Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Chen ZHAO Date

Quantitative trait locus analysis of molecular phenotypes in the GTEx cohort

By

Chen Zhao

Master of Science in Public Health

Biostatistics and Bioinformatics

_____

Zhaohui (Steve) Qin, PhD

(Thesis Advisor)

_____

Xiangqin Cui, PhD

(Reader)

Quantitative trait locus analysis of molecular phenotypes in the GTEx cohort

By

Chen Zhao

B.A.

University of Science and Technology of China

2013

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD

Reader: Xiangqin Cui, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University in partial fulfillment of the requirements

for the degree of Master of Public Health

In Biostatistics and Bioinformatics Department

2020

Abstract

Quantitative trait locus analysis of molecular phenotypes in the GTEx cohort

By Chen Zhao

Background: One of the main purposes of human genetic research is to understand the function of genetic variants. Expression quantitative trait loci (eQTL) analyses have been successfully carried out to identify variants that affect the expression level of their target gene. Due to computation cost, existing analyses focus on cis eQTLs, and only evaluate the variants effect on individual genes which may be affected by the excessive uncertainties and noise in the gene expression measurements.

Method: In this study, we study the impact of genetic variants on the overall expression levels of biological pathways using data from the Genotype-Tissue Expression (GTEx) consortium. We applied the GSVA and combined Z-score methods to transform the gene expression data to pathway-level expression scores. Then we utilized these scores instead of the raw expression data for QTL analysis to find the SNPs with significant association *p*-values, and their corresponding pathway and tissue.

Results: We found eight significant pathway/tissue pairs with genome-wide significant QTLs.: Folate Biosynthesis / Adipose Subcutaneous, Folate Biosynthesis / Muscle Skeletal, Sulfur Metabolism / Muscle Skeletal, Taste Transduction / Skin - Sun Exposed (Lower leg) for GSVA method, and Glycosaminoglycan Biosynthesis Chondroitin Sulfate / Brain Frontal Cortex BA9, Glyoxylate and Dicarboxylate Metabolism / Brain Frontal Cortex BA9, Folate Biosynthesis / Adipose Subcutaneous, Olfactory Transduction / Adrenal Gland for Z-score method.

Conclusion: Our analysis identified significant QTLs related to biological pathways in multiple tissues. Many of these QTLs are located in the coding regions of the gnome. These findings may help us to better understand the biological functions of genes, pathways and their connections with genetic variants.

Keywords: Pathway, Expression quantitative trait loci, Single Nucleotide Polymorphism

Quantitative trait locus analysis of molecular phenotypes in the GTEx cohort

By

Chen Zhao

B.A.

Emory University

2020

Thesis Committee Chair: Zhaohui (Steve) Qin, PhD

Reader: Xiangqin Cui, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University in partial fulfillment of the requirements

for the degree of Master of Public Health

In Biostatistics and Bioinformatics Department

2020

Introduction

To understand the impact of the genetic variants is a fundamental objective in human genetics research. As the human genome sequencing project completed, we have the opportunity to study the impact of the variants on various types of molecular phenotypes. Over the last fifteen years, many variants have been found to have a strong association with diseases such as Alzheimer's disease (AD) using genome-wide association studies (1,2) (Gatz et al., 1997; Wingo, Lah, Levey, & Cutler, 2012). But most of the variants are located in the non-coding region of the genome, we cannot identify the feature for variants directly. Thus, it is a grand challenge to study the functional impact of the variants. Meanwhile, next generation sequencing technologies give us a chance to study the transcriptome patterns across tissue types. The GTEx (3,4) (Consortium, 2015; Carithers & Moore, 2015) project provides transcriptome data from multiple tissues at the population–level, giving us the opportunity to conduct a comprehensive study on the impact of genetic variants on gene expression. Based on GTEx database, we get many new findings such as expression quantitative trait loci (eQTL). Using the approach of eQTL, we can make novel findings for complex diseases such as AD.

Although eQTL information can help the identification of the target genes of GWAS-identified variants (5-7) (Hormozdiari et al., 2016; Ratnapriya et al., 2019; Gamazon et al., 2018), it has three limitations. First, there are noises produced with the high-throughput technologies and uncertainties in the measurement procedure. Second, considering the time and computing cost, trans-eQTL is often not considered. Third, eQTL analysis is only conducted at the single gene-level, the combination of the genes function is not being considered in this approach.

In our study, to overcome these limitations, we turn to the pre-defined, expert-curated molecular pathways that have been cataloged in databases such as KEGG (8,9) (Kanehisa & Goto, 2000; Du et al., 2014). An obvious benefit of using these pathway-based methods is interpretability. Given that gene function may deviate and can be influenced by the environment or the disease state, using the gene set method can yield a stable and intuitive result to evaluate the biological impact of genetic variants.

Our research utilizes gene set variance analysis (GSVA) method (10) (Hänzelmann, Castelo, & Guinney, 2013) to derive a quantitative summative assessment of a pathway's activity at the single sample level, which then enables QTL analysis to be extended from the single gene level to the pathway level. GSVA has been shown to be an effective way to summarize pathway activities at the individual level from transcriptome profiling data. GSVA calculates gene enrichment score, using the approach of comparing the gene inside and outside of the pathway, and evaluate the variance of the enrichment score over samples. In our study, using data from the GTEx consortium, we first calculate the GSVA score for each of the 186 KEGG pathways in each individual. Next, we conduct an genome-wide QTL scan using these GSVA values. Our research intends to analyze the relationship between the pathway activity QTL we identified and the relationships with GWAS and eQTL results of nearby SNPs.

Method

First, we obtained the data from the Genotype-Tissue Expression (GTEx) consortium, which includes gene expression, genotype, and clinical data for 449 human donors across 44 tissues. Generally, the whole dataset contains three parts: gene expression data, the SNP (Single Nucleotide Polymorphism) genotype data and the covariates data. In our study, we use the gene expression levels data from the GTEx v7 release. Each row for the gene expression data corresponds to one gene, and each column corresponds to a sample. For SNP genotype data, each row represents one sample, and each column represents one SNP. We use SNP in a broader sense, which include SNP and other type of variants like short indels, that are profiled in the GTEx study. The covariate data is a matrix with the rows of the covariate details and columns of samples.

We first sorted and separated the gene expression and covariate data depending on the tissue sources of the samples. Next, based on KEGG pathway database, we applied GSVA method on the gene expression data to convert the gene-level data to pathway-level enrichment scores for each tissue.

Now we introduce the algorithm of GSVA method. First, we have an input of a matrix $X = \{x_{ij}\}_{p\times n}$, which represents the normalized expression values for p genes by n samples. Also, we have a collection of gene sets $\Gamma = \{\gamma_1, \dots, \gamma_m\}$. We denote x_i for the expression profile of the *i*-th gene and x_{ij} corresponds to the specific expression value for *j*-th sample in *i*-th gene. γ_k represents a pathway in the collection Γ, $|\gamma_k|$ is the number of genes in γ_k.

Next, we evaluate the expression level of a gene *i* in sample *j* in the context of sample population distribution. We calculate an expression-level statistic. For each gene expression profile $x_i = \{x_{i1}, \dots, x_{in}\}$, a Gaussian kernel is used to calculate the non-parametric kernel estimation of the cumulative density function for the expression profile. The formula is:

$$\widehat{F_{h_i}}(x_{ij}) = \frac{1}{n}\sum_{k=1}^{n}\int_{-\infty}^{\frac{x_{ij}-x_{ik}}{h_i}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt, \quad (1)$$

Where h_i is the parameter of bandwidth that controls the resolution of the kernel estimation. $h_i = s_i/4$, where s_i is the sample standard deviation of the *i*-th gene. In terms of RNA-seq data, we use a discrete Poisson kernel:

$$\widehat{F_r}(x_{ij}) = \frac{1}{n}\sum_{k=1}^{n}\sum_{y=0}^{x_{ij}} \frac{e^{-(x_{ik}+r)}(x_{ik}+r)^y}{y!}, \quad (2)$$

Where $r = 0.5$, in order to set the mode of Poisson kernel at each x_{ik}.

Let z_{ij} denote $\widehat{F_{h_i}}(x_{ij})$, or $\widehat{F_r}(x_{ij})$, depending on whether x_{ij} are continuous microarray or discrete count RNA-seq values. In order to reduce the effect of the outliers, we convert z_{ij} to ranks $z_{(i)j}$ for each sample *j*. We denote $r_{ij} = |p/2 - z_{(i)j}|$, to centralize the statistics. Then we calculate the Kolmogorov-Smirnov (KS) like random walk statistic:

$$\nu_{jk}(\ell) = \frac{\sum_{i=1}^{\ell}|r_{ij}|^{\tau} I(g_i \in \gamma_k)}{\sum_{i=1}^{p}|r_{ij}|^{\tau} I(g_i \in \gamma_k)} - \frac{\sum_{i=1}^{\ell} I(g_i \notin \gamma_k)}{p - \gamma_k}, \quad (3)$$

where τ is a parameter describing the weight of the tail in random walk, with the default value 1, γ_k represents the *k*-th gene set, $I(g_i \in \gamma_k)$ is the indicator function shows whether the *i*-th gene

(the gene corresponding to the *i*-th ranked expression-level statistic) is in the pathway γ_k, $|\gamma_k|$ is the number of genes in γ_k, and *p* is the number of genes in the whole dataset.

The next step is turning the KS like statistic into an enrichment statistic, named GSVA score. Under the null hypothesis that no change in pathway activity throughout the sample population, we provide a standard Gaussian distribution of enrichment scores, and the ES score is defined like:

$$ES_{jk}^{diff} = \left|ES_{jk}^{+}\right| - \left|ES_{jk}^{-}\right| = \max_{\ell=1,\ldots,p}(0, \nu_{jk}(\ell)) - \min_{\ell=1,\ldots,p}(0, \nu_{jk}(\ell)) \quad (4)$$

The biological interpretation for this statistic is: it can show the degree of how genes in pathways activate in one direction, either over-expressed or under-expressed. If the pathway contains genes that are acting in both directions, the value will cancel out. This ES score is unimodal and approximately normal.

Alternatively, for the purpose of validation, we adopt another unsupervised GSE method called calculate single sample pathway summaries of expression, the combined z-score method (11). It standardized first expression profiles into z-scores over samples and combines them together for each gene set at each individual sample as follows. Given a gene set $\gamma = \{1, \ldots, k\}$ with z-scores $Z_1, \ldots, Z_k$ for each gene, the combined z-score Z_γ for the gene set γ is defined as:

$$Z_\gamma = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \quad (5)$$

In this paper, using the R package "GSVA", we applied the GSVA approach to turn the gene-level expression data into GSVA enrichment score data with pathway level. In addition, we also consider the combined z-score method. Next, we used eQTL method to identify the SNPs which

are significantly associated with the corresponding GSVA scores or the score of the combined z-score method.

The basic idea for the eQTL analysis is using linear regression and ANOVA models to find the association between expression and SNP genotype. For each gene-SNP pair, the codes for SNP is 0,1, and 2 corresponding to the minor allele number. Then we build a linear regression model between GSVA score g and genotype s, with the covariate x like:

$$g = \beta_0 + \beta_1 s + \beta_2 x + \epsilon, \qquad where\ \epsilon \sim i.i.d.N(0, \sigma^2) \tag{6}$$

We choose one tissue with the covariates, and one pathway-level score to test all the SNPs p-value for β_1, and check the association between the SNP genotypes and the GSVA scores each time. Then we repeat the whole procedure to test all tissues and pathways that we are interested in.

Normally, the eQTL analysis is known to be time consuming and computationally intensive, because the genotype measured over millions of SNPs with over ten billion tests. In this paper, we use the method named MatrixeQTL (12) . It used matrix operator to optimize the calculation algorithm and speed up the computation.

Additionally, we adopt some data pre-processing steps in order to remove some SNPs that contain too much noise, or with queationable qualities. There are three criteria that we use to filter SNPs. In particularly, we remove SNPs that:

1. Have minor allele frequency less than 0.1
2. Have missing genotyping rate higher than 0.2
3. Have Hardy-Weinberg Equilibrium exact test (13) p-value less than 0.05

Finally, we report the SNPs with the significant *p*-value and its corresponding pathway, tissue, analyze the result we got.

Result

First, we calculated each samples' GSVA score and Z-score for each KEGG pathway based on the GTEx gene expression data, treating them as the response variables. Before the filtering, we have 79,457,242 SNPs in total, then after the data cleaning procedure, there are 4,362,883 SNPs left for the eQTL analysis. There are 186 KEGG pathway set and 39 tissues types. Because of the high computation cost, using p-value threshold of 5×10^{-8} , a commonly used significance threshold for genome-wide association studies, we first check which pairs of pathway/tissue combinations can generate more significant results, and we can focus on analyzing the property for those pairs.

With this threshold, we find 2204 SNPs as significant QTLs in total using GSVA scores involving 171 pathways and all 39 tissue types, and 2181 SNPs as significant QTLs using the Z-scores involving 179 pathways and 38 tissue types. They are shown as heatmap in Figure 1. and Figure 2. There are 5 SNPs significantly associated with pathway/tissue in both methods. They are rs411828 (One Carbon Pool By Folate / Brain Frontal Cortex BA9), rs12549084 (Folate Biosynthesis / Adipose – Subcutaneous), rs229081 (Glycosylphosphatidylinositol GPI Anchor Biosynthesis / Small Intestine - Terminal Ileum), rs2217861 (Glyoxylate And Dicarboxylate Metabolism / Cells - EBV-transformed lymphocytes), and rs869309398 (Pancreatic Cancer / Brain - Spinal cord (cervical c-1)). Among the pathways, pathway Folate Biosynthesis has the most SNPs, 379. Among the tissues, tissue Muscle Skeletal has the most SNPs--491.

Correspondingly, for the Z-score method, the most SNPs pathway is Olfactory Transduction and the tissue is Brain Frontal Cortex BA9, the number is 198 and 278 respectively. Moreover, the smallest p-value is 2.59×10^{-13}, which belongs to the combination of Taste Transduction / Skin - Sun Exposed (Lower leg) with GSVA method, and 4.42×10^{-12}, corresponding to the pair Olfactory Transduction / Adrenal Gland. In addition, the distributions of these two methods are not similar even though the total number are close. We can clearly find out through the heatmap that the significant SNPs in the method GSVA are more concentrated in several and those in the method Z-score are dispersion.

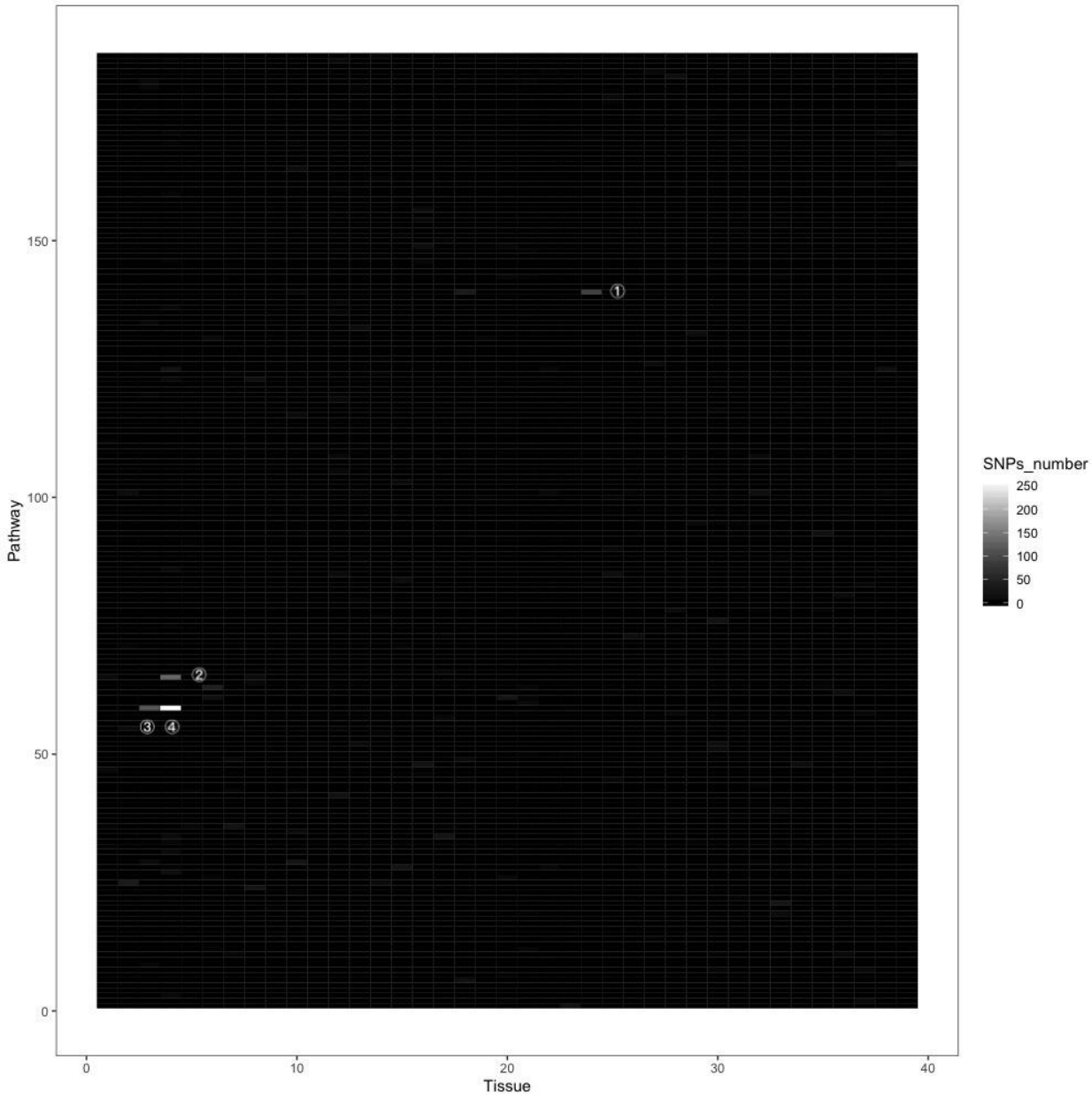

Figure 1. The heatmap for SNP numbers with GSVA Method

① Pathway: Taste Transduction, Tissue: Skin - Sun Exposed (Lower leg)
② Pathway: Sulfur Metabolism, Tissue: Muscle Skeletal
③ Pathway: Folate Biosynthesis, Tissue: Adipose Subcutaneous
④ Pathway: Folate Biosynthesis, Tissue: Muscle Skeletal

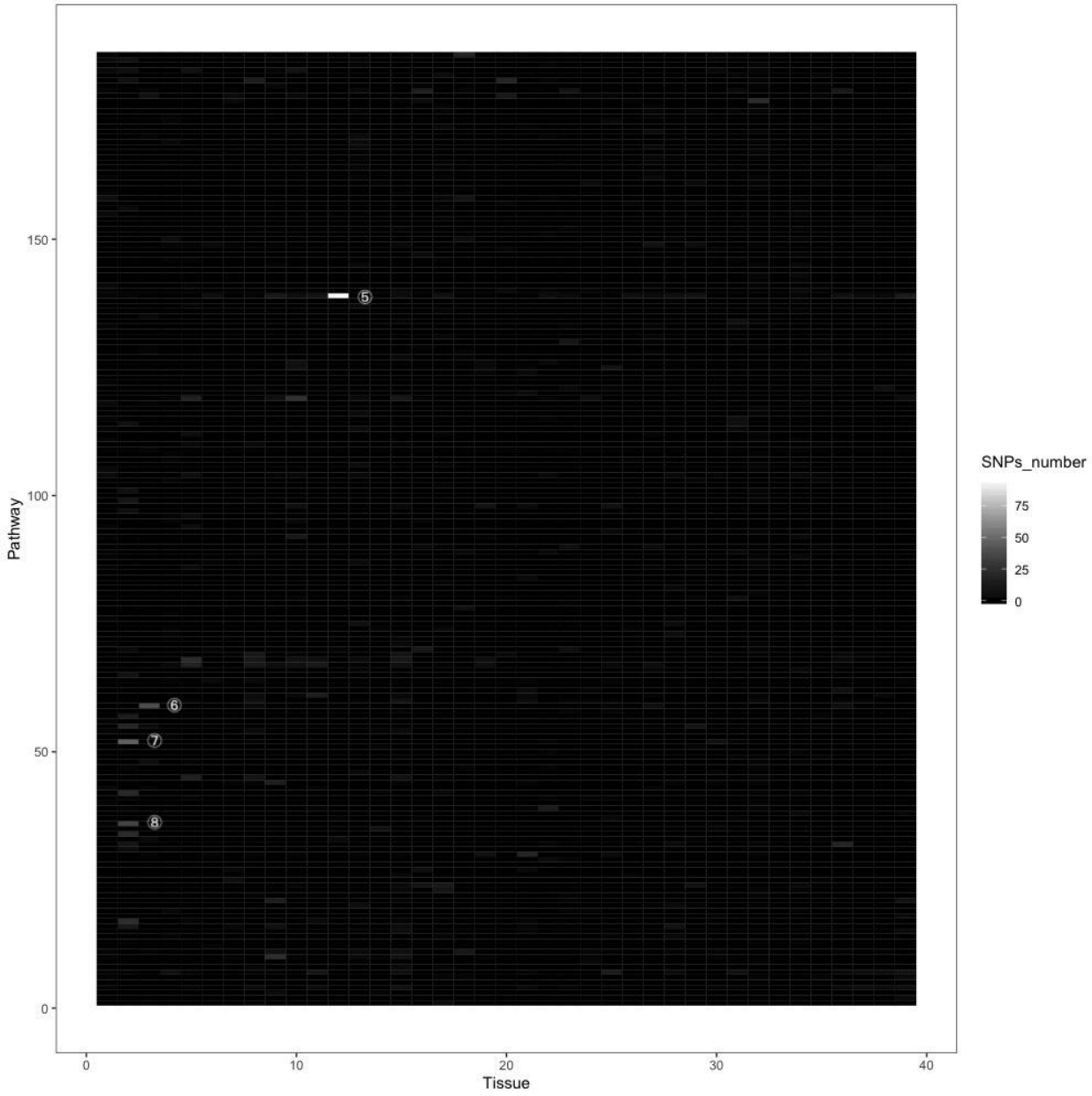

Figure 2. The heatmap for SNP numbers with Z-score Method

⑤ Pathway: Olfactory Transduction, Tissue: Adrenal Gland
⑥ Pathway: Folate Biosynthesis, Tissue: Adipose Subcutaneous
⑦ Pathway: Glyoxylate and Dicarboxylate Metabolism, Tissue: Brain Frontal Cortex BA9
⑧ Pathway: Glycosaminoglycan Biosynthesis Chondroitin Sulfate, Tissue: Brain Frontal Cortex BA9

With the demonstration of Figure 1 and Figure 2, we select 4 representative combinations for each method. For GSVA, all 4 the pairs we selected, which are Folate Biosynthesis / Adipose Subcutaneous, Folate Biosynthesis / Muscle Skeletal, Sulfur Metabolism / Muscle Skeletal, and Taste Transduction / Skin - Sun Exposed (Lower leg) have over 90 significant SNPs *p*-value. For Z-score, all of the pairs we selected have over 30 significant SNPs *p*-value, which are Glycosaminoglycan Biosynthesis Chondroitin Sulfate / Brain Frontal Cortex BA9, Glyoxylate and Dicarboxylate Metabolism / Brain Frontal Cortex BA9, Folate Biosynthesis / Adipose Subcutaneous, Olfactory Transduction / Adrenal Gland.

We draw the Manhattan plots for each combination in Figure 3-10. In these plots, the x-axis represents the SNP location on the Chromosome, the y-axis represents the negative log 10 of the *p*-value for the corresponding SNP. In addition, the blue line is the threshold 1×10^{-5}, and the red line is the threshold 5×10^{-8}.

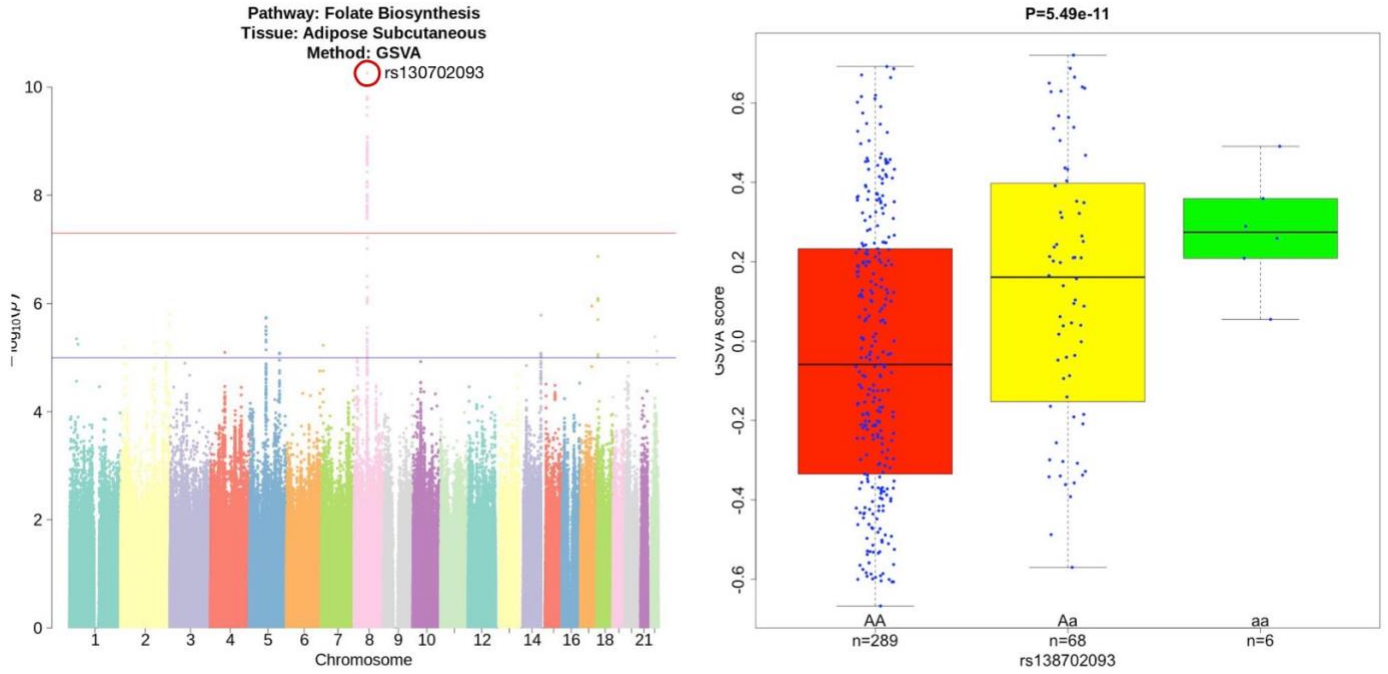

Figure 3. Manhattan Plot for the eQTL *p*-value with the boxplot of the peak SNP's genotype

Folate Biosynthesis / Adipose Subcutaneous with GSVA Method

In Figure 3, we find the SNP rs138702093 has the most significant *p*-value, 5.49×10^{-11}, which lies in the Chromosome 8. In Table 1, we show the top 10 significant SNPs of it.

SNP ID	SNP information	P-value
rs138702093	8_63902958_G_GT_b37_GT	5.49E-11
rs16930060	8_63906199_G_A_b37_A	1.52E-10
rs16930066	8_63907144_A_G_b37_G	1.68E-10
rs62508123	8_63908088_C_A_b37_A	1.68E-10
rs16930072	8_63911296_A_G_b37_G	1.68E-10
rs58554293	8_63884058_G_A_b37_A	2.35E-10
rs111684515	8_63885412_G_A_b37_A	3.30E-10
rs113984807	8_63891319_G_A_b37_A	8.24E-10
rs62508154	8_63912068_G_C_b37_C	8.47E-10
rs62508156	8_63912174_C_G_b37_G	8.47E-10

Table 1. The top 10 significant SNPs with *p*-values

Folate Biosynthesis / Adipose Subcutaneous with GSVA Method

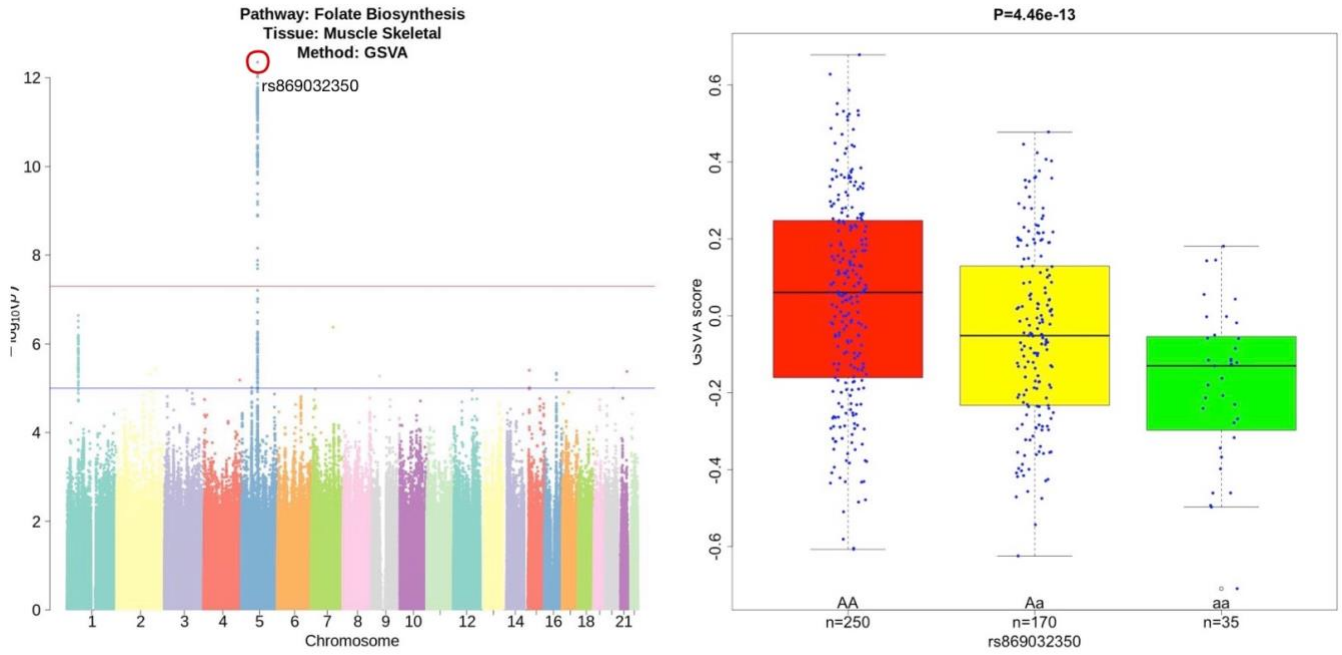

Figure 4. Manhattan Plot for the eQTL *p*-value with the boxplot of the peak SNP's genotype

Folate Biosynthesis / Muscle Skeletal with GSVA Method

In Figure 4, we find the SNP rs869032350 has the most significant *p*-value, 4.46×10^{-13}, which lies in the Chromosome 5. In Table 2, we show the top 10 significant SNPs of it.

SNP ID	SNP information	P-value
rs869032350	5_79902408_TTAATAA_T_b37_T	4.46E-13
rs1677670	5_79948654_A_T_b37_T	8.25E-13
rs1643662	5_79933235_A_G_b37_G	9.47E-13
rs1650688	5_79956129_G_A_b37_A	9.54E-13
rs1643646	5_79948641_C_T_b37_T	1.34E-12
rs1650692	5_79953393_C_T_b37_T	1.66E-12
rs1643652	5_79955079_G_A_b37_A	1.79E-12
rs1643645	5_79948540_G_A_b37_A	1.94E-12
rs860717	5_79947516_A_C_b37_C	1.99E-12
rs836819	5_79947763_G_A_b37_A	1.99E-12

Table 2. The top 10 significant SNPs with *p*-values

Folate Biosynthesis / Muscle Skeletal with GSVA Method

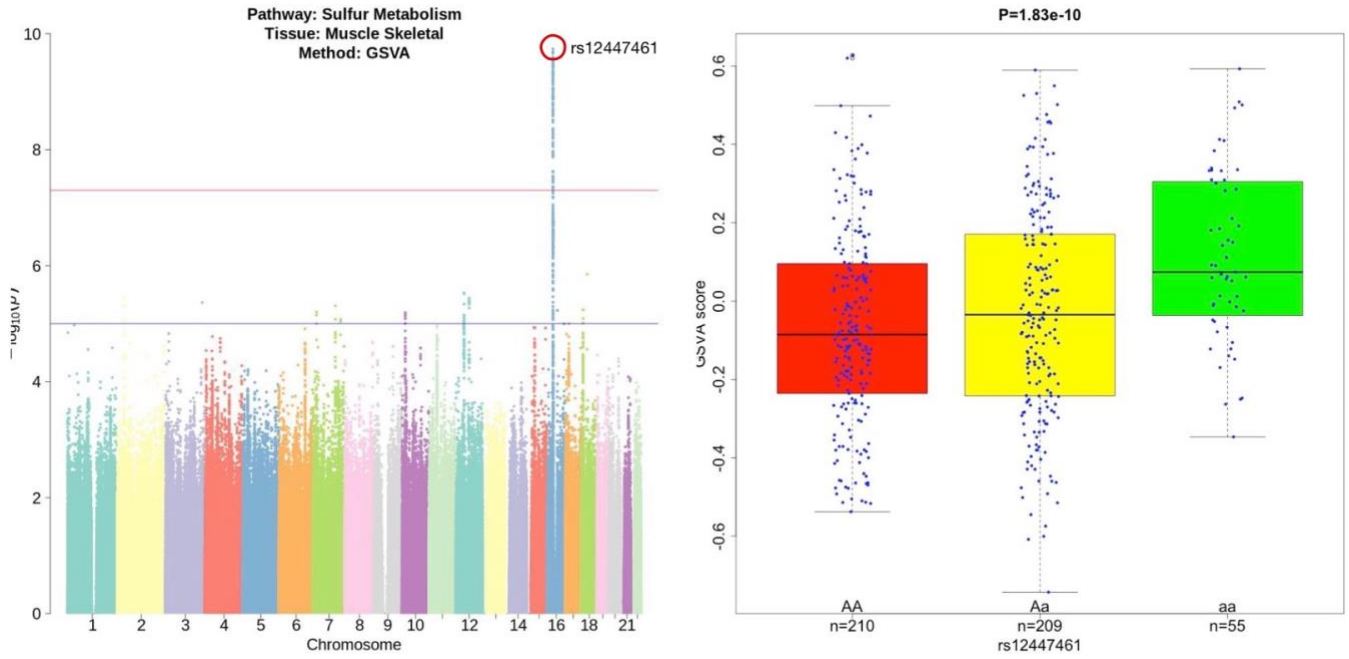

Figure 5. Manhattan Plot for the eQTL *p*-value with the boxplot of the peak SNP's genotype

Sulfur Metabolism / Muscle Skeletal with GSVA Method

In Figure 5, we find the SNP rs12447461 has the most significant *p*-value, 1.83×10^{-10}, which lies in the Chromosome 16. In Table 3, we show the top 10 significant SNPs of it.

SNP ID	SNP information	P-value
rs12447461	16_28582941_C_A_b37_A	1.83E-10
rs4788069	16_28616665_C_A_b37_A	2.06E-10
rs41278156	16_28618037_T_G_b37_G	2.18E-10
rs4788068	16_28616723_C_T_b37_T	2.54E-10
rs2925630	16_28619132_T_C_b37_C	2.92E-10
rs111384198	16_28617934_T_C_b37_C	3.01E-10
rs2925623	16_28618446_T_C_b37_C	3.04E-10
rs74459546	16_28617888_C_T_b37_T	3.09E-10
rs7191548	16_28614734_T_C_b37_C	3.15E-10
rs116840534	16_28617890_A_G_b37_G	3.29E-10

Table 3. The top 10 significant SNPs with *p*-values

Sulfur Metabolism / Muscle Skeletal with GSVA Method

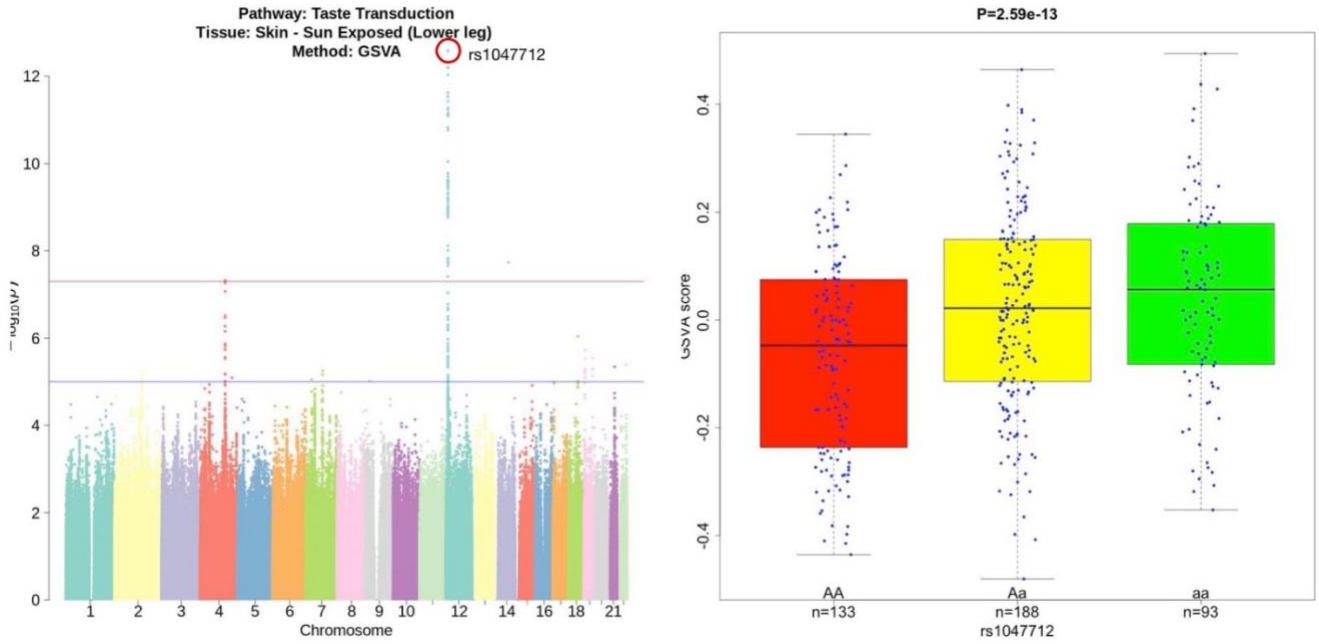

Figure 6. Manhattan Plot for the eQTL *p*-value with the boxplot of the peak SNP's genotype

Taste Transduction / Skin - Sun Exposed (Lower leg) with GSVA Method

In Figure 6, we find the SNP rs1047712 has the most significant *p*-value, 2.59×10^{-13}, which lies in the Chromosome 12. In Table 4, we show the top 10 significant SNPs of it.

SNP ID	SNP information	P-value
rs1047712	12_11324364_G_A_b37_G	2.59E-13
rs2416549	12_11325804_G_A_b37_G	2.62E-13
rs2416548	12_11324176_C_A_b37_C	6.15E-13
rs1047709	12_11324344_C_T_b37_C	6.15E-13
rs7488095	12_11323939_G_C_b37_G	6.39E-13
rs7488102	12_11323994_G_A_b37_G	6.39E-13
rs1863848	12_11328653_C_T_b37_C	9.24E-13
rs7350611	12_11331094_T_C_b37_T	2.40E-12
rs6488357	12_11331973_C_A_b37_C	2.40E-12
rs71057704	12_11337456_A_AACAAAC_b37_A	2.93E-12

Table 4. The top 10 significant SNPs with *p*-values

Taste Transduction / Skin - Sun Exposed (Lower leg) with GSVA Method

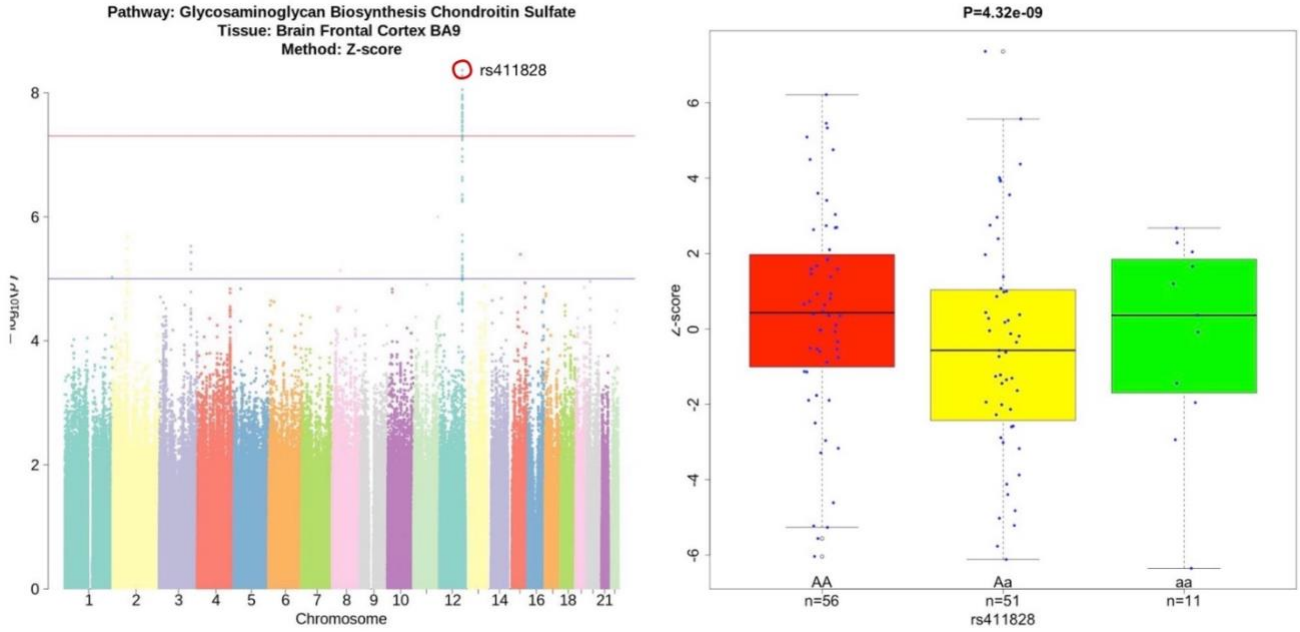

Figure 7. Manhattan Plot for the eQTL *p*-value with the boxplot of the peak SNP's genotype Glycosaminoglycan Biosynthesis Chondroitin Sulfate / Brain Frontal Cortex BA9 with Z-score Method

In Figure 7, we find the SNP rs411828 has the most significant *p*-value, 4.32×10^{-9}, which lies in the Chromosome 12. In Table 5, we show the top 10 significant SNPs of it.

SNP ID	SNP information	P-value
rs411828	12_118741714_A_G_b37_G	4.32E-09
rs428073	12_118682751_C_T_b37_C	5.29E-09
rs795480	12_118617641_C_G_b37_C	5.66E-09
rs811247	12_118622836_G_T_b37_G	5.66E-09
rs353895	12_118633715_T_C_b37_T	5.66E-09
rs1726390	12_118645384_A_G_b37_A	5.66E-09
rs1699160	12_118756336_C_A_b37_C	8.92E-09
rs795479	12_118618027_G_A_b37_G	1.08E-08
rs464781	12_118689816_T_C_b37_T	1.11E-08
rs459229	12_118684748_A_C_b37_A	1.17E-08

Table 5. The top 10 significant SNPs with *p*-values Glycosaminoglycan Biosynthesis Chondroitin Sulfate / Brain Frontal Cortex BA9 with Z-score Method

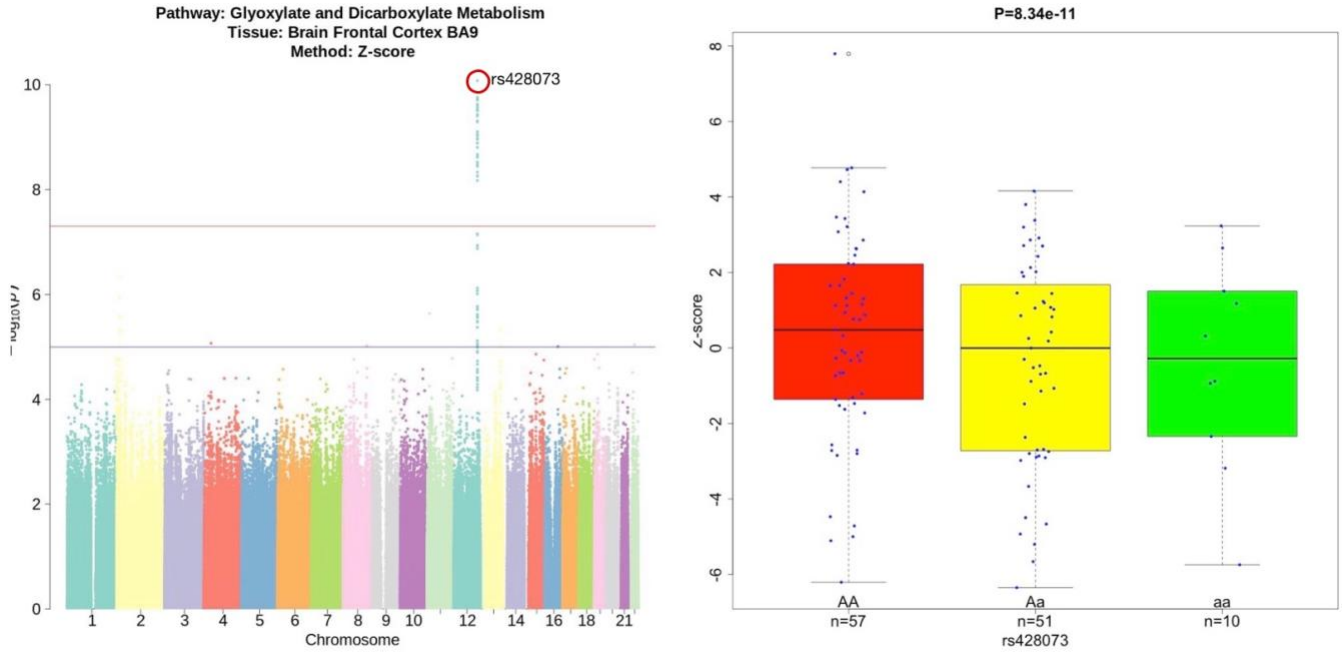

Figure 8. Manhattan Plot for the eQTL *p*-value with the boxplot of the peak SNP's genotype Glyoxylate and Dicarboxylate Metabolism / Brain Frontal Cortex BA9 with Z-score Method

In Figure 8, we find the SNP rs428073 has the most significant *p*-value, 8.34×10^{-11}, which lies in the Chromosome 12. In Table 6, we show the top 10 significant SNPs of it.

SNP ID	SNP information	P-value
rs428073	12_118682751_C_T_b37_C	8.34E-11
rs464227	12_118722997_G_A_b37_A	1.74E-10
rs795480	12_118617641_C_G_b37_C	1.90E-10
rs811247	12_118622836_G_T_b37_G	1.90E-10
rs353895	12_118633715_T_C_b37_T	1.90E-10
rs1726390	12_118645384_A_G_b37_A	1.90E-10
rs1726392	12_118598925_A_G_b37_A	1.94E-10
rs1277441	12_118605989_G_A_b37_G	1.94E-10
rs2454757	12_118668359_T_C_b37_T	2.33E-10
rs1151900	12_118686282_A_G_b37_A	2.43E-10

Table 6. The top 10 significant SNPs with *p*-values Glyoxylate and Dicarboxylate Metabolism / Brain Frontal Cortex BA9 with Z-score Method

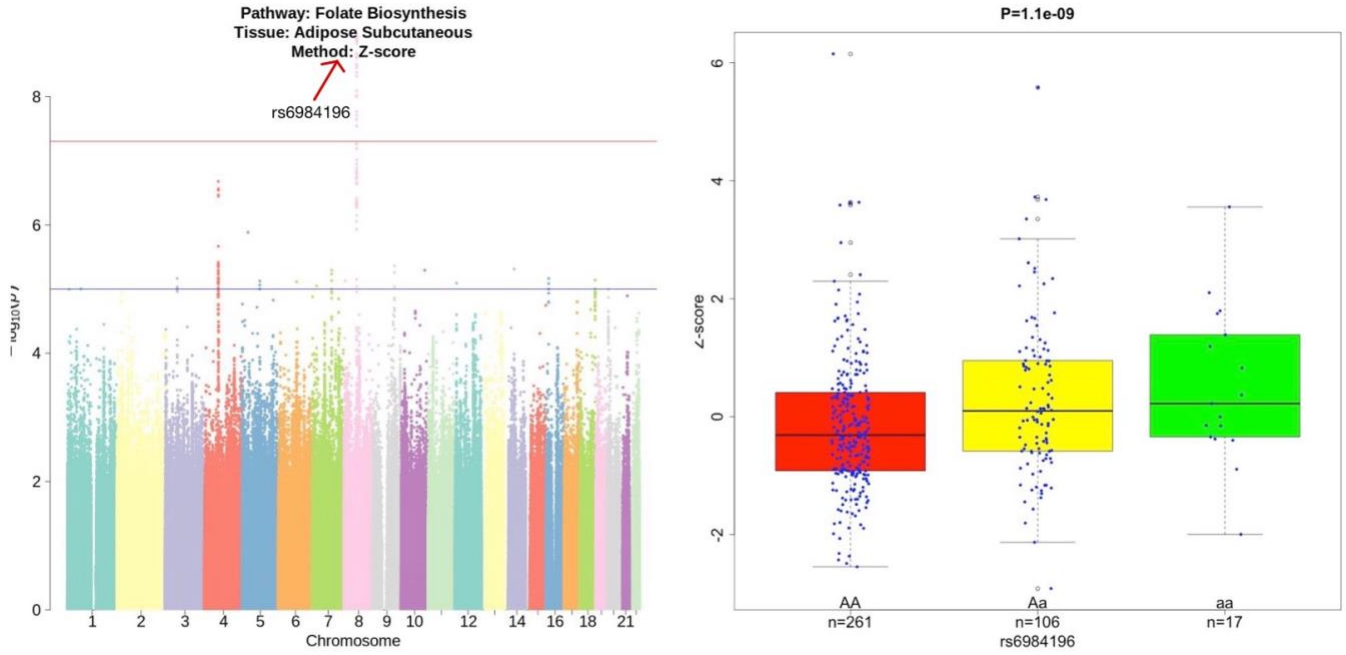

Figure 9. Manhattan Plot for the eQTL p-value with the boxplot of the peak SNP's genotype

Folate Biosynthesis / Adipose Subcutaneous with Z-score Method

In Figure 9, we find the SNP rs6984196 has the most significant *p*-value, 1.10×10^{-9}, which lies in the Chromosome 8. In Table 7, we show the top 10 significant SNPs of it.

SNP ID	SNP information	P-value
rs6984196	8_63909975_G_T_b37_T	1.10E-09
rs60408006	8_63918525_T_C_b37_C	1.21E-09
rs7004312	8_63919102_G_C_b37_C	1.21E-09
rs16930070	8_63910434_T_C_b37_C	1.22E-09
rs16930062	8_63906216_A_G_b37_G	1.34E-09
rs62508120	8_63906382_G_T_b37_T	1.34E-09
rs60782904	8_63906763_A_T_b37_T	1.34E-09
rs1031553	8_63906828_G_A_b37_A	1.34E-09
rs62508122	8_63908069_C_G_b37_G	1.34E-09
rs76746695	8_63908240_A_AG_b37_AG	1.34E-09

Table 7. The top 10 significant SNPs with *p*-values

Folate Biosynthesis / Adipose Subcutaneous with Z-score Method

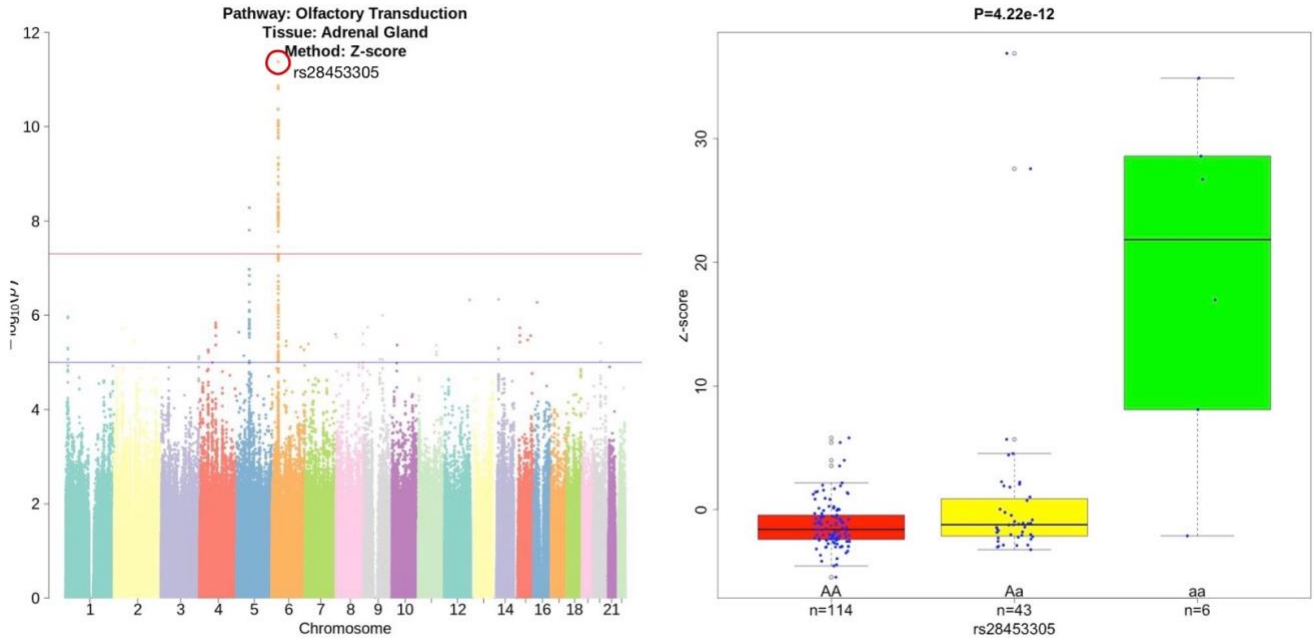

Figure 10. Manhattan Plot for the eQTL *p*-value with the boxplot of the peak SNP's genotype

Olfactory Transduction / Adrenal Gland with Z-score Method

In Figure 10, we find the SNP rs28453305 has the most significant *p*-value, 4.22×10^{-12}, which lies in the Chromosome 6. In Table 8, we show the top 10 significant SNPs of it.

SNP ID	SNP information	P-value
rs28453305	6_32501842_G_A_b37_G	4.22E-12
rs72849276	6_32469075_C_T_b37_C	1.34E-11
rs35350417	6_32513687_C_T_b37_C	1.51E-11
rs67020511	6_32513415_C_T_b37_C	1.58E-11
rs34072909	6_32513966_A_T_b37_A	4.25E-11
rs72849280	6_32469166_A_T_b37_A	4.28E-11
rs71545459	6_32469173_G_A_b37_G	4.28E-11
rs72844103	6_32513249_A_G_b37_A	7.33E-11
rs72844104	6_32513250_C_G_b37_C	7.33E-11
rs66717749	6_32513377_C_T_b37_C	8.35E-11

Table 8. The top 10 significant SNPs with *p*-values

Olfactory Transduction / Adrenal Gland with Z-score Method

In the boxplots above, we can see the monotone trend of the genotypes in most of these SNPs and the scatter points are normally distributed with the larger n. That provides solid evidence for the SNPs we selected.

To summarize, using GSVA and combined Z-score method working on pathways, we find some meaningful SNPs, for some specific pathways and tissues. The *p*-values for these SNPs are extraordinary from the other SNPs, showing the strong association between the genotype of these SNPs with these pathway-level scores. We list all the peak SNPs with the detail information in Table 9.

SNP ID	P-value	Chromosome	Pathway	Tissue	Method
rs138702093	5.49E-11	8	Folate Biosynthesis	Adipose Subcutaneous	GSVA
rs869032350	4.46E-13	5	Folate Biosynthesis	Muscle Skeletal	GSVA
rs12447461	1.83E-10	16	Sulfur Metabolism	Muscle Skeletal	GSVA
rs1047712	2.59E-13	12	Taste Transduction	Skin - Sun Exposed (Lower leg)	GSVA
rs411828	4.32E-09	12	Glycosaminoglycan Biosynthesis Chondroitin Sulfate	Brain Frontal Cortex BA9	Z-score
rs428073	8.34E-11	12	Glyoxylate and Dicarboxylate Metabolism	Brain Frontal Cortex BA9	Z-score
rs6984196	1.10E-09	8	Folate Biosynthesis	Adipose Subcutaneous	Z-score
rs28453305	4.22E-12	6	Olfactory Transduction	Adrenal Gland	Z-score

Table 9. Summary of the peak of the significant SNPs

Discussion

Compared the these two methods, we found that one pathway/tissue pair in both methods' selection. It is the pathway Folate Biosynthesis / Adipose Subcutaneous, and the significant SNPs are in Chromosome 8 at the same time, and the location for these SNPs are even close. In future research, we can analyze the association between the significant SNPs and the pathway, tissue.

Also, we can observe there are some Manhattan plots showing the second peak, some of them are even high than the genome-wide threshold 5×10^{-8}. In the future, we plan to follow up with these result.

Furthermore, there are some special Chromosomes that emerge several times. For instance, Chromosome 12 emerges three times in our analysis. We should find out the mutation frequency on this Chromosome and check out why this Chromosome appeared frequently.

We found some biologica evidence that can solid our findings. The SNPs found by GSVA or Z-score method to be associated with Folate Biosynthesis in Adipose, such as rs138702093 (GSVA) and rs6984196 (Z-score) are located in the exon of NKAIN3 gene, which is a coding SNP. The protein name of NKAIN3 is sodium/potassium transporting ATPase interacting 3. These SNPs are also shown as significant eQTLs pf the GGH gene which is a member of the Folate biosynthesis pathway and located near NKAIN3 gene. This shows that GGH gene may play a critical role inside the pathway. The SNPs associated with Folate Biosynthesis and Muscle

Skeletal are located in DHFR. Dihydrofolate reductase (DHFR) is a well-known enzyme of the folate metabolic pathway and it is a validated drug target for leishmaniasis. The SNP rs12447461 is located in gene NPIPL1, CCDC101, SGF29. eQTL of SULT1A1, SULT1A2, NUPR1, IL27, SH2B1. Rs1047712 near many TAS2R genes.

In addition, for Z-score method, the SNP rs428073 is located inside TAOK3 gene coding exon 5, which is a GWAS SNP (14, 15). Rs28453305 is eQTL of HLA-DRB5, HLA-DRB1, associated with the pathway Olfactory transduction. These above give solid evidence for the novel method we developed.

Besides, Yifan Han's work is using another two methods calculating pathway-score to find SNPs. We can have a comparison. If there are some SNPs are lied in common pairs of both our work, that should give strong evidence for this new method.

Reference:

1. Gatz, M., Pedersen, N. L., Berg, S., Johansson, B., Johansson, K., Mortimer, J. A., . . . Ahlbom, A. (1997). Heritability for Alzheimer's disease: the study of dementia in Swedish twins. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 52*(2), M117-M125.
2. Wingo, T. S., Lah, J. J., Levey, A. I., & Cutler, D. J. (2012). Autosomal recessive causes likely in early-onset Alzheimer disease. *Archives of neurology, 69*(1), 59-64.
3. Consortium, G. T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science, 348*(6235), 648-660. doi:10.1126/science.1262110
4. Carithers, L. J., & Moore, H. M. (2015). The genotype-tissue expression (GTEx) project. In: Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.
5. Hormozdiari, F., Van De Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., . . . Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics, 99*(6), 1245-1260.
6. Ratnapriya, R., Sosina, O. A., Starostik, M. R., Kwicklis, M., Kapphahn, R. J., Fritsche, L. G., . . . Pietraszkiewicz, A. (2019). Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nature genetics, 51*(4), 606-610.
7. Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., . . . Aguet, F. (2018). Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics, 50*(7), 956-967.
8. Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research, 28*(1), 27-30.
9. Du, J., Yuan, Z., Ma, Z., Song, J., Xie, X., & Chen, Y. (2014). KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Molecular bioSystems, 10*(9), 2441-2447.
10. Hänzelmann, S., Castelo, R., & Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics, 14*(1), 7.
11. Guinney, J., & Guinney, M. J. (2013). Package 'GSVA'.
12. Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics, 28*(10), 1353-1358.
13. Wigginton, J. E., Cutler, D. J., & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. The American Journal of Human Genetics, 76(5), 887-893.
14. Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., ... & Lambourne, J. J. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell, 167*(5), 1415-1429.
15. Gateva, V., Sandling, J. K., Hom, G., Taylor, K. E., Chung, S. A., Sun, X., ... & Gunnarsson, I. (2009). A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus. *Nature genetics, 41*(11), 1228.