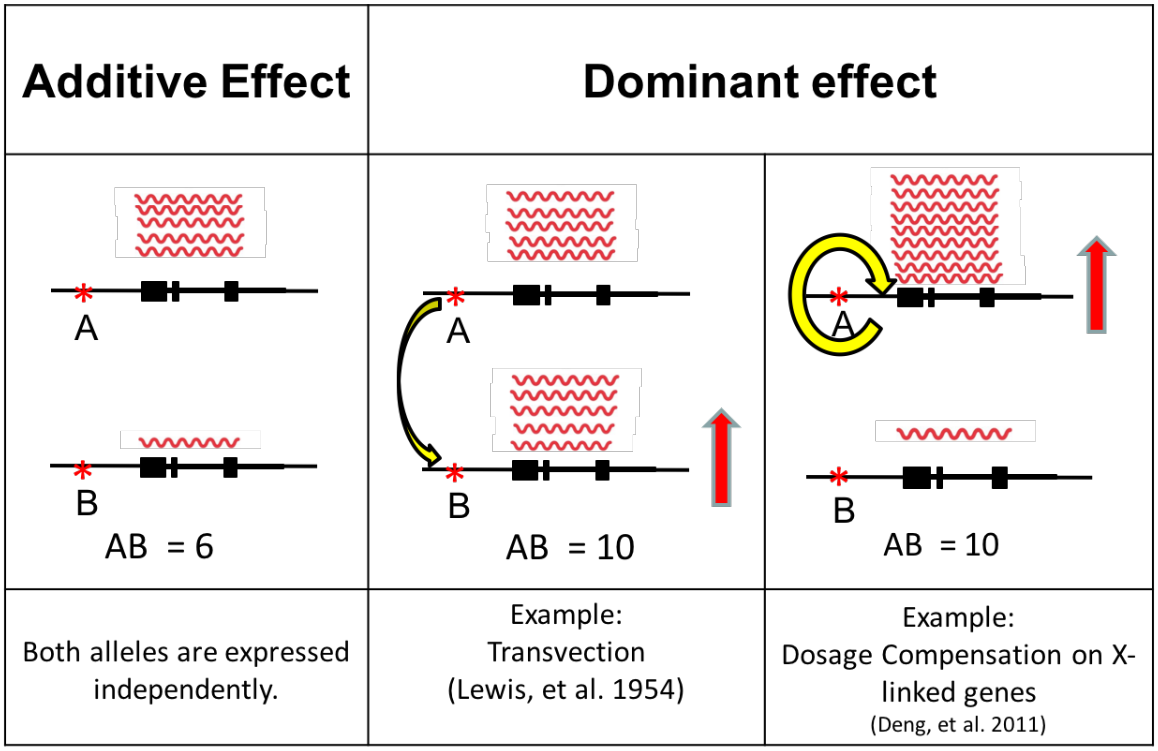


**Figure 1.2 An example of eQTL with a dominant association.** Schematic representation of gene expression levels against number of non-reference alleles. The black arrows indicate that in dominant associations heterozygous individuals have higher or lower expression than expected.

As an example suppose allele A produces 5 transcripts and allele B produces 1 transcript (Figure 1.3). When both alleles are independently expressed, the heterozygote genotype, AB, should generate 6 transcripts. However, under a dominant model, the expression of the heterozygote AB might be much higher (e.g. 10), which is close to individuals that are homozygous for high-expression alleles. One possible mechanism for this dominant effect is interallelic interaction, which causes the low-expression allele B to increase its expression of RNA transcripts. Lewis proposed this mechanism as “transvection” in 1954, when he observed intra-allelic complementation in *Drosophila* *melanogaster* (Lewis 1954). A second possible mechanism could be that the high-expression allele A upregulates itself to compensate for the low expression of allele B. There has been evidence showing that X-linked genes are upregulated in mammals, C. elegans and Drosophila, using both microarray and RNA-Seq data (Deng et al. 2011; Nguyen and Disteche 2006; Adler et al. 1997).



Both alleles are expressed independently.

Example:
Transvection
(Lewis, et al. 1954)

Example:
Dosage Compensation on X-linked genes
(Deng, et al. 2011)

**Figure 1.3 A schematic representation for theoretical mechanisms to explain dominant effects.** The red curves represent gene transcripts. Yellow arrows show the interaction that potentially causes dominant associations. Red arrows indicate the change in transcript abundance.

Several studies have shown dominant patterns in transcript levels across multiple model organisms. As shown in Table 1.1, there are considerable differences in the percentages of genes that show dominant effects in different organisms. These differences may be biological, but could also potentially reflect different methodologies and definitions of dominance between stidues.

We hypothesize that there also exist dominant genetic associations with transcript abundance in the human genome. A previous study on human samples identified 208 eQTLs (~1% of the genes tested) with dominant effects on gene expression levels, by quantifying whole blood gene expression using microarrays (Powell et al. 2013). Compared to microarrays, RNA-Seq is known to have a broader dynamic range and higher sensitivity and specificity to detect rare transcripts. Here we use RNA-Seq data to generate gene expression profiles and develop a statistical model to identify gene expression quantitative trait loci with dominant effects (dominant eQTLs) in the human genome. This may help us better understand how human genetic variantion influences the transcriptome.

**Table 1.1 A summary of studies that detect genes showing dominant effects across multiple organisms**

| **Paper** | **Organisms** | **Percentage of genes showing dominant effects** |
|---|---|---|
| Gibson et al. 2004 | *Drosophila melanogaster* | ~ 40% |
| Vuylsteke et al. 2005 | *Arabidopsis thaliana* | 1% - 40% |
| Cui et al. 2006 | Mice | <1% |
| Stupar et al. 2007 | Maize | ~ 10% |
| Powell et al. 2013 | Human | ~ 1% |

# II. Statistical model and computational approaches

## A. Statistical Model

We developed a multiple linear regression model to detect dominant associations between SNPs and gene expression levels.

$$\text{Gene Expression (E)} = \beta_0 + \beta_1 G_A + \beta_2 G_D + \varepsilon \tag{1}$$

$$\varepsilon \sim N(0, \sigma^2)$$

Here, $G_A$ stands for the number of non-reference alleles. For a bi-allelic SNP, $G_A = 0$ if genotype is reference homozygous; $G_A = 1$ if genotype is heterozygous; and $G_A = 2$ if genotype is non-reference homozygous. To allow for dominant effects, we introduce an additional variable $G_D$, where $G_D = 1$ if genotype is heterozygous; and $G_D = 0$ if the genotype is either reference or non-reference homozygous. "E" denotes the observed gene expression levels. Our model assumes that the noise across samples to is normally distributed. To apply our model, we first transformed the gene expression levels to be a standard normal distribution.

Our alternative hypothesis is there is a dominant genetic association such that the effect size of $G_D$ ($\beta_2$) is not equal to zero. On the other hand, the null hypothesis is that there is no evidence for a dominant association and $\beta_2$ is not significantly different from zero. Under the null hypothesis, SNPs can have either no association ($\beta_1$=0) or an additive association with gene expression levels ($\beta_1 \neq 0$). An ANOVA model can be used instead of a multiple linear regression to detect both additive and dominant effects of the genotype; however, the F-test used in the ANOVA model aims to test if at least one beta is significantly different from zero. This does not fulfill our objective, as we specifically wish to test whether there is evidence showing that $\beta_2$ is not equal to zero. Alternatively, a t-test can be used to test the effect sizes of $G_A$ and $G_D$ separately.

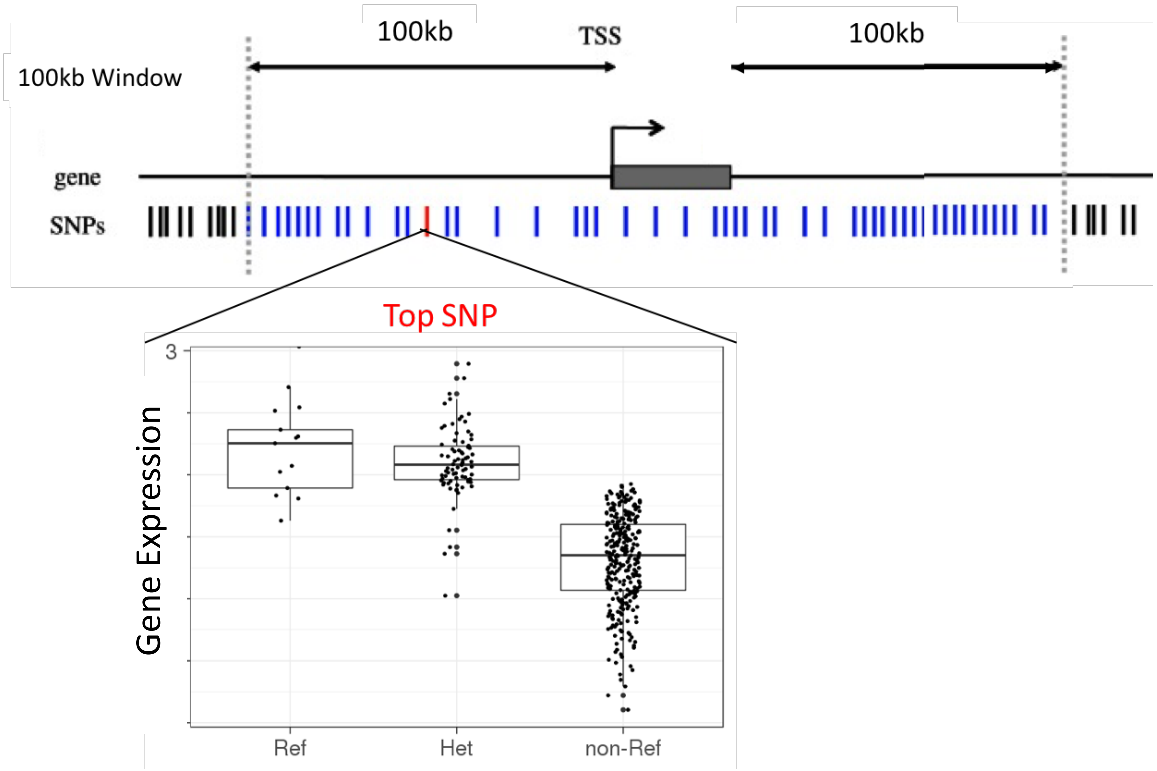


**Figure 2.1 An illustration of eQTL analysis.** An example of a cis-eQTL where SNPs in a 100 kb window from the start and end of a gene are tested for association with the gene's expression levels. The red bar represents the top SNP with the strongest association. The boxplot shows the distribution of expression values separately for each genotype group (Figure adapted from Nica and Dermitzakis 2013).

We chose a window size of 100kb to test if SNPs nearby genes have any dominant associations with gene expression levels (Figure 2.1). It has been shown that most eQTLs with strong additive associations are relatively close (<100kb) to the genes they are associated with (Yvert et al. 2003; Brem et al. 2002; Morley et al. 2004). To save computational time and avoid the complexity of distal interactions, we focused on SNPs that are nearby genes. Furthermore, we limited ourselves to common SNPs (≥5% minor allele frequency) that are bi-allelic and located on the autosomal chromosomes. We did not use SNPs on the X or Y chromosomes because it can be difficult to interpret results from the sex chromosome due to the effects of imprinting and X inactivation.

### B. Highly correlated SNPs cause hypothesis tests to be non-independent.

For each gene, there could be hundreds to thousands of SNPs being tested for dominant associations. Without correction, the family-wise error rate would be much bigger than the preset significance threshold of $\alpha$= 0.05. Our approach to control for the false positive rate ($\alpha$) is to calculate the false discovery rate (FDR) using the step-up method of Benjamini & Hochberg (1995). FDR is defined as the expected proportion of erroneous rejections among all the rejections (Benjamini, Y., and Hochberg, Y. 1995). The underlying assumption for this method is that tests are independent to each other, however, SNPs that are near to each other in the human genome have highly correlated genotypes due to linkage disequilibrium (International HapMap Consortium 2005). The non-independence of the SNPs means that correcting for false positive rates is very complicated. As an alternative approach, we decided to correct the false positive rate only for the top SNP with the lowest P-value (strongest association) for each gene. It is convenient to only have one SNP to work with per gene, even though a limitation is that we are assuming each gene has either 0 or 1 associations, when in fact some genes may have more than one association.

### C. Permutation scheme for obtaining the empirical null distribution

By definition, p-values of independent tests under the null should be uniformly distributed. However, the distribution of the lowest p-values chosen for each gene are skewed toward much smaller values even under the null. Therefore, obtaining the empirical null distribution for the lowest p-values after each permutation is required to adjust the observed lowest P-value for each gene. Typically, direct permutations are used to approximate the null distribution. With R permutations, an adjusted P-value cannot be smaller than 1/(R+1), which indicates a large number of permutations are necessary to achieve a low P-value. For instance, a million permutations are required to obtain a P-value of around $10^{-6}$. Therefore, direct permutation scheme for genome-wide analysis is very computationally intensive to reach low P-values.

We adopted a more efficient permutation scheme, which uses beta distribution to approximate the smallest P-values obtained through permutations (Ongen et al. 2016). It has been shown that order statistics of independent identically distributed random variables form a beta distribution (Jones et al. 2009). Based on the assumption that ranked $k^{th}$ P-values from each round of permutation are also beta distributed, Ongen found that the lowest P-values obtained from L tests form a beta distribution with shape parameters k and n:

$$U \sim \mathrm{Beta}\ (k, n) \qquad (2)$$

Both shape parameters can be estimated by maximizing the log-likelihood given a null set of P-values ($\{p_1, p_2, p_3, p_4, ...p_n\}$) obtained from permutations. Then an adjusted P-value $P_b$ can be computed as

$$p_b = P\ (U \leq p_n) \qquad (3)$$

This approach is efficient because approximating the tail of null distribution instead of directly sampling from it requires far fewer permutations to achieve the desired P-value. Ongen et al. found that 500 permutations allowed them to obtain accurate p-values by this method. Therefore, we applied their permutation scheme with 1000 permutations to obtain P-values that are well-calibrated under the null hypothesis.

### D. Speed up dominant eQTL analysis via matrix multiplication

Even using the above scheme to reduce the number of required permutations, our analysis pipeline took around 6 days to run on a single GTEx tissue using a compute cluster (196 cores). To examine multiple tissues, we need to improve the computation time required for genome-wide analysis. In particular, we performed multiple linear regression using the lm function from the R stats package and we wondered if it was possible to speed up this aspect of the pipeline. Matrix eQTL takes advantage of R's implementation of large matrix operations to achieve faster speed (Shabalin 2012), and we adapted their approach to increase the speed of our calculations. Instead of performing multiple linear regression for each permuted phenotype and repeating 1000 times, we first generated a matrix of 1000 permuted gene expression levels and then applied matrix multiplication between the gene expression matrix and the genotype matrix to obtain correlation coefficients. We first standardized the genotype and expression variables such that they have zero mean and unit sum of squares. Then the calculation of sample correlations can be simplified as the inner product between each permuted gene expression vector and genotype vector via

$$\text{cor(e, g)} = \frac{\Sigma(e-\bar{e})(g-\bar{g})}{\sqrt{\Sigma(e-\bar{e})^2\Sigma(g-\bar{g})^2}} = \Sigma eg \ = < \hat{e}, \hat{g} > \qquad (4)$$

Therefore, one large matrix multiplication generates all of the gene-SNP correlations, which are further used to compute t-statistics (Figure 2.2). By adapting the matrix multiplication algorithms to our own model, we are able to shorten the step of linear regression from 6 days to less than 3 hours for the whole genome. The algorithm for the analysis of each gene is as follows:

(1) Permute the gene expression values 1000 times to obtain a gene expression matrix E

(2) Center variables e, $g_A$, $g_D$ by subtracting their means to remove the intercept $\beta_0$

(3) Orthogonalize $g_D$ with respect to $g_A$ to remove the effect of $g_D$ that can be explained by $g_A$

$$\widetilde{g_D} = g_A - \frac{<g_A, g_D>}{<g_A, g_A>} g_A \qquad (5)$$

(4) Standardize e, $g_A$, $\widetilde{g_D}$

(5) Compute test statistics $r_1^2$, $r_2^2$, $R^2 = r_1^2 + r_2^2 = <g_A, e>^2 + <\widetilde{g_D}, e>^2$

(6) Calculate the T-statistic $T_1 = \frac{r1\sqrt{n-k-1}}{\sqrt{1-R^2}}$ for testing $\beta_1$ and $T_2 = \frac{r2\sqrt{n-k-1}}{\sqrt{1-R^2}}$for testing $\beta_2$, where k = 2 for the number of variables tested ($g_A$, $\widetilde{g_D}$).

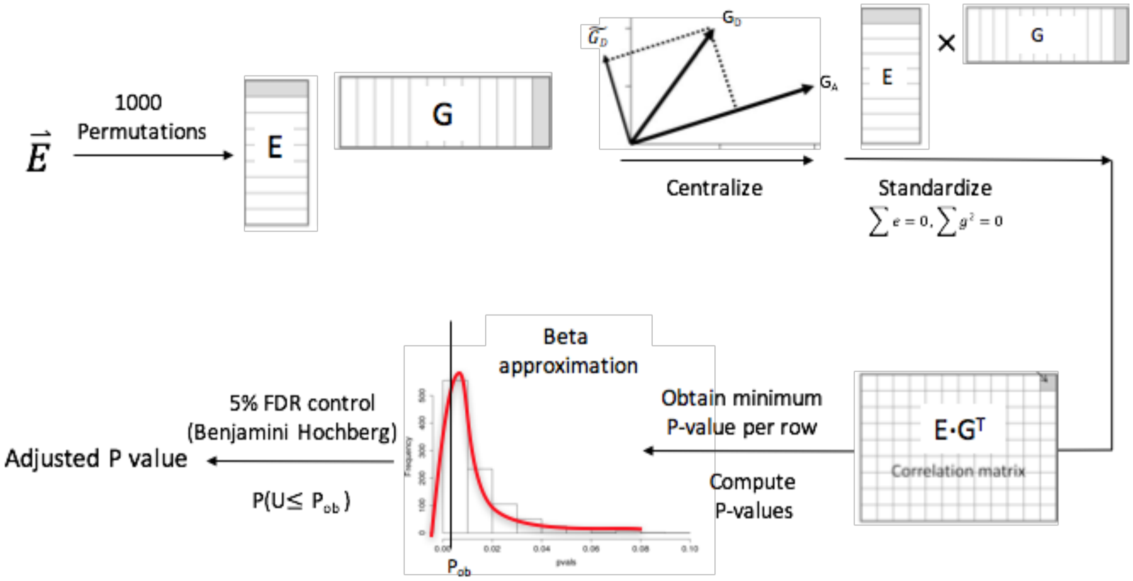


**Figure 2.2 The workflow of using matrix multiplication and beta approximation as permutation scheme to speed up the running time.** Figure adapted from Shabalin et al. 2012.

# III. Application to the GTEx data set

## A. The GTEx data set

The Genotype Tissue Expression dataset (version: v6.p1.c1) was downloaded from dbGAP. Our analysis was focused on 10 tissue types: adipose (subcutaneous), tibial artery, lung, muscle (skeletal), tibial nerve, skin (sun exposed), thyroid, esophagus (mucosa), cells transformed fibroblasts, and whole blood. The sample size for each of these tissues is at least 280, which maximizes our statistical power to detect dominant associations. The GTEx Consortium genotyped all samples using their blood-derived DNA and performed 76-base pair paired-end mRNA sequencing on RNA extracted from each tissue (GTEx Consortium. 2015). GTEx performed data preprocessing involving RNA-seq alignment and genotyping, which is described in the supplementary materials of their paper (GTEx Consortium. 2015).

## B. Dominant eQTL Analysis

The gene expression levels quantified using RNA-seq were first normalized into reads per kilobase of transcript per million mapped reads (RPKM) and the minimum threshold was set to be RPKM > 0.1 in at least 10 individuals. For each gene, the gene expression levels were inverse quantile normalized to a standard normal distribution across samples. Common SNPs were extracted by filtering out SNPs with minor allele frequency less than 5%. The principal components (PCs) of the genotype matrix were computed to account for differences in genetic ancestry among GTEx samples. The top 3 genotype PCs, five observed covariates (gender, age, race, ethnicity and BMI), and hidden covariates (such as those caused by batch effects) inferred via Probabilistic

Estimation of Expression Residuals (PEER) (Stegle et al. 2012) were regressed out to generate a residual expression matrix.

## C. Results

We applied multiple linear regression between genotypes of common SNPs that are +/- 100 kb away from genes and their corresponding gene expression levels. Figure 3.1A shows a quantile-quantile (Q-Q) plot of the observed p-values against the p-values expected under the null hypothesis for all ten tissues. By comparing the Q-Q plots of $\beta_1$ and $\beta_2$'s p-values, we found that the $\beta_1$'s p-values deviate from the null expectation much earlier than the $\beta_2$'s p-values (Figure 3.1A). This agrees with previous findings that a large fraction of SNPs have an additive association with the expression level of a nearby gene (Yvert et al. 2003; Brem et al. 2002; Morley et al. 2004). However, there are also a substantial fraction of SNPs that show evidence for dominant effects on gene expression. Interestingly, whole blood has more SNPs with dominant associations than all the other tissues, with 156 SNP-gene pairs with non-zero effect sizes for $\beta_2$ under FDR = 5%.

The standardization step for both genotype and expression variables transforms the effect size into a correlation coefficient between expression and genotype values, which therefore ranges from -1 to 1. The effect sizes of additive eQTLs form a more symmetrical distribution than those of dominant eQTLs (Figure 3.1B). The effect sizes of dominant eQTLs ($\beta_2$) are skewed to the positive side, which indicates an unbalanced direction for dominant effects. This implies that the gene expression levels of most heterozygotes that exhibit dominance tend to be closer to those of the homozygotes with high-expression alleles than the low-expression alleles. The overall effect size of additive eQTLs is larger than that of dominant eQTLs.

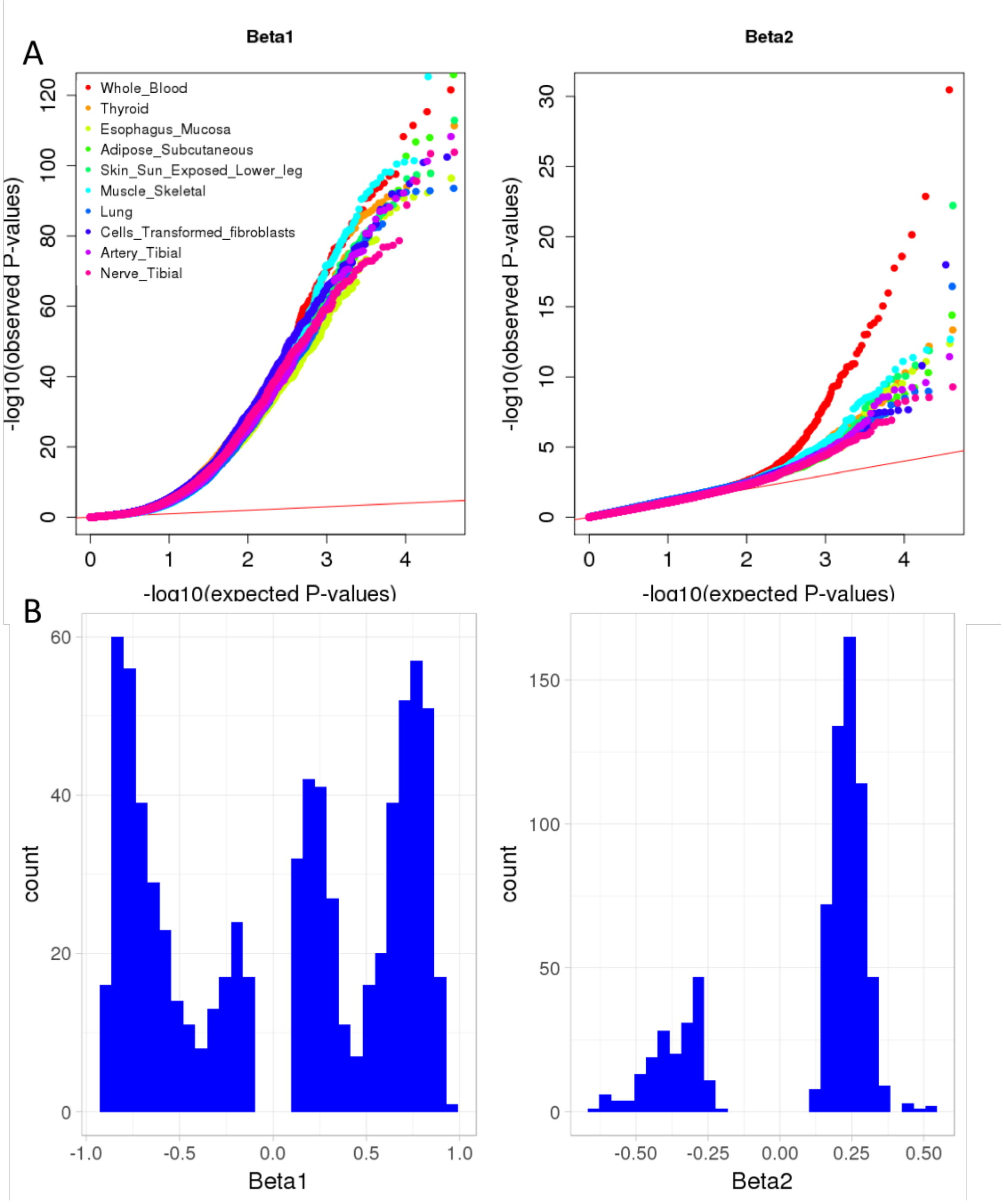
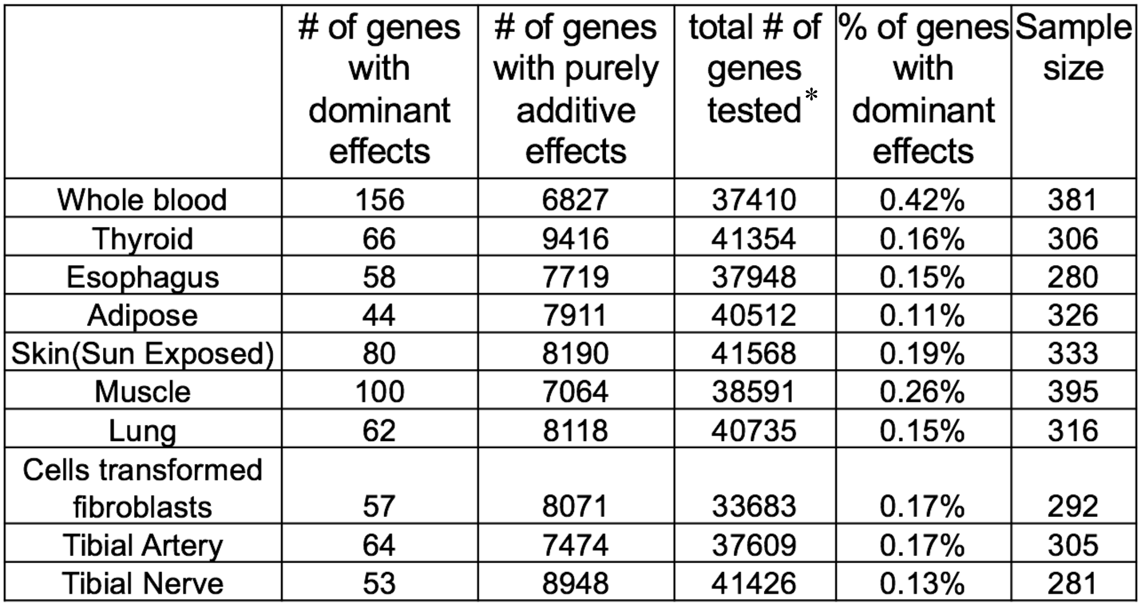


**Figure 3.1 Quantile-quantile plots for P-values and distribution of effect sizes for both additive and dominant variables**. (A) The observed p-values from testing whether the effect sizes of additive ($\beta_1$) and dominant ($\beta_2$) variables are significantly different from 0 are plotted against expected p-values under the null hypothesis, which are uniformly distributed between 0 and 1. The red line is the null expectation. (B) Histograms of effect sizes for $\beta_1$ and $\beta_2$. The distribution of $\beta_1$ is more symmetric.

Table 3.1 provides a summary of results from 10 tissues. Whole blood has a much higher percentage of genes that show dominant effects, which could potentially be explained by the large sample size of this tissue (n = 381) and the fact that the statistical power of our method is affected by sample size. However, the muscle tissue has a similar sample size (n= 395) but has far fewer dominant eQTLs. Surprisingly, muscle and whole blood tissues have the highest percentage of genes with dominant effects but do not have a higher proportion of genes with pure additive effects (18.2% and 18.3% respectively, compared to the average proportion of 20.4%).

**Table 3.1 Number of genes that show dominant effects or pure additive effects as well as sample sizes for each tissue**



| | # of genes with dominant effects | # of genes with purely additive effects | total # of genes tested* | % of genes with dominant effects | Sample size |
|---|---|---|---|---|---|
| Whole blood | 156 | 6827 | 37410 | 0.42% | 381 |
| Thyroid | 66 | 9416 | 41354 | 0.16% | 306 |
| Esophagus | 58 | 7719 | 37948 | 0.15% | 280 |
| Adipose | 44 | 7911 | 40512 | 0.11% | 326 |
| Skin(Sun Exposed) | 80 | 8190 | 41568 | 0.19% | 333 |
| Muscle | 100 | 7064 | 38591 | 0.26% | 395 |
| Lung | 62 | 8118 | 40735 | 0.15% | 316 |
| Cells transformed fibroblasts | 57 | 8071 | 33683 | 0.17% | 292 |
| Tibial Artery | 64 | 7474 | 37609 | 0.17% | 305 |
| Tibial Nerve | 53 | 8948 | 41426 | 0.13% | 281 |

* The genes tested for genetic associations include non-coding RNAs and pseudogenes.

For purely additive eQTLs with no evidence for a significant dominant effect, there is a clear linear association between SNP genotypes and normalized gene expression values (Figure 3.2A) and the median of expression values for heterozygotes is halfway in between those of homozygotes. When there are additional dominant effects, depending on the direction of the effect, the median of heterozygotes' expression values is either higher or lower than expected under the linear model (Figure 3.2 B, C).

The additive model typically used in mapping eQTLs may fail to identify some genetic associations with gene expression. When the effect size ($\beta_1$) for the additive variable $G_A$ is zero, some of the eQTLs that we identify still show significant effect sizes ($\beta_2$) for the dominant variable $G_D$ (Figure 3.2 D, E). When we examined some of these eQTLs, we found that the median expression value of heterozygotes was the highest or lowest of the three genotype classes, and we categorized them as over-dominant or under-dominant eQTLs.

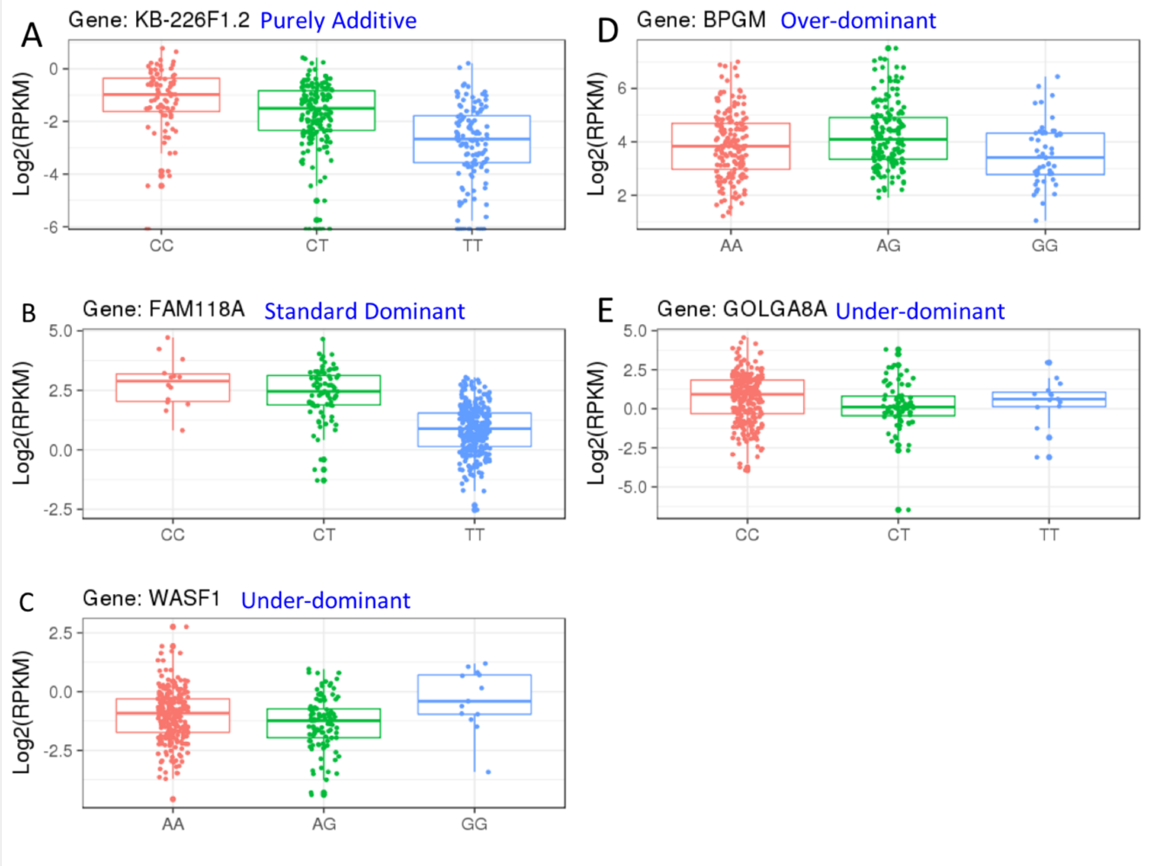


**Figure 3.2 Boxplots of log2 of gene expression values in RPKM against genotype groups.** Individuals were divided into three genotype groups based on the genotype of the SNP with the strongest association with the expression level of the gene. The genotype to the left is from reference homozygotes, while the genotype to the right is from non-reference homozygotes. Based on the way how the types of dominant associations were defined, each example was labeled with their own dominant type shown in blue.

To further examine the types of dominant associations that we observe, we grouped dominant eQTLs into three categories—standard dominant, over-dominant and under-dominant—based on the following procedure:

1) Compute the median gene expression values for all the three genotype groups;

2) Choose the homozygous genotype group that has a median expression value that is closest to that of the heterozygous group;

3) Apply a two-sided t-test to test whether the homozygous group and the heterozygous group have significantly different mean expression values.

When the mean expression values between the tested two groups are significantly different, there are three possible classifications for the eQTL depending on the median expression value of heterozygotes. If the heterozygous group has the highest median expression among the three genotype groups, the eQTL is over-dominant; if the heterozygous group has the lowest expression, the eQTL is under-dominant; otherwise, it is classified as a standard dominant eQTL. When there is no evidence that the two tested groups have different mean expression values, we again classify the eQTL as standard dominant.

The majority of the dominant eQTLs identified across the 10 tissues, show a standard dominant effect on gene expression (Figure 3.3). Dominant eQTLs with over-dominant or under-dominant effects are fairly rare, and are not well-explained by the theoretical mechanisms proposed in Figure 1.3. The under- and over- dominant eQTLs therefore require further investigation.

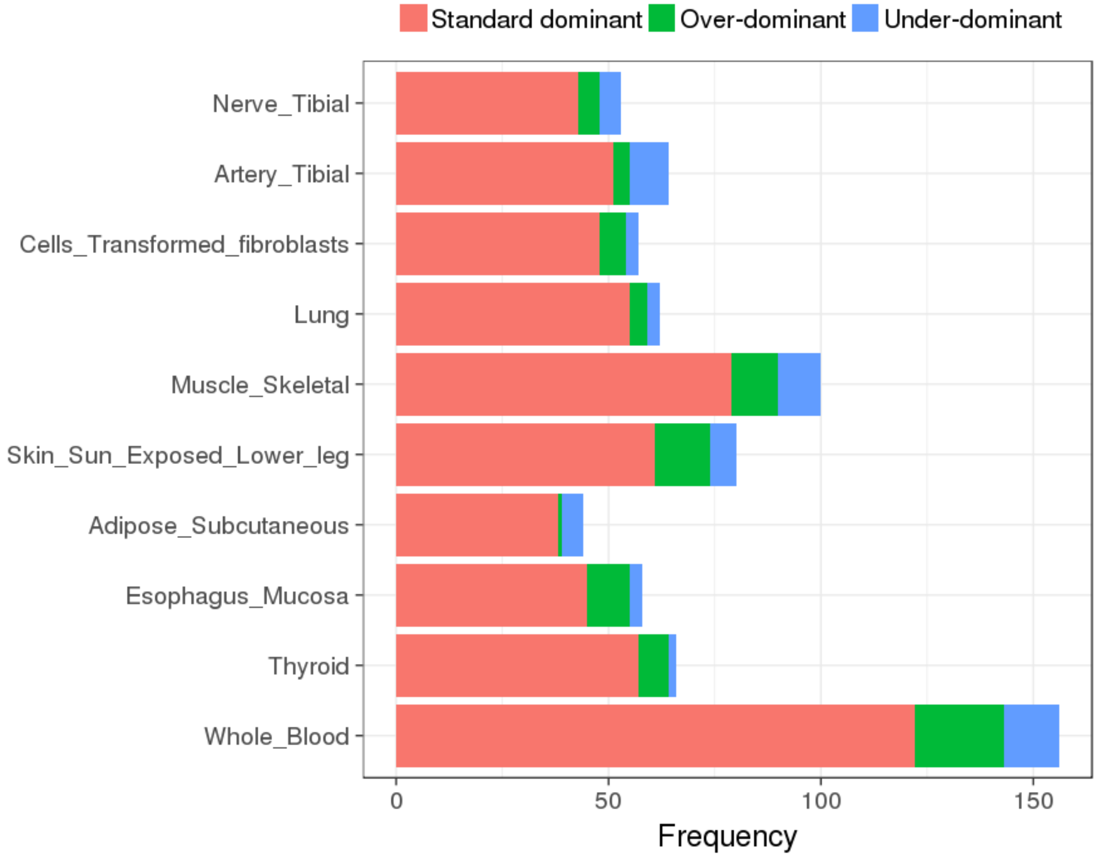


**Figure 3.3 Distribution of dominant types across multiple tissues.**

The positions of dominant eQTLs relative to the transcription start site (TSS) of the genes they are associated with may give us some insight into how they function. As expected, additive eQTLs with no dominant effects are centered around the TSS (Figure 3.4A), which is consistent with previous findings that eQTLs are enriched in close proximity to the genes they are associated with (Battle et al. 2014; Veyrieras et al. 2008). A similar trend was observed for dominant eQTLs, which are located close to the gene TSSs (Figure 3.4B). It would be interesting to further compare the positions of the eQTL SNPs with the positions of SNPs with similar characteristics that have no association with gene expression levels.

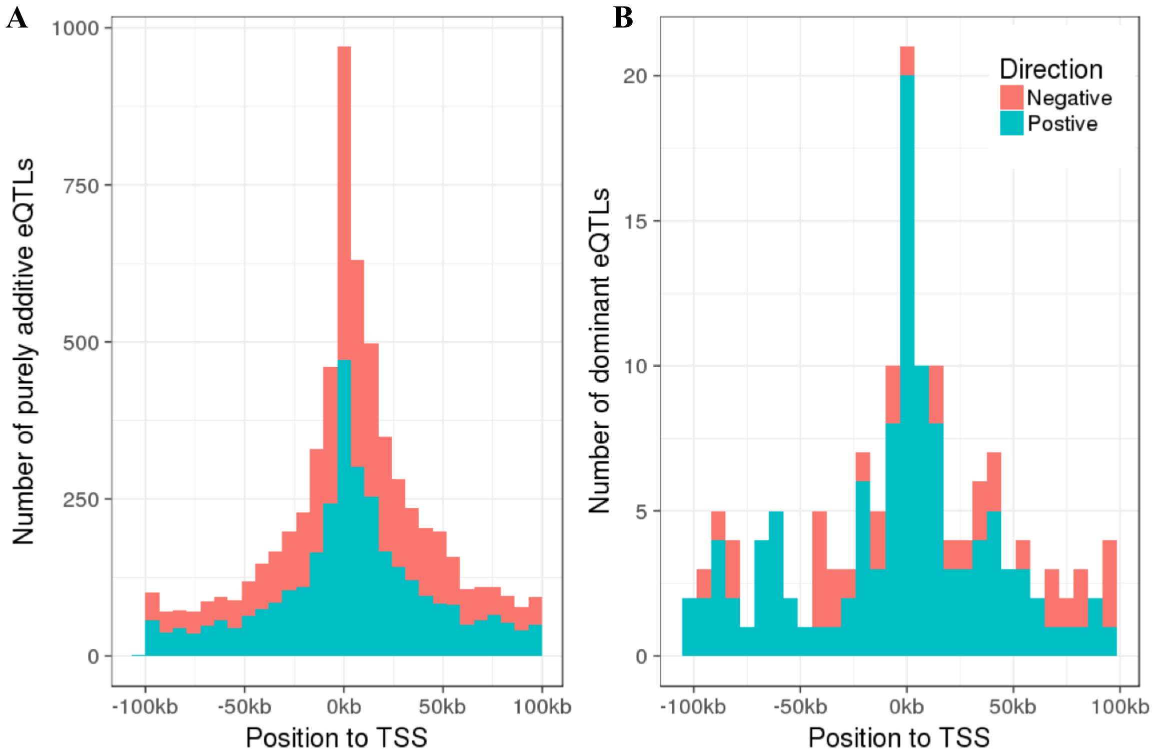


**Figure 3.4 Positions of purely additive eQTLs and dominant eQTLs from their corresponding genes' transcription start sites.** (A) eQTLs that show purely additive effects, were divided into two groups based on the sign of the additive effect size ($\beta_1$), turqoise for $+\beta_1$ and red for $-\beta_1$. (B) The same color scheme was used for the dominant effect size ($\beta_2$).

As our method has enabled us to identify a substantial fraction of genes that show dominant effects in different tissues, we would like to know how many of these genes have dominant associations in multiple tissues. The majority of genes with dominant effects only occur in one of the tissues, which is probably due to incomplete statistical power and the limited amount of tissues that have relatively large sample sizes (Figure 3.5A). Among the 13 genes that occur in at least 7 tissues, 4 of them are pseudogenes and 5 are HLA genes. (Figure 3.5B). In total, 20.7% of the genes that show dominant effects across tissues are pseudogenes, which is significantly higher than the proportion in genes that show additive effects (12.1%;  $P = 8.1\times10^{-15}$ by Fisher's exact test).

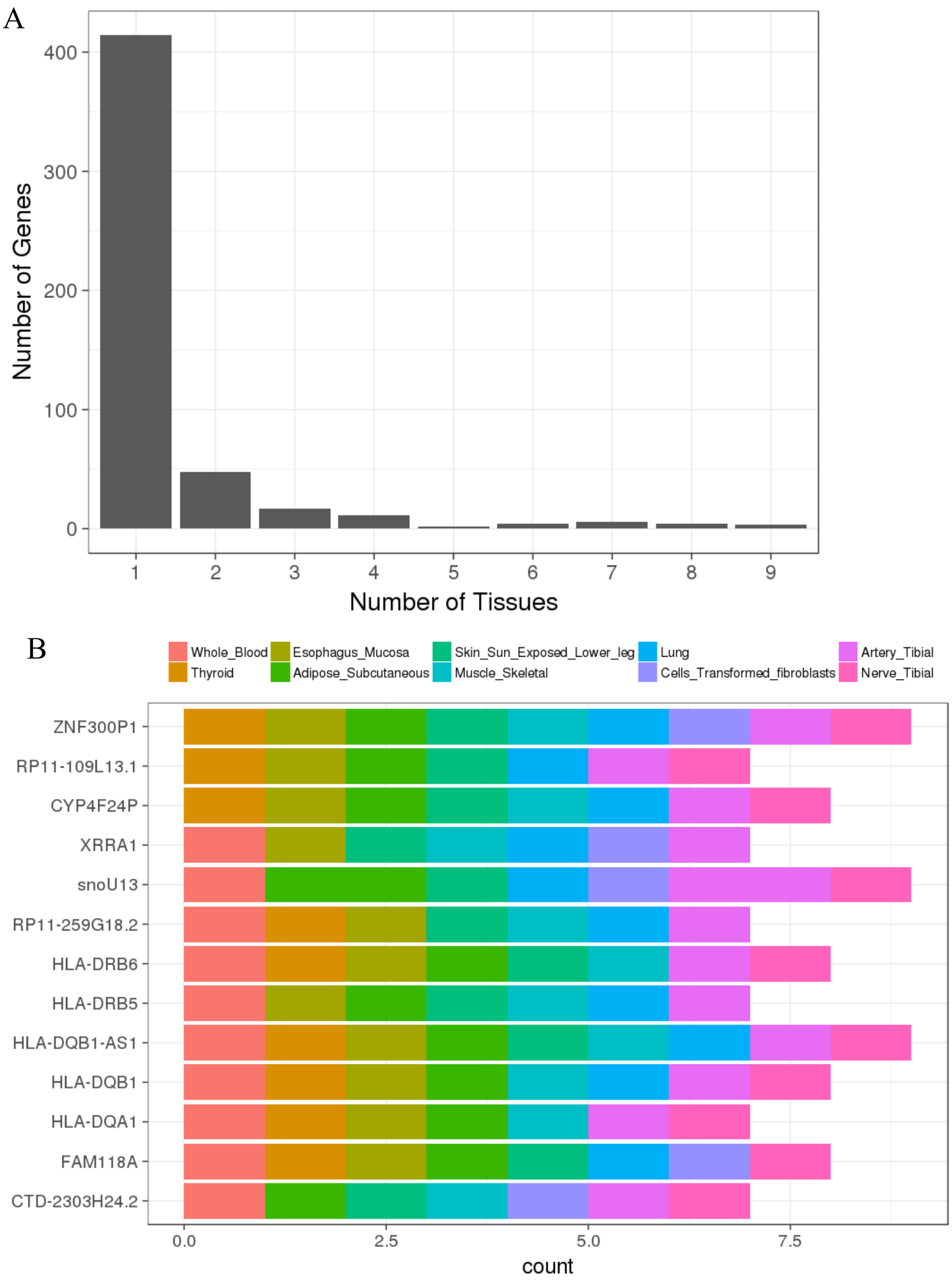


**Figure 3.5 Number of genes showing dominant effects across tissues.** (A) The number of tissues where the same gene is found to have dominant association with gene expression levels. (B) The identities of genes that occur in at least 7 tissues.

To understand the function of dominant eQTLs on gene regulation, we looked into the average expression values across individuals. Interestingly, genes that show dominant effects have lower average gene expression levels compared to those that only show additive effects. This observation was consistent across all the tissues tested.

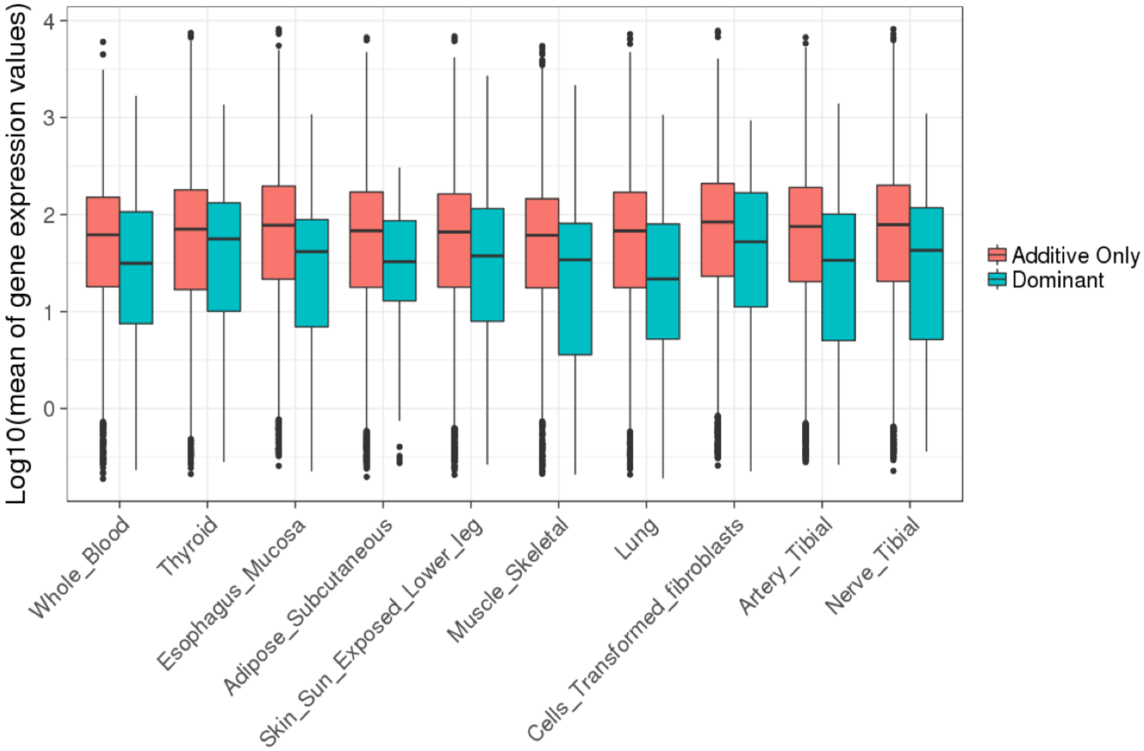


**Figure 3.6 Comparison of average gene expression levels between eQTLs that only show additive effects and dominant eQTLs across tissues.** The box-plot shows log10 of mean gene expression values across individuals.

We further grouped genes into gene ontology (GO) categories and tested for an enrichment of genes associated with dominant eQTLs using Fisher's exact test (Ashburner et al. 2000). Using this method, we found that 8 out of 10 tissues have dominant genes significantly enriched in the major histocompatibility complex (MHC). For instance, genes with dominant effects in the whole blood tissue are significantly enriched for genes from the MHC GO category ($p=5.5\times10^{-9}$; $FDR=2\times10^{-5}$. After removing genes in the MHC GO category, none of the other GO categories in biological process, cellular component, or molecular function are significant at the FDR=0.05 level.

# IV. Discussion

We examined common SNPs that were +/- 100kb away from genes on autosomal chromosomes and applied a multiple linear regression model to detect dominant genetic associations with gene expression levels. We corrected for the SNP with the lowest P-value for each gene using a beta approximation permutation scheme, which requires fewer permutations to achieve desired p-values (Figure 2.2). To further speed up our pipeline, we performed large matrix operations instead of repeatedly calling the lm function in the R stats package. With these optimizations, we were able to run our method on large samples with a much higher efficiency. By applying a multiple linear regression model, we found that 0.19% of all the genes tested (including long non-coding RNAs and pseudogenes) show dominant genetic associations in ten human tissues. The effect sizes of dominant variables ($\beta_2$) are on average smaller than those of additive variables ($\beta_1$) and are highly skewed to positive numbers (Figure 3.1). This implies that, for most dominant eQTLs, the expression levels of heterozygotes are closer to that of the high-expression homozygous genotype than expected. In other words, in most dominant associations, the expression levels of heterozygotes were upregulated. The majority of dominant eQTLs are classified as standard dominant for each tissue, rather than over-dominant or under-dominant. Few of the genes that are associated with dominant eQTLs in one tissue are associated with dominant eQTLs in other tissues (Figure 3.5). This may reflect our incomplete power to detect dominant eQTLs or indicate that many dominant eQTLs are tissue-specific. Dominant eQTLs are located in close proximity to the genes they are associated with. Finally, genes with dominant eQTLs tend to have lower mean expression values, and this observation is consistent across tissues (Figure 3.6). It is unclear to us why genes with lower mean expression values are more prone to have dominant effects but it is possible that genetic variants exert dominant effects to compensate for the low expression of genes.

Gene ontology (GO) analysis enabled us to find that 8 out of 10 tissues have dominant genes enriched in the major histocompatibility complex (MHC). The variants nearby MHC genes are highly polymorphic, which may give rise to mapping bias toward reference alleles. Potentially this bias could distort linear relationship between genotypes and expression values, however it is not clear if mapping bias would result in false dominant associations and dominant associations have previously been observed in the MHC region. Specifically, Lenz et al. discovered that genetic variants in the MHC region have a dominant association with the risk of autoimmune diseases such as rheumatoid arthritis and celiac disease (Lenz et al. 2015). While this raises the intriguing possibility of a link between dominant gene expression within the MHC region and autoimmune disease risk, further study is required to rule out the possibility of mapping artifacts within this highly polymorphic region.

In the future, allele-specific expression analysis can be performed to see if reads overlapping heterozygous sites come equally from both alleles or are biased to one allele. This approach can help us better understand the mechanism of dominant associations (Figure 1.3) and can reveal whether dominant eQTLs act in cis or trans to regulate gene expression. It would also be interesting to see if dominant eQTLs are enriched with other functional and genomic annotations such as enhancers and transcription factor binding sites. Due to our limited statistical power to detect dominant associations, a larger dataset with more tissues may help to discover more genes with dominant genetic associations.

# REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics, 25(1), 25–29.

Adler, D. A., E. I. Rugarli, P. A. Lingenfelter, K. Tsuchiya, D. Poslinski, H. D. Liggitt, V. M. Chapman, R. W. Elliott, A. Ballabio, and C. M. Disteche. 1997. “Evidence of Evolutionary up-Regulation of the Single Active X Chromosome in Mammals Based on Clc4 Expression Levels in Mus Spretus and Mus Musculus.” Proceedings of the National Academy of Sciences of the United States of America 94 (17): 9244–48.

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB,Levinson DF, Koller D. 2014. “Characterizing the Genetic Basis of Transcriptome Diversity through RNA-Sequencing of 922 Individuals.” Genome Research 24 (1): 14–24.

Brem, Rachel B., Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. 2002. “Genetic Dissection of Transcriptional Regulation in Budding Yeast.” Science 296 (5568): 752–55.

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

Consortium, International Hapmap, and Others. 2005. “A Haplotype Map of the Human Genome.” Nature 437 (7063). NIH Public Access: 1299.

Cui, Xiangqin, Jason Affourtit, Keith R. Shockley, Yong Woo, and Gary A. Churchill. 2006. “Inheritance Patterns of Transcript Levels in F1 Hybrid Mice.” Genetics 174 (2): 627–37.

Deng X, Hiatt JB, Nguyen DK, Ercan S, Sturgill D, Hillier LW, Schlesinger F, Davis CA, Reinke VJ, Gingeras TR, Shendure J, Waterston RH, Oliver B, Lieb JD, Disteche CM. 2011. “Evidence for Compensatory Upregulation of Expressed X-Linked Genes in Mammals, Caenorhabditis Elegans and Drosophila Melanogaster.” Nature Genetics 43 (12): 1179–85.

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K. 2008. "Genetics of Gene Expression and Its Effect on Disease." Nature 452 (7186): 423–28.

Gibson, G., Riley-Berger, R., Harshman, L., Kopp, A., Vacha, S., Nuzhdin, S., & Wayne, M. (2004). Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. Genetics, 167(4), 1791–1799.

GTEx Consortium. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science, 348(6235), 648–660.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America, 106(23), 9362–9367.

Jones, M.C. Kumaraswamy's distribution: a beta-type distribution with some tractability advantages. Stat. Methodol. (2009): 6, 70–81.

Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PA, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Geuvadis Consortium, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501(7468), 506–511.

Lenz TL, Deutsch AJ, Han B, Hu X, Okada Y, Eyre S, Knapp M, Zhernakova A, Huizinga TW, Abecasis G, Becker J, Boeckxstaens GE, Chen WM, Franke A, Gladman DD, Gockel I, Gutierrez-Achury J, Martin J, Nair RP, Nöthen MM, Onengut-Gumuscu S, Rahman P, Rantapää-Dahlqvist S, Stuart PE, Tsoi aJT, Gregersen PK, Schumacher J, Rich SS, Wijmenga C, Sunyaev SR, de Bakker PI, Raychaudhuri S, 2015. "Widespread Non-Additive and Interaction Effects within HLA Loci Modulate the Risk of Autoimmune Diseases.” Nature Genetics 47 (9): 1085–90.

Lewis, E. B. 1954. “The Theory and Application of a New Method of Detecting Chromosomal Rearrangements in Drosophila Melanogaster.” The American Naturalist 88 (841). [University of Chicago Press, American Society of Naturalists]: 225–39.

Morley, Michael, Cliona M. Molony, Teresa M. Weber, James L. Devlin, Kathryn G. Ewens, Richard S. Spielman, and Vivian G. Cheung. 2004. “Genetic Analysis of Genome-Wide Variation in Human Gene Expression.” Nature 430 (7001). nature.com: 743–47.

Nguyen, Di Kim, and Christine M. Disteche. 2006. “Dosage Compensation of the Active X Chromosome in Mammals.” Nature Genetics 38 (1): 47–53.

Nica, Alexandra C., Stephen B. Montgomery, Antigone S. Dimas, Barbara E. Stranger, Claude Beazley, Inês Barroso, and Emmanouil T. Dermitzakis. 2010. “Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations.” PLoS Genetics 6 (4): e1000895.

Ongen, Halit, Alfonso Buil, Andrew Anand Brown, Emmanouil T. Dermitzakis, and Olivier Delaneau. 2016. “Fast and Efficient QTL Mapper for Thousands of Molecular Phenotypes.” Bioinformatics 32 (10): 1479–85.

Powell, Joseph E., Anjali K. Henders, Allan F. McRae, Jinhee Kim, Gibran Hemani, Nicholas G. Martin, Emmanouil T. Dermitzakis, Greg Gibson, Grant W. Montgomery, and Peter M. Visscher. 2013. “Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data.” PLoS Genetics 9 (5): e1003502.

Shabalin, Andrey A. 2012. “Matrix eQTL: Ultra Fast eQTL Analysis via Large Matrix Operations.” Bioinformatics 28 (10): 1353–58.

Stegle, Oliver, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. 2012. “Using Probabilistic Estimation of Expression Residuals (PEER) to Obtain Increased Power and Interpretability of Gene Expression Analyses.” Nature Protocols 7 (3): 500–507.

Stupar, Robert M., Peter J. Hermanson, and Nathan M. Springer. 2007. “Nonadditive Expression and Parent-of-Origin Effects Identified by Microarray and Allele-Specific Expression Profiling of Maize Endosperm.” Plant Physiology 145 (2): 411–25.

Veyrieras, Jean-Baptiste, Sridhar Kudaravalli, Su Yeon Kim, Emmanouil T. Dermitzakis, Yoav Gilad, Matthew Stephens, and Jonathan K. Pritchard. 2008. "High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation." PLoS Genetics 4 (10): e1000214.

Yvert, Gaël, Rachel B. Brem, Jacqueline Whittle, Joshua M. Akey, Eric Foss, Erin N. Smith, Rachel Mackelprang, and Leonid Kruglyak. 2003. "Trans-Acting Regulatory Variation in Saccharomyces Cerevisiae and the Role of Transcription Factors." Nature Genetics 35 (1): 57–64.