Quantitative Biology 2021, 9(2): 151–167
https://doi.org/10.1007/s40484-020-0210-9

REVIEW

The statistical practice of the GTEx Project: from single to multiple tissues

Xu Liao[1,†], Xiaoran Chai[2,3,†], Xingjie Shi[1,4], Lin S. Chen[5], Jin Liu[1,*]

[1] Center for Quantitative Medicine, Health Services & Systems Research, Duke-NUS Medical School, Singapore 169857, Singapore
[2] Beijing Advanced Innovation Center for Genomics (ICG), Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, Peking University, Beijing 100871, China
[3] Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore
[4] Department of Statistics, Nanjing University of Finance and Economics, Nanjing 210023, China
[5] Department of Public Health Sciences, The University of Chicago, Chicago, IL 60637, USA
* Correspondence: jin.liu@duke-nus.edu.sg

Received January 21, 2020; Revised March 18, 2020; Accepted March 24, 2020

***Background*: The Genotype-Tissue Expression (GTEx) Project has collected genetic and transcriptome profiles from a wide spectrum of tissues in nearly 1,000 ceased individuals, providing an opportunity to study the regulatory roles of genetic variants in transcriptome activities from both cross-tissue and tissue-specific perspectives. Moreover, transcriptome activities (*e.g.*, transcript abundance and alternative splicing) can be treated as mediators between genotype and phenotype to achieve phenotypic alteration. Knowing the genotype associated transcriptome status, researchers can better understand the biological and molecular mechanisms of genetic risk variants in complex traits.**
***Results*: In this article, we first explore the genetic architecture of gene expression traits, and then review recent methods on quantitative trait locus (QTL) and co-expression network analysis. To further exemplify the usage of associations between genotype and transcriptome status, we briefly review methods that either directly or indirectly integrate expression/splicing QTL information in genome-wide association studies (GWASs).**
***Conclusions*: The GTEx Project provides the largest and useful resource to investigate the associations between genotype and transcriptome status. The integration of results from the GTEx Project and existing GWASs further advances our understanding of roles of gene expression changes in bridging both the genetic variants and complex traits.**

Keywords: the Genotype-Tissue Expression Project; quantitative trait loci (QTL); transcriptome-wide association studies; genome-wide association studies

Author summary: In the genetic area, people have made extensive efforts to investigate the associations between genetic variants and disease traits. However, we are lacking the knowledge of underlying biological mechanisms through which the genetic factors could affect the phenotypic outcome. Genotype-Tissue Expression (GTEx) Project provided us several angles to think about this question, including quantitative trait locus, alternative spicing patterns, and tissue-specific effect of genetic variants, and so on. In this article, we are providing a comprehensive review of their methods and results, and also suggest several down-stream analysis methods (*e.g.*, TWAS, co-expression network) by which we can go deeper into the regulatory mechanisms triggered by genetic factors.

[†] These authors contributed equally to this work.

INTRODUCTION

To characterize and interpret the function of genetic risk variants across the human genome remains essential in the post genome-wide association studies (GWASs). More efforts are required to interrogate the regulatory roles of genetic variants in complex traits [1–4]. There are several downstream processes that could be affected by genetic variants, such as DNA methylation, histone modification, transcription factor binding, isoform splicing, gene expression change. Usually, we can treat these processes as mediators between genetic variants and phenotype. Nowadays, several projects have been built to comprehensively study these mediators and annotate the functional elements in human genome across tissues and cell types, *e.g.*, ENCODE Project [5] and Roadmap Epigenomics Project [6], and the Genotype-Tissue Expression (GTEx) Project [7]. Both ENCODE and Roadmap projects primarily focus on whole blood or blood cell types due to the ease of accessibility and culturabilty. In contrast, the GTEx Project have collected a wide spectrum of tissues from ceased individuals and focused on investigating the associations between genotype and transcriptome status, including gene/isoform expression, allele specific expression, splicing pattern and others. Here, we use transcriptome QTL to represent this class of associations. In the most recent V8 release, the GTEx Project provides genetic profiles for 948 individuals and collects gene expression in 54 tissues from these individuals [8], as shown in Fig. 1.

Expression quantitative trait loci (eQTLs) are genomic loci that explain all or a fraction of variations in expression levels of mRNAs [9] and eQTL mapping offers a simple but powerful tool for identifying genetic variants which can affect gene expression [10]. Two major methods that could be used to conduct single-tissue eQTL analysis are Matrix eQTL [11] and FastQTL [12]. Several studies suggest that eQTLs are tissue-specific [13,14]. That means a genetic variant might have different regulatory effects on gene expression in different tissues. Multi-tissue eQTLs were not available until the launch of the GTEx Project. Compared to the single-tissue eQTLs,

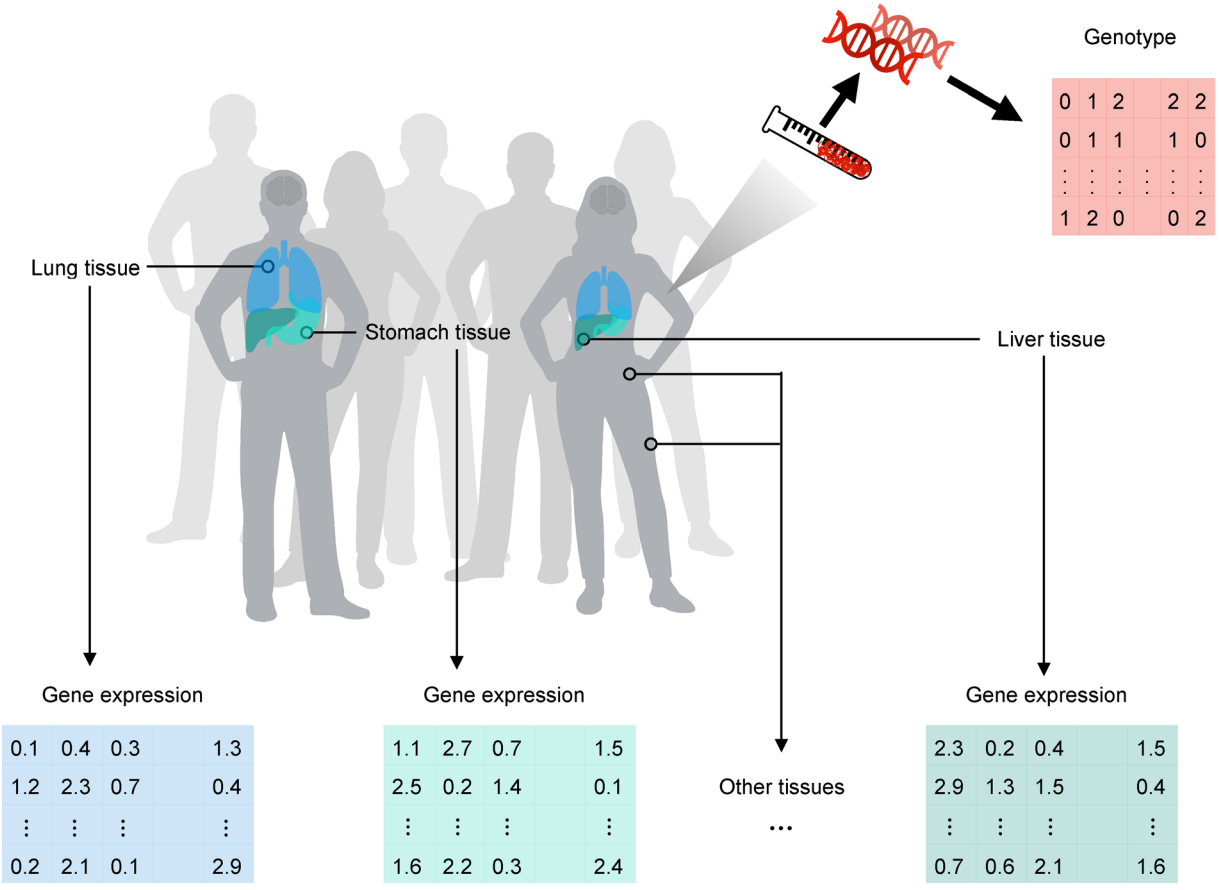


Figure 1. GTEx data framework. Data collection framework on the GTEx: genotypes and gene expression data across different tissue types, such as liver tissue, lung tissue, stomach tissue, are collected from donors.

eQTLs from multiple tissues provide us an opportunity to examine the influence of shared and tissue-specific regulatory effects of a single variant. More recently, a variety of methods have been proposed to analyze eQTLs from multiple tissues, *e.g.*, Meta-Tissue [15], MT-eQTL [16] and multivariate adaptive shrinkage (mash) [17]. Additionally, other types of transcriptome QTL can also be easily accessed in the GTEx data, *e.g.*, allele specific expression (ASE) and splicing quantitative trait locus (sQTL). Moreover, GTEx V8 release provides haplotype-level ASE as complementary to other eQTL data, where they reported a high correlation between the effect sizes of eQTL and those from both SNP-level and haplotype-level ASE [18].

Recent eQTL studies suggest that gene expression changes play a key role in bridging both the genetic variants and complex traits [19,20]. Aside from conventional eQTL analysis, multi-tissue data sets from the GTEx Project have been used as a resource to enhance the understanding of the genetic basis in complex traits either directly or indirectly. On the one hand, transcriptome-wide association studies (TWASs) were proposed to leverage the genetic regulatory information from eQTL directly in widely available GWAS results [21]. Many TWAS methods have been proposed to model the genetic effects on phenotype outcome through the gene expression based on either a single tissue or multiple tissues, *e.g.*, PrediXcan [21], TWAS [22], CoMM [23] for single-tissue TWAS, and MultiXcan [24], UTMOST [25], and TisCoMM [26] for multi-tissue TWAS. On the other hand, as genetic risk variants are not distributed equally across the genome, gene expression can serve as an indirect resource in GWAS analysis by putting more weights onto expression associated variants. Many methods have been proposed in this category including conditional FDR [27], GPA [28], EPS [29], and LSMM [30].

In this article, we first conduct empirical studies to demonstrate the genetic architecture of gene expression, showing the sparsity of gene expression traits and the heritability estimates across different tissues. We then briefly review single-tissue and multi-tissue eQTL methods followed by examining the gene co-expression network in the GTEx data. To illustrate the importance of leveraging transcriptome QTL information in GWASs, we review methods that integrate transcriptome information in GWASs either directly or indirectly. We briefly review PrediXcan, CoMM and TisCoMM to show the direct usage of eQTL information, where PrediXcan and CoMM is built for single-tissue analysis but TisCoMM is built for multi-tissue analysis as well as the examination of tissue specific effects. Finally, we briefly review GPA as an indirect use of eQTL information followed by a real data analysis using eSNPs in muscle skeletal tissue.

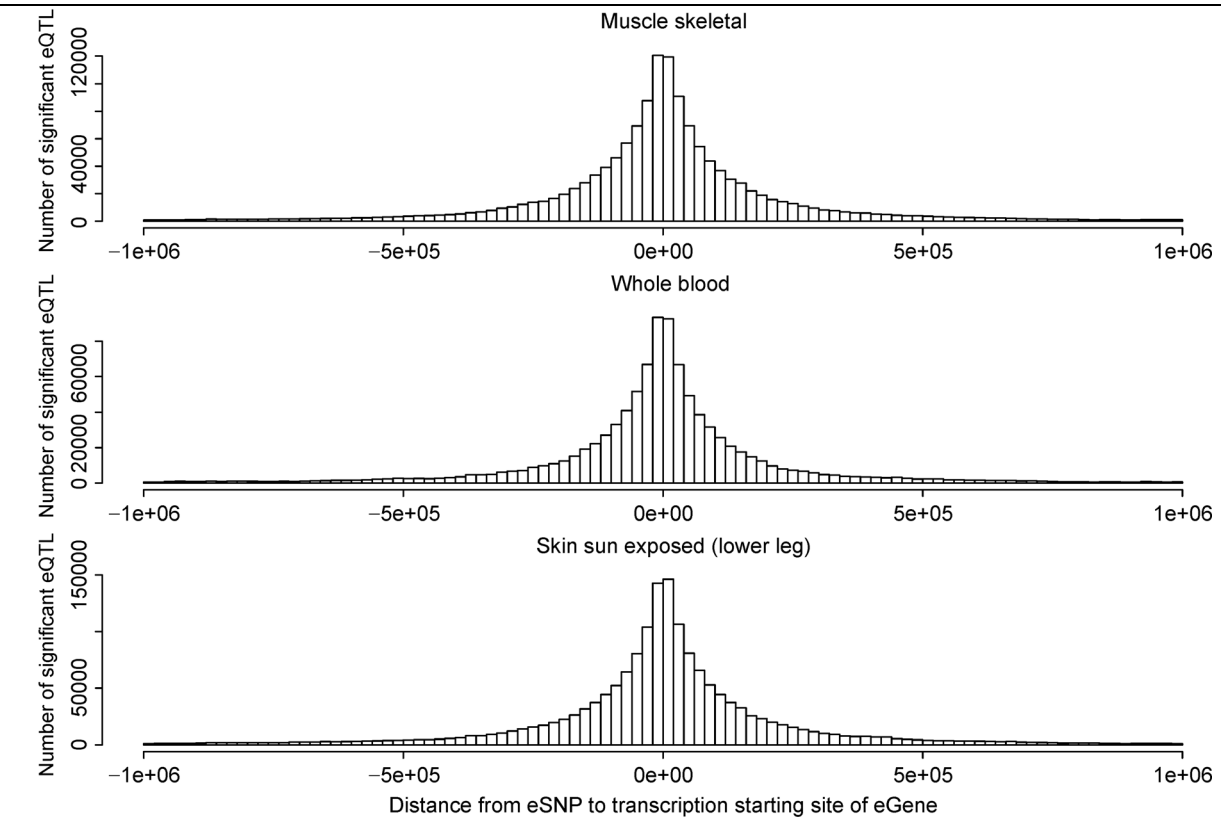
GENETIC ARCHITECTURE OF GENE EXPRESSION TRAITS

The GTEx data

The GTEx Project was launched in 2010 aiming to provide a resource to scientific community with which to examine human gene expression and how it could be regulated by genetic variation in multiple human tissues. This project collects biospecimens from 54 human tissues from nearly 1,000 postmortem donors who are both densely genotyped and subjected to RNA sequencing, and creates standards and protocols for optimizing postmortem tissue collection and donor recruitment [31,32], biospecimen processing [31], and data sharing (refer to the website:www.gtexportal.org/).

There are two types of eQTLs in terms of the distance between loci and target gene: cis-eQTLs (local eQTLs) and trans-eQTLs (distant eQTLs) [10], where cis-eQTLs are in close proximity (typically within 1 Mb) to the target gene and trans-eQTLs are typically located >1 Mb away from the target gene. Considering the relatively smaller sample sizes, most human eQTL studies have focused on how cis-eSNPs affect gene expression because of the reduced multiple testing burden. More importantly, the cis-SNPs located in the gene promoter region could have direct regulatory effects on the nearby gene, whereas the transSNPs may affect gene expression in an indirect manner. As we can see from the cis-eQTL results in GTEx, the number of significant eQTLs decreases with increasing distance between eSNP and its corresponding eGene (Fig. 2A). This phenomenon suggests that eSNPs are trend to be enriched in the promoter region of its related eGene.

To better understand the genetic architecture of gene expression traits, we conducted empirical studies for GTEx tissues from two perspectives, one from sparsity of eSNPs and the other from heritability of each gene across genome. Despite the fact that many complex traits are highly polygenic [4,33], the genetic architecture of gene expression traits are not well studied. Lots of variants with small effects contributing to gene expression variability is defined as polygenic architecture, while a small amount of variants with large effects contributing to variability is defined as sparse one. Following the survey [34], we use Bayesian sparse linear mixed model (BSLMM) [35] to evaluate the sparse and polygenic components of an expression trait. BSLMM assumes the genetic effects come from a mixture of two normal distributions, one for the sparse component and the other for the polygenic contributions. The effect sizes from the polygenic component is smaller than the ones from the sparse component. BayesR [36] extended the mixture of two components to the one with multiple components.



B

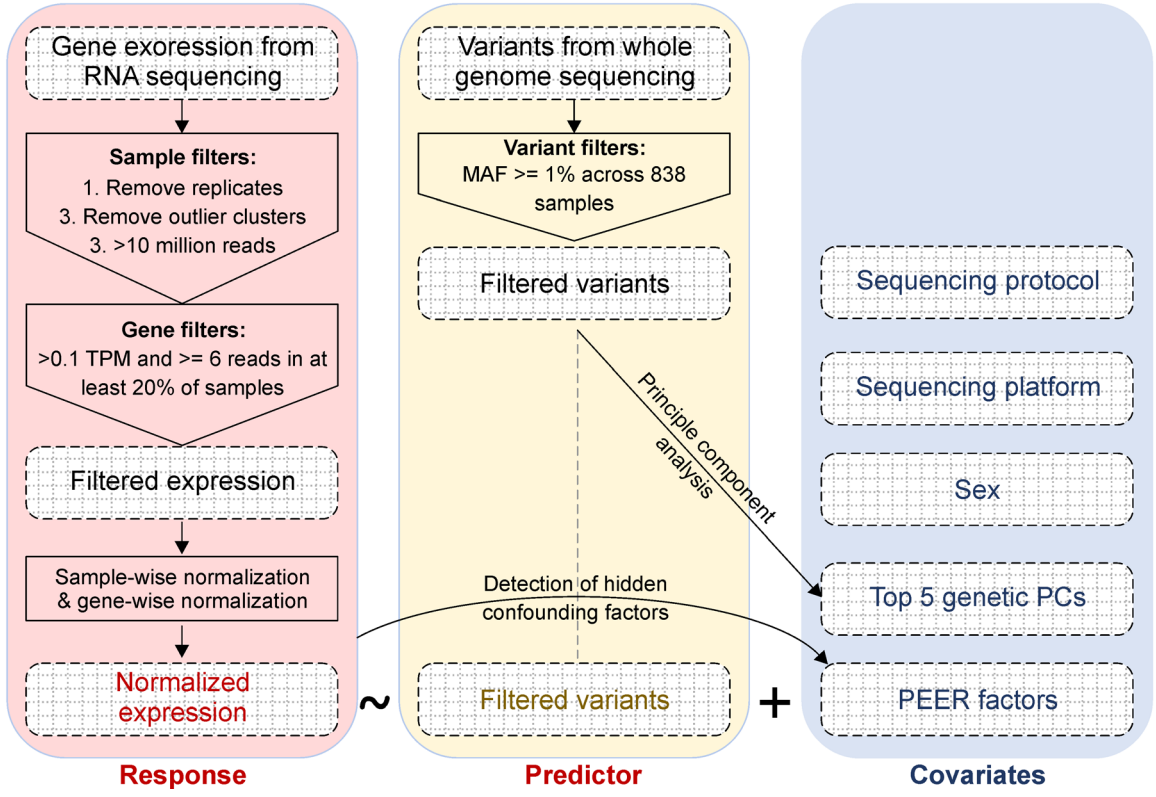


Figure 2. Distribution of cis-eQTL and workflow of FastQT. (A) Distributions of cis-eQTL in three GTEx tissues (muscle skeletal, whole blood and skin sun exposed) centered at the transcription starting site of the eGenes. Here we only plotted the distribution for SNPs and discarded all the INDELs. (B) Workflow of FastQTL.

Nonetheless, since it is of interest to explore the sparse structure of gene expression, we conducted the analysis using two components only. The linear mixed model (LMM) method estimates the heritability based on the assumption that effect sizes are normally distributed, which infers that estimated heritability of LMM represents the polygenic genetic architecture. For BSLMM, estimated heritability represents the sum of polygenic and sparse genetic architecture. Specifically, we conducted an analysis using both BSLMM and LMM to obtain estimates of heritability on 1,000 random selected genes across nine tissues. Figure 3 shows the violin plots of the estimated heritability for these 1,000 genes using both BSLMM and LMM. The estimates of heritability for BSLMM are higher than that for LMM across nine tissues, indicating that more sparse architecture may have more statistical power. Therefore, the genetic architecture of gene expression traits is more sparse rather than polygenic.

QTL discovery

The GTEx Project, in the most recent v8 release, provides DNA sequencing and other transcriptome measurements to conduct analysis for eQTL, allelic specific expression (ASE), and splicing QTL (sQTL). In this review, as the correlation between eQTL effect size and effect size measured using ASE data for a single genetic variant is quite large ($\rho = 0.838$) [18], we only focus on eQTL analysis. We note that the ASE data set from the GTEx Project is the largest to date and it makes haplotype-level data publicly available. It is, therefore, anticipated to have a larger impact on the understanding of regulatory variation across tissues in future studies.

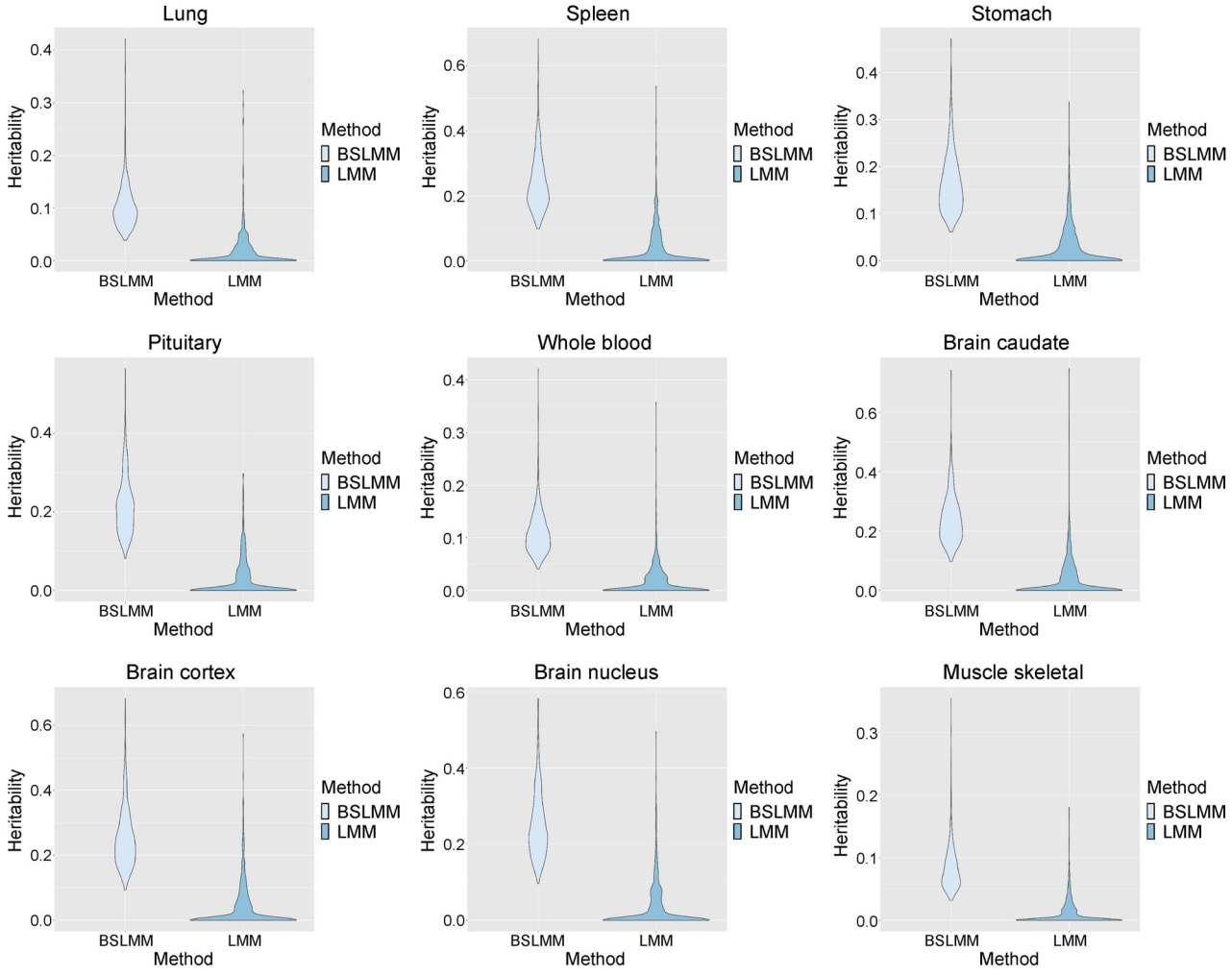


Figure 3. The sparse architecture of cis-eQTL in GTEx data. The comparison of estimated heritability between BSLMM and LMM by applying violin plots to show the sparse architecture of cis-eQTL in GTEx data.

eQTL analysis aims to identifying associations between genetic variants and gene expression across the genome [37]. Conventional eQTL analysis was performed by applying a large number of linear regressions across the genome, treating gene expression as the response variable and genotype as the predictor variable. A variety of methods have been developed to map QTL, including PANAMA [38], WASP [39], Matrix eQTL [11], FastQTL [12]. Before FastQTL was proposed, Matrix eQTL was considered as a gold standard that completes thousands of association tests in an acceptable amount of time.

Single-tissue based eQTL mapping

Due to an efficient implementation of linear regressions, less number of permutation steps and rapid data retrieval from indexed files, FastQTL was even faster and applied in many GTEx analysis [12]. Here, we roughly present the pipeline to conduct eQTL mapping and all the results are shared in GTEx portal (refer to the website: www.gtexportal.org/).

As illustrated in Fig. 2B, gene expressions and genetic variants are the two major inputs for performing FastQTL analysis as well as all the other eQTL mappings. Begin with the raw gene expressions calculated directly from RNA sequencing results, three sample filters were applied to remove the replicated samples, outlier samples, and samples with very low sequencing depth (<10 million reads). After that, one more gene filter was applied to keep the expressed genes only by requiring that gene expression $>$ 0.1 transcripts per million (TPM) and read count $\geqslant$ 6 in at least 20% of the samples. Then the sample-wise normalization was applied on the remaining gene expression values using weighted trimmed mean of M-values [40], and gene-wise normalization was performed using an inverse normal transformation. This normalized gene expressions were used as the response variables in the following FastQTL analysis. On the other hand, the second major input for eQTL mapping is genetic variants. Begin with the genotypes generated from whole genome sequencing results, only the common variants with minor allele frequency (MAF) $\geqslant$ 1% were kept and used as predictors in FastQTL. As we know that there exist a lot of confounding factors between genetic variants and gene expressions, a set of covariates were identified and incorporated into the linear model. Sequencing protocol and sequencing platform were included to correct for the batch effects, sex was included to correct for the gender effect, and top five genetic principle components (PCs) were included to correct for the population stratification. In addition to all of these, hidden confounding factors could be also identified from the normalized gene expression by PEER [41], and different number of PEER factors could be selected for different data sets according to their sample sizes.

The association between gene expression and genotype was determined by regressing the normalized expression on genotype and several confounding factors. As a result, a set of nominal association p-values was reported for each gene. It is known that the minimum nominal p-value for each gene follows a Beta distribution, and the distribution can be estimated by 1,000 permutations only [12]. Then the gene-level p-value could be obtained by locating the observed minimum p-value onto the estimated Beta distribution. In a traditional permutation test, we have to perform 10,000 permutations to get a p-value of 10^{-4}. However using the beta approximation in FastQTL, any p-value can be calculated using only 1,000 permutations. To correct for multiple testing, gene-level FDR was then calculated using Storey & Tibshirani correction [42]. Significant eGenes were selected if gene-level FDR $\leqslant$0.05. Gene-level p-value corresponding to FDR threshold 0.05 was then used as a global cut-off to fetch out all the significant eSNPs for each eGene. Specifically, for every eGene, its significant eSNPs were selected if the nominal association p-value $\leqslant F_{\beta}^{-1}$ (gene level p-value cut-off; α, β), where F_{β}^{-1} is the inverse function of Beta cumulative distribution, and α and β are two estimated parameters for Beta distribution.

In addition to cis-eQTLs, GTEx Project also reported trans-eQTL discoveries [42]. Focusing on the variants and genes located on different chromosomes or $\geqslant$5 Mb apart, they identified 126 trans-eGenes based on 10% FDR cut-off. By comparing across tissues, testis was highlighted as a more important tissue carrying most trans-eGenes. And another interesting observation of trans-eQTL is that it had higher tissue-specificity compared to ciseQTLs and thus are generally harder to replicate across studies.

Multi-tissue based eQTL mapping

Since the GTEx Project measured RNA sequencing for multiple postmortem tissues from donors, an ad-hoc analysis for such datasets is to perform single-tissue analysis across all tissues and then make comparisons for significant findings among them. However, such an analysis strategy does not make full use of the sharing patterns among multiple tissues and thus misses the power gains that come from sharing information at both common study individuals and shared cell types across tissues. Thus, performing joint multi-tissue eQTL that explicitly take advantages of sharing patterns is optimal to study both tissue-specific and shared patterns across tissues. Unlike linear models in tissue-by-tissue methods, Meta-Tissue [15] not only harnessed results from LMM to account for correlations of tissues from the overlapped samples and adjusted their effect sizes but also utilized the random-effects model to account for heterogeneity. On

the other hand, several methods have been proposed to allow for tissue-specific effects, sharing patterns among tissues, and heterogeneity in shared effects [16,17,43,44]. Among them, MT-eQTL used an empirical Bayes approach to explicitly model the patterns of effect sizes across multiple tissues and to perform inferences [16] while mash [17] is more flexible to combine the most attractive features of existing methods to improve effect estimates but overcome their major limitations. Overall, Table 1 summarizes the major methods for eQTL analysis either in single tissue or multiple tissues.

Multivariate adaptive shrinkage [17] is a computationally tractable method for dozens of tissues in the GTEx application by generating candidate covariance matrices in both data-driven and canonical ways to assess effect-size heterogeneity among tissues. In the empirical Bayesian framework, correlations and effects among different tissues are captured by a mixture model of multivariate normal distribution. The true effects for an eQTL across T tissues denoted by $\boldsymbol{\mu}$ are modeled with the mixture model,

$$p(\boldsymbol{\mu};\boldsymbol{\pi},U)=\sum_{k=1}^{K}\sum_{l=1}^{L}\pi_{k,l}N_T(\boldsymbol{\mu};\mathbf{0},\omega_l U_k), \qquad (1)$$

where $N_T(\cdot\,;\mathbf{0},\omega_l U_k)$ denotes the multivariate Gaussian density in T dimensions with mean $\mathbf{0}$ and variance-covariance matrix $\omega_l U_k$; $\boldsymbol{U}=(U_1,\ldots,U_K)$ are series of covariance matrices which can be divided into data-driven type and canonical type to capture patterns of effects; each ω_l is a coefficient to scale a effect size; the weight $\pi_{k,l}$ represents the contribution of each Gaussian component. Typically, $\boldsymbol{\omega}=(\omega_1,\ldots,\omega_L)$ are set values on a fixed dense grid from a small value to a large enough value.

In the model of mash, there are two major parameters, $\boldsymbol{U}$ and $\boldsymbol{\pi}$. The parameter $\boldsymbol{U}$ can be constructed not only in a data-driven way based on principal components analysis (PCA) and sparse factor analysis (SFA), but also in a canonical type by harnessing identity matrix, rank -1 matrix $\mathbf{1}\mathbf{1}^T$ and other typical matrices. With the generated parameter $\boldsymbol{U}$, $\boldsymbol{\pi}$ can be estimated by maximizing likelihood from tissue-by-tissue results. We implemented mash with eQTL results from GTEx V7 data based on FastQTL which are available on the GTEx Portal. The estimated effect sizes and corresponding standard errors of cis-eQTL from 16,851 genes across nine tissues (Fig. 4A) were the input of mash. For each gene, eQTL data with the largest absolute Z-Statistic was selected as the input of the model. Therefore, $16,851\times 9$ effect sizes and standard errors were used in our experiment. In order to show the comparison between mash and the single-tissue method (FastQTL in our experiment), we randomly selected 20 genes and presented their effect sizes and standard errors across nine tissues in Fig. 4A. According to Fig. 4A, it can be concluded that mash has a corrective effect on the input effect sizes and has a contraction effect on the standard errors of the input. Additionally, we further provide Fig. 4B and Fig. 4C, to illustrate the pairwise sharing patterns among nine tissues based on the effect sizes estimated from FastQTL and mash. Each cell of the heatmap demonstrates the pairwise proportion of effect sizes that is within factor 2 in size and in the same sign. The pairwise pattern for the original eQTL results is shown in Fig. 4B, and the pairwise proportions across tissues are not as large as that of mash results which are shown in Fig. 4C. The pairwise pattern obtained by mash is consistent with basic biological mechanisms. For example, the pairwise proportions among brain brain-caudate, brain-cortex, and brain-nucleus are close to 1, revealing different parts of the brain are highly correlated.

Co-expression network analysis

Reconstruction of gene co-expression networks is a powerful tool to better understand the co-regulation patterns among thousands of genes [45]. The network can be represented as undirected graphical models (UGM). Let nodes represent genes and connected nodes represent significantly corelated gene pairs. The co-expression network is not only useful in identifying gene-gene interactions, but also help us to build the possible regulatory pathways of trans-eSNPs. For example, if a trans-eSNP for gene A is also a cis-eSNP for transcription factor B, and A and B are co-expressed in

Table 1 Single-tissue and multi-tissue methods of eQTL analysis on the GTEx data

Approach	Method/algorithm	Languages
Single-tissue methods		
Matrix eQTL	Matrix multiplication	R, Matlab
FastQTL	Adaptive permutation	R, C ++
Multi-tissue methods		
Meta-tissue	Efficient mixed-model association (EMMA)	Java
MT-eQTL	Hierarchical Bayes	Matlab
mash	Empirical Bayes	R

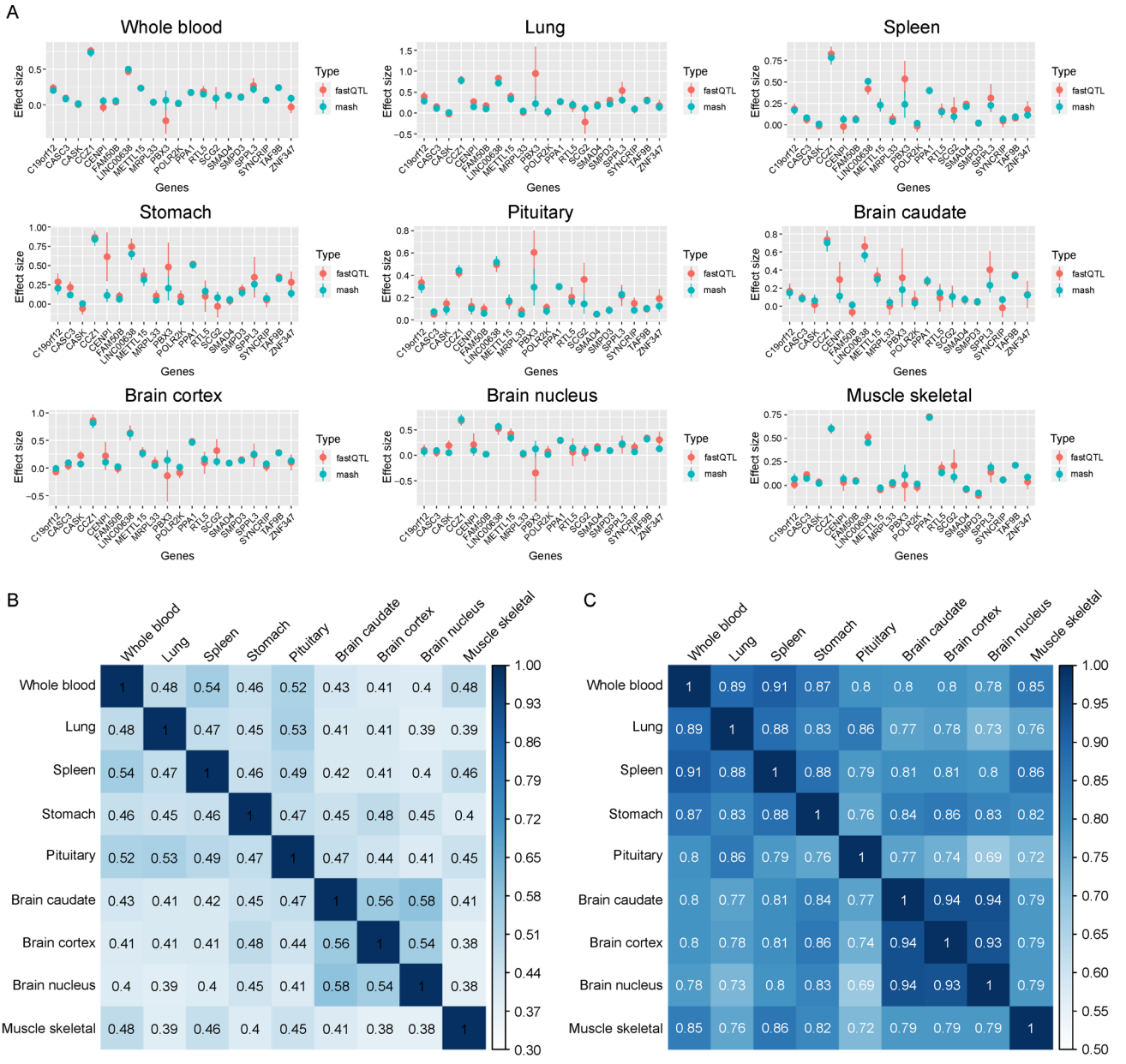


Figure 4. Sharing patterns among multiple tissues. (A) The comparison of effect sizes and standard errors between mash and FastQTL (the input for mash) across nine tissues. (B) Pairwise sharing of nine tissues according to the effect size estimated from original eQTL data. (C) Pairwise sharing of nine tissues according to the effect size estimated from mash. The element of heatmap represents the pairwise proportion of effect sizes which is within factor 2 and in the same sign.

the network, it is possible that the SNP affects the expression of A through the regulation of its transcription factor B.

WGCNA [46] is a popular R package that builds co-expression network based on genegene correlation. WGCNA provides construction for both weighted and unweighted networks with hard and soft threshold. The flowchart and illustration of gene co-expression network analysis using WGCNA can be found in the tutorial [47].

On the other hand, Gaussian graphical model is usually used to recover edges in UGM, *e.g.*, GeneNet [48] and graphical Lasso [49]. Gaussian graphical model assumes gene expression levels for each individual come from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. If the ij-th component of (Σ^{-1}) is zero, then the i-th gene and the j-th gene are conditionally independent. Denote K the inverse covariance matrix, also known as the precision matrix. Thus, the sparsity

level of precision matrix corresponds to the complexity of the network, sparser graphs indicates simpler networks with less edges in UGM. Graphical Lasso considers estimating sparse graphs by a Lasso penalty applied to the precision matrix [49]. Suppose we have n observations of dimension p (genes) from multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ. Graphical Lasso can be formulated as

$$\text{logdet}(K) - \text{tr}(SK) - \rho||K||_1, \qquad (2)$$

where S is the empirical covariance matrix, $\det(\cdot)$ denotes determinant of a square matrix, $\text{tr}(\cdot)$ is trace operator, and $||K||_1$ is the L_1 norm — the sum of the absolute values of the elements in Σ^{-1}.

Here, we performed the analysis of co-expression network for gene expression from muscle skeletal tissue in the GTEx data using both WGCNA and graphical Lasso. The adjacency heatmap and correlation heatmap for 100 random selected protein coding genes are shown in Fig. 5A and Fig. 5B. As shown in Fig. 5A, WGCNA can provide the visualization of gene dendrogram, module assignment and gene expression with heatmap where each cell represents topological overlap between two genes. Genes with the same color were designated as the same module by clustering. With the overlap score among genes, the co-expression network was established to investigate and leverage relationships among selected genes. The selected gene names and their corresponding modules are listed in Table 2. Comparison of correlation heatmaps with (the lower triangle) and without (the upper triangle) graphical Lasso regularization is shown in Fig. 5B. Due to the regularization of graphical Lasso, the correlation heatmap patttern is more sparse than the original one. In that case, utilizing graphical Lasso regularization is conducive to prune the whole co-expression network to find out the core genes (hub nodes in the network). To further explore the sharing and specific patterns of co-expression networks across multiple tissues in the GTEx data, several studies were conducted, including GNAT [50], eMAGMA [51] and others. We resort their original articles for the analysis results.

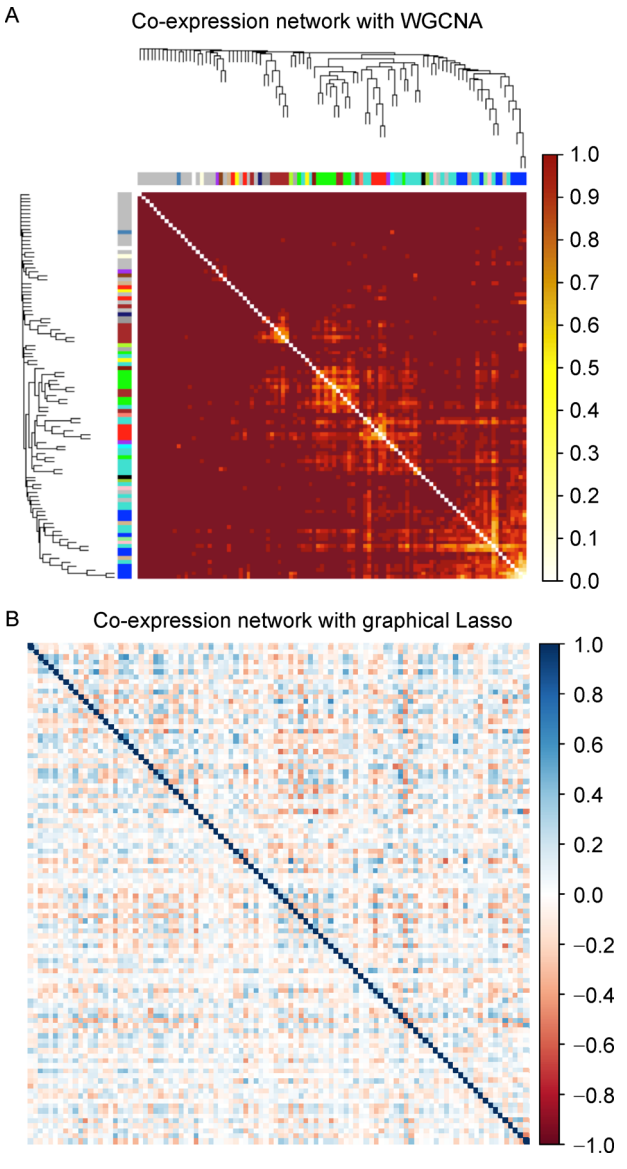


Figure 5. Co-expression network by applying WGCNA and graphical Lasso. (A) Gene dendrogram, module assignment and gene expression heatmap. In the cell of heatmaps, the darker color represents higher overlap between genes. (B) Correlation heatmaps with (the lower triangle) and without (the upper triangle) graphical Lasso regularization. To better visualize the heatmap, the overlap score was set to be zero for the same gene, and the correlation of the same gene was set to be 1.

ANALYSIS THAT LEVERAGES GENETICALLY REGULATORY INFORMATION

Direct use of transcriptome information in GWAS

Recent studies indicate the importance of gene expression changes in mediating the influence of genetic variants on complex traits. eQTL data sets from the GTEx Project can be taken as a reference data set to reflect the regulatory roles of genetic variants on gene expression in multiple tissues. Various methods for TWASs have been proposed to leverage the SNP-gene associations identified in a single tissue to infer significant gene-trait associations, *e.g.*, PrediXcan [21], TWAS [22], and CoMM [52]. To make full use of publicly available GWAS results calculated from large cohort studies, S-PrediXcan [53] and CoMM-S2 [23] extend PrediXcan and CoMM to take GWAS summary statistics as the input, respectively. The

major difference of CoMM and PrediXcan is that CoMM considers the imputation uncertainty in gene expression and thus improves the statistical power in an unified probabilistic model. To leverage the substantial sharing of eQTLs across tissues in the GTEx data sets, several methods have been proposed, *i.e.*, MultiXcan [24], UTMOST [25] and TisCoMM [26]. In the following subsections, we briefly discuss PrediXcan, CoMM and TisCoMM and resort their papers for technical details.

PrediXcan

Denote $\mathcal{D}_1 = \{\mathbf{Y},\mathbf{X}_1\}$ the reference eQTL data that contains expression matrix $\mathbf{Y} \in \mathbb{R}^{n_1 \times G}$ and genotype matrix $\mathbf{X}_1 \in \mathbb{R}^{n_1 \times M}$, where n_1 is the sample size, G is the number of genes, and M is the number of genotypes. Denote $\mathcal{D}_2 = \{\mathbf{z},\mathbf{X}_2\}$ the GWAS data that contains an $n_2 \times 1$ vector of phenotypic values $\mathbf{z}$ and genotype matrix $\mathbf{X}_2 \in \mathbb{R}^{n_2 \times M}$, where n_2 is the GWAS sample size, and M is the number of variants genotyped in the GWAS. Generally, in the TWAS, samples in $\mathcal{D}_1$ and $\mathcal{D}_2$ are disjoint and the genome-wide scan is performed for each gene. Suppose we work on gene g and let M_g be the number of variants within the g-th gene, $\mathbf{y}_g \in \mathbb{R}^{n_1 \times 1}$ is the g-th column of expression matrix $\mathbf{Y}$ representing the expression of in gene g across all the eQTL samples, and $\mathbf{X}_{1g} \in \mathbb{R}^{n_1 \times M_g}$ denotes the corresponding genotype matrix for this gene in the eQTL data set. Similarly, let $\mathbf{X}_{2g} \in \mathbb{R}^{n_2 \times M_g}$ denotes the corresponding genotype matrix for gene g in the GWAS data. Without loss of generality, we assume $\mathbf{z}$ denotes centered phenotypic vector, $\mathbf{X}_{1g}$ and $\mathbf{X}_{2g}$ denote standardized genotype data, and $\mathbf{y}_g$ denotes expression data adjusted for confounding factors. Basically, there are three steps in PrediXcan.

Table 2 Selected 100 gene names and corresponding modules

Module	Color	Count	Gene name
Module 1	Purple	2	*ORC3, ANAPC4*
Module 2	Grey	23	*LINGO4, ARHGAP10, CCDC18, WDR61, FAM49A, RAR-RES3, FGF17, RDH16, COX20, TRMT9B, CDYL2, NECAB2, KANTR, TERF2IP, CALHM6, GCFC2, TMEM68, MUC6, ETV2, RHBDF1, EIF1B, NHLH1, CATSPER1*
Module 3	Brown	9	*AHR, SRGAP1, RAB27A, CYP4F12, AFAP1, SH2D4A, GSN, MYMX, CHST11*
Module 4	Green	9	*PTPRN2, EMP3, CRISPLD1, TP53I3, B3GALNT1, P2RX7, PNMA8A, DLG3, SORCS1*
Module 5	Midnightblue	1	*CST7*
Module 6	Black	1	*RTL3*
Module 7	Greenyellow	1	*SURF1*
Module 8	Yellowgreen	1	*OIP5*
Module 9	Turquoise	19	*NUFIP2, CAVIN3, PET117, STAT5A, AQP11, ATG3, MAN2B1, MAFG, IKZF2, MOCS1, GID8, DDHD1, ZNF384, UBAP1L, NFX1, SLC19A2, LRIG2, KDM4C, CLASRP*
Module 10	Blue	10	*AES, RAB8B, CCND3, PKIB, GAK, NEK3, DYRK1B, FHIT, ACAA1, NME4*
Module 11	Yellow	2	*TMEM158, CACNG4*
Module 12	Lightyellow	1	*ABHD16A*
Module 13	Pink	2	*DERL3, SREK1*
Module 14	Darkred	1	*P2RX1*
Module 15	Saddlebrown	1	*EEF1E1*
Module 16	Grey60	2	*RASGRF2, ITGA6*
Module 17	Red	6	*POU2F1, C20orf24, SMG5, HDGF, GARNL3, PNMA1*
Module 18	Steelblue	1	*ATF2*
Module 19	Darkgrey	1	*ZNF865*
Module 20	Lightgreen	1	*ISOC1*
Module 21	Tan	3	*RBM20, ATP6V1E2, RCN2*
Module 22	Cyan	1	*SMARCA5*
Module 23	Salmon	1	*FADD*
Module 24	White	1	*ORAI3*

(a) Step 1: It uses elastic net to build the predictive model,

$$(\hat{\mu}_0,\hat{\boldsymbol{\mu}}) = \arg\min_{\mu_0,\boldsymbol{\mu}} ||\mathbf{y}_g - \mu_0 - \mathbf{X}_{1g}\boldsymbol{\mu}||^2 + \lambda\alpha||\boldsymbol{\mu}||_1 + \lambda(1-\alpha)||\boldsymbol{\mu}||_2^2, \quad (3)$$

where α = 0.5 (by default) and λ is tuned by cross-validation.

(b) Step 2: Gene expression levels for the individuals in the GWAS data are predicted as,

$$\hat{\mathbf{y}}_{2g} = \hat{\mu}_0 + \mathbf{X}_{2g}\hat{\boldsymbol{\mu}},$$

where $\mathbf{X}_{2g}$ is the corresponding genotype matrix in $\mathcal{D}_2$.

(c) Step 3: Conducting association analysis between $\mathbf{z}$ and $\hat{\mathbf{y}}_{2g}$ by simple linear regression,

$$\mathbf{z} = \beta_0 + \beta\hat{\mathbf{y}}_{2g} + \epsilon_z,$$

and standard statistical inference can be obtained for the parameter of interest, β.

PrediXcan has several advantages. First, compared with methods that identify differentially expressed genes using only transcriptome data, PrediXcan can be applied on GWAS data set, in which the sample size is several orders of magnitude larger than that of a transcriptome data set, *i.e.*, $n_2 \gg n_1$. Secondly, compared with single-variant-based approaches, PrediXcan reduces multiple testing burden from the order of 10^6 tests at the variant level to only 10^4 tests at the gene level. Thirdly, unlike indirect use of transcriptome information that enhances the identification of genetic associations between genetic variants and complex traits, PrediXcan is able to investigate the associations between genes and complex traits by leveraging transcriptome data.

CoMM

In trait-gene association scan, PrediXcan simply treated the predicted expression values $\hat{\mathbf{y}}_{2g}$ as if they were observed without error. This leads to underestimation of the coefficient β, known as the attenuation bias in measurement error models (MEM) [54]. To fix this issue, CoMM was proposed to jointly fit the predictive and association models in a principled manner. Focused on the g gene, the relationship between eQTL and GWAS data sets can be modeled using the following equations

$$\mathbf{y}_g = \mathbf{X}_{1g}\boldsymbol{\mu} + \boldsymbol{\epsilon}_1, \mathbf{z} = \beta\mathbf{X}_{2g}\boldsymbol{\mu} + \boldsymbol{\epsilon}_2, \quad (4)$$

where $\boldsymbol{\mu}$ is an $M_g \times 1$ vector of genetic effects on gene expression, $\boldsymbol{\epsilon}_1$ is an $n_1 \times 1$ vector of independent noises for gene expression, $\boldsymbol{\epsilon}_2$ is an $n_2 \times 1$ vector of independent noises for trait, and β is a scalar coefficient of interest representing the genetically regulated gene effect on phenotype. CoMM considers models (4) with a simple priori,

$$\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0},\sigma_\mu^2\mathbf{I}_{M_g}), \quad \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0},\sigma_{\epsilon_1}^2\mathbf{I}_{n_1}),$$

$$\boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0},\sigma_{\epsilon_2}^2\mathbf{I}_{n_2}), \quad (5)$$

where $\mathbf{I}_n$ denotes $n \times n$ identity matrix, and $\{\sigma_\mu^2,\sigma_{\epsilon_1}^2,\sigma_{\epsilon_2}^2\}$ is the set of unknown variance components to be estimated. To remove the effects from population stratification and other confounding factors, such as age and sex in the GWAS data, models (4) can be extended as,

$$\mathbf{y}_g = \mathbf{X}_{1g}\boldsymbol{\mu} + \boldsymbol{\epsilon}_1, \quad \mathbf{z} = \mathbf{W}\boldsymbol{\beta}_w + \beta\mathbf{X}_{2g}\mu + \boldsymbol{\epsilon}_2, \quad (6)$$

where $\mathbf{W}$ denotes the top genetic principal components (PCs) and other demographic factors, and β_w is a vector containing the coefficients for these covariates. To solve both models (4) and (6), a parameter-expanded expectation-maximization (PX-EM) algorithm [55] was developed to speed up the computation and the statistical inference for β was achieved by using the likelihood ratio test (LRT).

We conducted TWAS analysis on the GWAS data from the Northern Finland Birth Cohort 1966 (NFBC1966) and eQTL data from muscle-skeletal tissue in the GTEx Project using both PrediXcan and CoMM, where NFBC1966 data contains ten quantitative traits, *i.e.*, body mass index (BMI), triglyceride (TG), total cholesterol (TC), systolic blood pressure (SysBP), low-density lipoprotein cholesterol (LDL-C), insulin, high-density lipoprotein cholesterol (HDL-C), glucose, diastolic blood pressure (DiaBP), and c-reactive protein (CRP), and the qq-plots of p-values for these ten traits are shown in Fig. 6A. As shown in Fig. 6B, for most traits CoMM has higher statistical power than PrediXcan except insulin. The test statistics may degenerate to zero for CoMM in a few cases, causing the deflation phenomenon.

TisCoMM

Note that PrediXcan and CoMM can only be applied to eQTL data in a single tissue. To further increase the statistical power of target gene identification, multi-tissue TWAS methods have been proposed by integrating eQTL data from multiple tissues, *e.g.*, MultiXcan, UTMOST, and TisCoMM. The hypothesis testings performed in all these methods aim to prioritize gene-trait associations. Compared to MultiXcan and UTMOST, TisCoMM not only accounts for the imputation uncertainty, but also detects trait associated genes in a tissue specific manner. By conditioning on the correlated gene expression patterns in multiple trait-relevant tissues, false positive detections in single tissue analysis can be largely avoided using TisCoMM.

Here, we briefly introduce the TisCoMM method. Given the g-th gene, TisCoMM considers $\mathbf{Y}_g \in \mathbb{R}^{n_1 \times T}$ as the expression matrix for n_1 samples across T tissues for gene g, and $\mathbf{B}_g \in \mathbb{R}^{M_g \times T}$ as the corresponding coefficient matrix across T tissues. Other notations remain the same as those in CoMM. The model of TisCoMM can be written as

$$\mathbf{Y}_g = \mathbf{X}_{1g}\mathbf{B}_g + \mathbf{E}_g, \quad \mathbf{z} = \mathbf{X}_{2g}\mathbf{B}_g\boldsymbol{\alpha}_g + \boldsymbol{\epsilon}_z, \qquad (7)$$

where $\mathbf{E}_g \in \mathbb{R}^{n_1 \times T}$ is random error matrix from multivariate normal distribution $\mathcal{N}(0, \mathbf{V}_e)$, $\boldsymbol{\epsilon}_z \sim \mathcal{N}(0, \mathbf{I}_{n_2}\sigma^2)$ is an $n_2 \times 1$ vector of independent errors associated with the trait, and $\boldsymbol{\alpha}_g \in \mathbb{R}^{T \times 1}$ is an unknown parameter vector for the gene-trait effects among T tissues in gene g. Some techniques are utilized to enable TisCoMM computationally feasible, including both the factorization and an adaptive weighting strategy for $\mathbf{B}_g$. To extract regulatory information from all the relevant tissues, TisCoMM assumes that $\mathbf{B}_g$ is factorizable. To further make TisCoMM identifiable, an adaptive weighting strategy is applied. A parameter expanded EM algorithm was developed to solve the TisCoMM model and LRT is used to make statistical inference over the parameters of interest ($\boldsymbol{\alpha}_g$).

We implemented multi-tissue TWAS analyses for summary statistics GWAS data for peripheral vascular disease (PVD) from UK Biobank (the website, cnsgenomics.com/data.html). The transcriptome data is from the GTEx V7 data containing gene expression across six tissues (muscle-skeletal, lung, adipose subcutaneous, thyroid, artery tibial, and skin sunexposed lower leg). The reference panel are European subsamples from the 1000 Genomes Project. To compare the performance of TisCoMM-S^2, S-MultiXcan, and UTMOST, the qq-plot of these methods is shown in Fig. 6B. The Manhattan plot for p-values among gene-trait associations is presented in Fig. 6C.

Indirect use of transcriptome information in GWAS

TWAS-type analysis directly integrates SNP-gene associations evaluated in eQTL studies with SNP-trait associations evaluated in GWASs to prioritize risk genes for complex traits in a fashion closely related to instrumental variable analysis or Mendelian randomization [56]. Another way to use eQTL results from the GTEx data is to weight genetic variants according to their genomic regulatory effects across the genome. For example, we want to put more weights to the genetic variants which could have functional impacts on gene products.

Conventionally, GWAS analysis is conducted without incorporating any prior information. Recent studies show that the functional importance of genetic variants may not be equal [57] and significant GWAS signals are highly enriched in genomic regulatory regions such as promoters and enhancers in trait relevant cells or tissues [58]. Hence, post GWAS analysis that integrates genomic regulatory information cannot only efficiently increase the statistical power to prioritize risk variants but also identify trait-relevant tissues. To perform such analysis, QTLs (*e.g.*, eQTL or sQTL) from the GTEx data can be served as functional annotations to genetic variants indicating whether the variant can affect gene expression or splicing pattern. A variety of methods have been proposed to integrate variant annotations in this manner, including the ones to conduct SNP-based analysis, *e.g.*, PAINTOR [59], CAVIAR [60], fgwas [61], GPA [28], LSMM [30] and others, and the ones to conduct genebased analysis, *e.g.*, EPS [29]. Additionally, many efforts have been made to colocalize the signals either from QTLs and GWAS results or from different QTLs, *e.g.*, coloc [62], enloc [63], and moloc [64].

Genetic analysis incorporating pleiotropy and annotation (GPA)

GPA is a statistical method that integrates multiple GWAS analysis and functional annotations for genetic variants to prioritize GWAS signals. Suppose we have conducted genomewide hypothesis testing in a GWAS study for all M genetic variants and their corresponding p-values are denoted as p_m,

$$\text{Null hypothesis:} \quad H_0^{(1)}, ..., H_0^{(M)},$$
$$p\text{-values:} \quad p_1, ..., p_M.$$

We consider the "two-groups model" [65] and assume that the observed p-values come from the mixture of null and non-null distribution, with probability π_0 and $\pi_1 = 1 - \pi_0$, respectively. Let $z_m \in \{0, 1\}$ be the latent variable indicating the association status for the m-th variant, either under null or non-null distribution, respectively. Moreover, we assume the prior for the latent variable z_m, $\pi_0 = \Pr(z_m = 0)$ and $\pi_1 = \Pr(z_m = 1)$. Then conditioned on the latent variable z_m, the two-groups model assumes the distributions of p-values are $p_m|z_m = 0 \sim \mathcal{U}(0, 1)$ and $p_m|z_m = 1 \sim \mathcal{B}(\alpha, 1)$, where the p-values from the null group follows a uniform distribution, and p-values from the non-null group follows a Beta distribution with shape parameters α $(0 < \alpha < 1)$ and 1. To incorporate SNP annotations, GPA extends the above basic "two-groups" model as follows. Suppose we have annotation matrix $\mathbf{A} \in \mathbb{R}^{M \times T}$ for T functional annotation sources across all M genetic variants. The most straightforward way is to annotate genetic variants using binary indicators. In this case, $A_{mt} \in \{0, 1\}$ indicates whether genetic variant m is

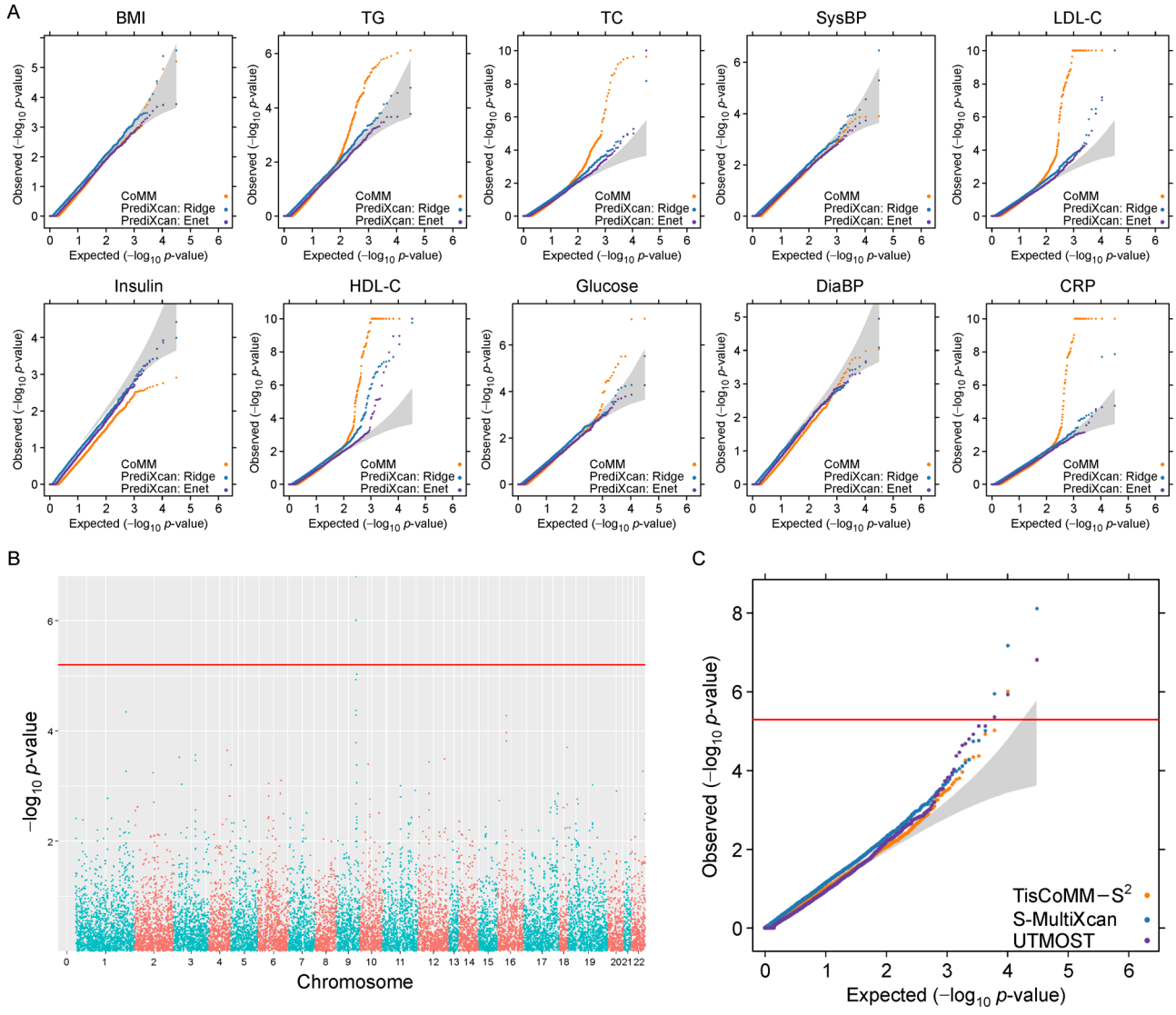


Figure 6. TWAS analysis of NFBC1966 data. (A) The qq-plots for ten quantitative traits in NFBC1966 with CoMM, PrediXcan: Ridge and PrediXcan: Enet by leveraging eQTL data of muscle-skeletal tissue in the GTEx Project; (B) and (C) The Manhattan plot for PVD using TisCoMM-S^2 and the qq-plot for PVD using TisCoMM-S^2, S-MultiXcan, and UTMOST.

Table 3 Single-tissue and multi-tissue methods of TWAS on the GTEx data

Method	Individual-level	Summary-statistics	Multi-tissue	Software
PrediXcan	√	×	×	https://github.com/hakyimlab/PrediXcan
S-PrediXcan	×	√	×	https://github.com/hakyimlab/MetaXcan
MultiXcan	√	√	√	https://github.com/hakyimlab/MetaXcan
UTMOST	√	√	√	https://github.com/Joker-Jerome/UTMOST
CoMM	√	×	×	https://github.com/gordonliu810822/CoMM
CoMM-S^2	×	√	×	https://github.com/gordonliu810822/CoMM
TisCoMM	√	√	√	https://github.com/XingjieShi/TisCoMM

annotated in the t-th annotation source. For example, if SNP j is a significant eSNP in liver tissue, we may annotate a genetic variant with A_{jt}= 1, otherwise we may assign 0 to A_{jt} indicating that it is not a significant eSNP in liver. Then the relationship between z_m and A_{jt} can be modeled as $A_{jt}|z_m = 0 \sim \text{Bernoulli}(q_{t0})$ and $A_{jt}|z_m =$

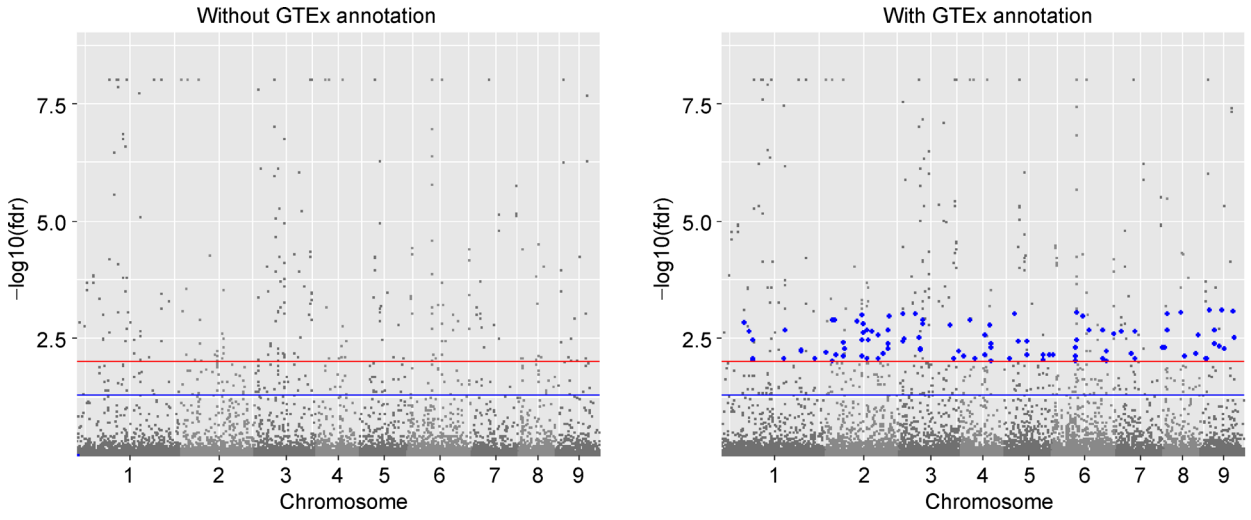


Figure 7. GPA analyses without and with annotation from eQTL in muscle skeletal tissue. The enhanced genetic variant signals were marked in blue.

$1 \sim \text{Bernoulli}(q_{t1})$, where q_{t0} and q_{t1} can be interpreted as the proportion of genetic variants being annotated in the t-th annotation for the null and non-null group, respectively. In this setting, the risk variants can be prioritized using the following posterior.

$$\Pr(\mathbf{z}|\mathbf{p},\mathbf{A}) = \frac{\Pr(\mathbf{p}|\mathbf{z})\Pr(\mathbf{z}|\mathbf{A})}{\sum_{\mathbf{z}}\Pr(\mathbf{p}|\mathbf{z})\Pr(\mathbf{z}|\mathbf{A})}. \quad (8)$$

In the framework of expectation-maximization (EM) algorithm, given an annotation tissue in the GTEx data, we may conduct hypothesis testing to examine whether the eQTLs identified in that tissue are enriched in some disease related GWAS hits: $H_0 : q_{t0} = q_{t1}$ vs. $H_1 : q_{t0} \neq q_{t1}$. More importantly, the integration of functional genomic annotation would assist identifying weak signals missed by the traditional single-variant analysis. We conducted GPA analysis on BMI of European Summary Statistics [66] with muscle skeletal tissue annotation. In details, we annotate a genetic variant as 1 if this variant was identified as an eSNP in eQTL analysis using muscle skeletal tissue in the GTEx; otherwise we annotate this variant as 0. Manhattan plots of local false discovery rate are shown in Fig. 7, where the enhanced genetic variants due to the incorporation of functional annotation were marked in blue.

CONCLUSION

The GTEx Project provides the largest resource for intra- and inter-individual genotype and transcriptome measurements across a spectrum of tissues. Its current V8 release contains 17,382 RNA-seq samples among 948 donors and also includes splicing quantitative trait loci (sQTL) in this version. We firstly explore the genetic architecture of gene expression traits, showing that cis-eSNP contribute a considerable amount of signals to the variation in expression and eSNP signals are sparse other than polygenic in expression traits. Secondly, the GTEx data forms a large genome and transcriptome database across multiple tissues in addition to whole blood, providing an opportunity to explore the tissue sharing patterns of eQTLs as well as its tissue-specific effects. Thirdly, investigators may explore the tissue sharing and tissue-specific patterns of co-expression network using the GTEx data.

Aside from transcriptome QTL and co-expression network analysis, the GTEx data can be used as a resource to leverage transcriptome QTL information directly or indirectly in genetic studies. TWAS has been widely used to integrate gene expression from eQTL studies with GWAS to prioritize trait-associated genes across the genome. Its prototype, PrediXcan, performs a step-wise analysis by conducting imputation for gene expressions and then performing subsequent association analysis. To account for uncertainty in the process of imputation, Yang *et al.* [52] proposed CoMM in a unified probabilistic model. To further explore the tissue-specific role of genes in complex traits, Shi *et al.* [26] proposed TisCoMM, a principled method to perform gene-trait joint and tissue-specific association tests across multiple tissues. Moreover, the transcriptome information across tissues can be utilized indirectly, as exemplified by GPA.

ACKNOWLEDGEMENTS

We would like to thank the two anonymous reviewers whose constructive comments have greatly improved this manuscript. This work was supported

by grant R-913-200-098-263 from the Duke-NUS Medical School, and AcRF Tier 2 (MOE2016-T2-2-029, MOE2018T2-1-046 and MOE2018-T2-2-006) from the Ministry of Education, Singapore. The computational work for this article was partially performed using resources from the National Supercomputing Centre, Singapore (https://www.nscc.sg).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Xu Liao, Xiaoran Chai, Xingjie Shi, Lin S. Chen and Jin Liu declare that they have no conflict of interests.

The article is a review article and does not contain any human or animal subjects performed by any of the authors.

REFERENCES

1. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R., Anttila, V., Xu, H., Zang, C., Farh, K., *et al.* (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet., 47, 1228–1235
2. Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science, 337, 1190–1195
3. Nica, A. C., Montgomery, S. B., Dimas, A. S., Stranger, B. E., Beazley, C., Barroso, I. and Dermitzakis, E. T. (2010) Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet., 6, e1000895
4. Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A. and Yang, J. (2017) 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet., 101, 5–22
5. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74
6. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. Nature, 518, 317–330
7. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. Nat. Genet., 45, 580–585
8. Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., *et al.* (2019) The GTEX consortium atlas of genetic regulatory effects across human tissues. bioRxiv, 787903
9. Rockman, M. V. and Kruglyak, L. (2006) Genetics of global gene expression. Nat. Rev. Genet., 7, 862–872
10. Gilad, Y., Rifkin, S. A. and Pritchard, J. K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet., 24, 408–415
11. Shabalin, A. A. (2012) Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics, 28, 1353–1358
12. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. and Delaneau, O. (2016) Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics, 32, 1479–1485
13. Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T. P., Meduri, E., Barrett, A., *et al.* (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet., 44, 1084–1089
14. Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., *et al.* (2010) New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. PLOS Comput. Biol., 6, e1000737
15. Sul, J. H., Han, B., Ye, C., Choi, T. and Eskin, E. (2013) Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. PLoS Genet., 9, e1003491
16. Li, G., Shabalin, A. A., Rusyn, I., Wright, F. A. and Nobel, A. B. (2018) An empirical Bayes approach for multiple tissue eQTL analysis. Biostatistics, 19, 391–406
17. Urbut, S. M., Wang, G., Carbonetto, P. and Stephens, M. (2019) Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat. Genet., 51, 187–195
18. Castel, S. E., Aguet, F., Mohammadi, P., GTEx Consortium, Ardlie, K. G., Lappalainen, T. (2019) A vast resource of allelic expression data spanning human tissues. bioRxiv, 792911
19. Albert, F. W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. Nat. Rev. Genet., 16, 197–212
20. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. and Lathrop, M. (2009) Mapping complex disease traits with global gene expression. Nat. Rev. Genet., 10, 184–194
21. Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., *et al.* (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet., 47, 1091–1098
22. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., de Geus, E. J., Boomsma, D. I., Wright, F. A., *et al.* (2016) Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet., 48, 245–252
23. Yang, Y., Shi, X., Jiao, Y., Huang, J., Chen, M., Zhou, X., Sun, L., Lin, X., Yang, C. and Liu, J. (2019) CoMM-S^2: a collaborative mixed model using summary statistics in transcriptome-wide association studies. bioRxiv, 652263
24. Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L. and Im, H. K. (2019) Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet., 15, e1007889
25. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M., Yu, Z., Li, B., Gu, J., Muchnik, S., Shi, Y., *et al.* (2019) A statistical framework for cross-tissue transcriptome-wide association analysis. Nat. Genet. 51, 568–576
26. Shi, X., Chai, X., Yang, Y., Cheng, Q., Jiao, Y., Huang, J., Yang, C. and Liu, J. (2019) A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. bioRxiv, 789396
27. Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S.,

Mattingsdal, M., Kelsoe, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., *et al.* (2013) Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional false discovery rate. PLoS Genet., 9, e1003455

28. Chung, D., Yang, C., Li, C., Gelernter, J. and Zhao, H. (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet., 10, e1004787
29. Liu, J., Wan, X., Ma, S. and Yang, C. (2016) EPS: an empirical Bayes approach to integrating pleiotropy and tissue-specific information for prioritizing risk genes. Bioinformatics, 32, 1856–1864
30. Ming, J., Dai, M., Cai, M., Wan, X., Liu, J. and Yang, C. (2018) LSMM: a statistical approach to integrating functional annotations with genome-wide association studies. Bioinformatics, 34, 2788–2796
31. Carithers, L. J., Ardlie, K., Barcus, M., Branton, P. A., Britton, A., Buia, S. A., Compton, C. C., DeLuca, D. S., Peter-Demchok, J., Gelfand, E. T., *et al.* (2015) A novel approach to high-quality postmortem tissue procurement: the GTEX project. Biopreserv. Biobank., 13, 311–319
32. Siminoff, L. A., Wilson-Genderson, M., Gardiner, H. M., Mosavel, M. and Barker, K. L. (2018) Consent to a postmortem tissue procurement study: distinguishing family decision makers' knowledge of the genotype-tissue expression project. Biopreserv. Biobank., 16, 200–206
33. The International Schizophrenia Consortium (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 460, 748–752
34. Wheeler, H. E., Shah, K. P., Brenner, J., Garcia, T., Aquino-Michaels, K., Cox, N. J., Nicolae, D. L., Im, H. K., and the GTEx Consortium. (2016) Survey of the heritability and sparse architecture of gene expression traits across human tissues. PLoS Genet., 12, e1006423
35. Zhou, X., Carbonetto, P. and Stephens, M. (2013) Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet., 9, e1003264
36. Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R. and Visscher, P. M. (2015) Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet., 11, e1004969
37. Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E. and Cox, N. J. (2010) Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet., 6, e1000888
38. Fusi, N., Stegle, O. and Lawrence, N. D. (2012) Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. PLOS Comput. Biol., 8, e1002330
39. van de Geijn, B., McVicker, G., Gilad, Y. and Pritchard, J. K. (2015) Wasp: allele-specific software for robust molecular quantitative trait locus discovery. Nat. Methods, 12, 1061–1063
40. Robinson, M. D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol., 11, R25
41. Stegle, O., Parts, L., Durbin, R. and 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
42. The GTEx Consortium (2017) Genetic effects on gene expression across human tissues. Nature, 550, 204–213
43. Flutre, T., Wen, X., Pritchard, J. and Stephens, M. (2013) A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet., 9, e1003486
44. Wei, Y., Tenzen, T. and Ji, H. (2015) Joint analysis of differential gene expression in multiple studies using correlation motifs. Biostatistics, 16, 31–46
45. Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol., 4, e17
46. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559
47. Langfelder, P. and Horvath, S. (2014) Tutorials for the WGCNA package
48. Ananko, E. A., Podkolodny, N. L., Stepanenko, I. L., Ignatieva, E. V., Podkolodnaya, O. A. and Kolchanov, N. A. (2002) Genenet: a database on structure and functional organisation of gene networks. Nucleic Acids Res., 30, 398–401
49. Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9, 432–441
50. Pierson, E., the GTEx Consortium, Koller, D. and Battle, A. (2015) Sharing and specificity of co-expression networks across 35 human tissues. PLOS Comput. Biol., 11, e1004220
51. Gerring, Z. F., Gamazon, E. R., Derks, E. M., the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2019) A gene co-expression networkbased analysis of multiple brain tissues reveals novel genes and molecular pathways underlying major depression. PLoS Genet 15, e1008245
52. Yang, C., Wan, X., Lin, X., Chen, M., Zhou, X. and Liu, J. (2019) CoMM: a collaborative mixed model to dissecting genetic contributions to complex traits by leveraging regulatory information. Bioinformatics, 35, 1644–1652
53. Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., *et al.* (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun., 9, 1825
54. Fuller, W. A. (2009) Measurement Error Models. Volume 305. New Jersey: John Wiley & Sons
55. Liu, C., Rubin, D. B., and Wu, Y-. N. (1998) Parameter expansion to accelerate EM: the PX-EM algorithm. Biometrika, 85, 755–770
56. Cheng, Q., Yang, Y., Shi, X., Yang, C., Peng, H. and Liu, J. (2019) MR-LDP: a two-sample Mendelian randomization for GWAS summary statistics accounting linkage disequilibrium and horizontal pleiotropy. bioRxiv, 684746

57. Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., Kelsoe, J. R., O'Donovan, M. C., Furberg, H., Schork, N. J., *et al.* (2013) All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet., 9, e1003449
58. Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) An expanded view of complex traits: from polygenic to omnigenic. Cell, 169, 1177–1186
59. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P. and Pasaniuc, B. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet., 10, e1004722
60. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. and Eskin, E. (2014) Identifying causal variants at loci with multiple signals of association. Genetics, 198, 497–508
61. Pickrell, J. K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet., 94, 559–573
62. Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C. and Plagnol, V. (2014) Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet., 10, e1004383
63. Wen, X., Pique-Regi, R. and Luca, F. (2017) Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. PLoS Genet., 13, e1006646
64. Giambartolomei, C., Zhenli Liu, J., Zhang, W., Hauberg, M., Shi, H., Boocock, J., Pickrell, J., Jaffe, A. E., Pasaniuc, B. and Roussos, P. (2018) A Bayesian framework for multiple trait colocalization from summary association statistics. Bioinformatics, 34, 2538–2545
65. Efron, B. (2008) Microarrays, empirical bayes and the two-groups model. Stat. Sci., 23, 1–22
66. Turcot, V., Lu, Y., Highland, H. M., Schurmann, C., Justice, A. E., Fine, R. S., Bradfield, J. P., Esko, T., Giri, A., Graff, M., *et al.* (2018) Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. Nat. Genet., 50, 26–41

