A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies

Xingjie Shi[1,2], Xiaoran Chai[3], Yi Yang[2], Qing Cheng[2], Yuling Jiao[4], Jian Huang[5], Can Yang[6], Jin Liu[2*]

[1]Department of Statistics, Nanjing University of Finance and Economics, Nanjing, China

[2]Centre for Quantitative Medicine, Health Services & Systems Research, Duke-NUS Medical School, Singapore

[3]School of Medicine, National University of Singapore, Singapore

[4]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China

[5]Department of Statistics and Actuarial Science, University of Iowa, USA

[6]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong, China

[*]To whom the correspondence should be addressed

Abstract

Transcriptome-wide association studies (TWAS) integrate expression quantitative trait loci (eQTLs) studies with genome-wide association studies (GWASs) to prioritize candidate target genes for complex traits. Several statistical methods have been recently proposed to improve the performance of TWAS in gene prioritization by integrating the expression regulatory information imputed from multiple tissues, and made significant achievements in improving the ability to detect gene-trait associations. The major limitation of these methods is that they cannot be used to elucidate the specific functional effects of candidate genes across different tissues. Here, we propose a tissue-specific collaborative mixed model (TisCoMM) for TWAS, leveraging the

[*]Correspondence should be addressed to Jin Liu (jin.liu@duke-nus.edu.sg)

co-regulation of genetic variations across different tissues explicitly via a unified probabilistic model. TisCoMM not only performs hypothesis testing to prioritize gene-trait associations, but also detects the tissue-specific role of candidate target genes in complex traits. To make use of widely available GWAS summary statistics, we extend TisCoMM to use summary-level data, namely, TisCoMM-S^2. Using extensive simulation studies, we show that type I error is controlled at the nominal level, the statistical power of identifying associated genes is greatly improved, and false positive rate (FPR) for non-causal tissues is well controlled at decent levels. We further illustrate the benefits of our methods in applications to summary-level GWAS data of 33 complex traits. Notably, apart from better identifying potential trait-associated genes, we can elucidate the tissue-specific role of candidate target genes. The follow-up pathway analysis from tissue-specific genes for asthma shows that the immune system plays an essential function for asthma development in both thyroid and lung tissues.

Introduction

Over the last decade, GWASs have achieved remarkable successes in identifying genetic susceptible variants for a variety of complex traits [1]. However, the biological mechanisms to understand these discoveries remain largely elusive as majority of these discoveries are located in non-coding regions [2]. Recent expression quantitative trait loci (eQTLs) studies indicate that the expression regulatory information may play a pivotal role bridging both genetic variants and traits [3, 4, 5]. Cellular traits in comprehensive eQTL studies can serve as reference data, providing investigators with an opportunity to examine the regulatory role of genetic variants on gene expression. For example, the Genotype-Tissue Expression (GTEx) Project [6] has provided DNA sequencing data from 948 individuals and collected gene-expression measurements of 54 tissues from these individuals in the recent V8 release.

Transcriptome-wide association studies (TWAS) has been widely used to integrate the expression regulatory information from these eQTL studies with GWAS to prioritize genome-wide trait-associated genes [7, 8, 9]. A variety of TWAS methods have been proposed using different prediction models for expression imputation, including the parametric imputation models, e.g., PrediXcan [7], TWAS [8], CoMM [10] and CoMM-S^2 [11], and the nonparametric imputation model, e.g., Tigar [12]. These methods have been used for analyzing many complex traits with expression profiles from different tissues, successfully enhancing the discovery of genetic risk loci for complex traits [13, 9]. To further improve the power of identifying potential target genes, two recent studies were proposed by leveraging the substantial shared eQTLs across different tissues, i.e., MultiXcan [14] and UTMOST [15]. They use a step-wise procedure by first conducting imputation for gene expressions across multiple tissues and then performing subsequent association analysis using a multivariate regression that pools information across different tissues. Compared to single-tissue methods, these multi-tissue strategies enhance the imputation accuracy for gene expression and thus improve the power of identifying potential target genes.

Despite their successes, the existing multi-tissue methods have several limitations. First, MultiXcan and UTMOST cannot be used to identify the tissue-specific gene-trait associations. Many studies have shown that genes associated with complex traits are always regulated in a tissue-specific manner [16, 17, 18, 9]. For example, a recent study across 44 tissues confirmed this phenomenon in 18 complex traits [19], implying the persuasive role of tissue-specific

regulatory effects in a wide range of complex traits. Using a single-tissue test, one can easily reach a false conclusion regarding which tissue that a gene affects traits through. Second, both MultiXcan and UTMOST rely on a step-wise inference framework, ignoring the uncertainty in the process of expression imputation and thus losing power, especially when cellular-heritability is small [10]. Recently, CoMM [10] and its variant for summary-level data, CoMM-S^2 [11], have been proposed to account for uncertainty in the process of expression imputation. Third, MultiXcan and UTMOST do not make efficient use of the shared patterns of eQTLs across tissues, where MultiXcan uses principal component analysis (PCA) regularization on the predicted expression data, and UTMOST uses penalized regularization on coefficients for eQTL effects. A study of GTEx revealed these shared patterns [20], and later many efforts have been made to take advantage of them in the analysis for GTEx data. For example, Urbut et al. proposed statistical methods for estimating and testing eQTL effects explicitly incorporating this extensively tissue-shared patterns [21], shedding light on how to account for the tissue-shared eQTLs in statistical modeling successfully.

To overcome these limitations, we propose a tissue-specific collaborative mixed model (TisCoMM) for TWAS, providing a principled way to perform gene-trait joint and tissue-specific association tests across different tissues. Our method allows us not only to perform hypothesis testing to prioritize gene-trait association but also to uncover the tissue-specific role of candidate genes. By conditioning on the trait-relevant tissues, one could largely remove the spurious associations due to highly correlated gene expressions among multiple tissues. As a unified model, TisCoMM jointly conducts the "imputation" and the association analysis, pooling expression regulatory information across multiple tissues explicitly. Furthermore, we extend TisCoMM to use summary statistics from a GWAS, namely, TisCoMM-S^2. In simulations, we show that both TisCoMM and TisCoMM-S^2 provide correctly controlled type I error and are more powerful than existing multi-tissue methods. More importantly, our methods can be used to test for the tissue-specific role of candidate genes. We illustrate the benefits of our methods using summary-level GWAS data in 33 complex traits. Results show that our findings have biologically meaningful implications. The follow-up pathway analysis from tissue-specific genes for asthma shows that the regulated immune system in both thyroid and lung tissues could have significant impact on asthma development.

Results

Method overview

Our method, TisCoMM, jointly integrates expression regulatory information across multiple tissues by considering two models. The first one models the relationship between genetic factors and gene expressions across multiple tissues in the eQTL data set,

$$\mathbf{Y}_g = \mathbf{X}_{1g}\mathbf{B}_g + \mathbf{E}_g, \tag{1}$$

where $\mathbf{Y}_g \in \mathbb{R}^{n_1 \times T}$ is expression matrix of n_1 samples across T tissues for gene g, $\mathbf{X}_{1g} \in \mathbb{R}^{n_1 \times M_g}$ is the standardized genotype matrix corresponding to M_g nearby single nucleotide polymorphisms (SNPs) of gene g in the eQTL data, $\mathbf{B}_g$ is an $M_g \times T$ matrix of the corresponding effect sizes across T tissues and $\mathbf{E}_g$ is an $n_1 \times T$ matrix for random errors from a multivariate normal distribution $\mathcal{N}(0, \mathbf{V}_e)$. Here, $\mathbf{V}_e$ captures the correlations among tissues from the same individual. Then we assume that phenotypic value $\mathbf{z}$ and standardized genotype $\mathbf{X}_{2g}$ in GWAS are related by

$$\mathbf{z} = \mathbf{X}_{2g}\mathbf{B}_g\boldsymbol{\alpha}_g + \mathbf{e}_z, \tag{2}$$

where $\mathbf{z}$ is an $n_2 \times 1$ vector of phenotypic values, $\mathbf{X}_{2g} \in \mathbb{R}^{n_2 \times M_g}$ is the standardized genotype matrix corresponding to M_g nearby variants of gene g in the GWAS data, $\boldsymbol{\alpha}_g$ is a $T \times 1$ unknown parameter vector of interest that represents the effect sizes of "imputed" gene expression across T tissues for gene g, and $\mathbf{e}_z \sim \mathcal{N}(0, \sigma^2)$ is an $n_2 \times 1$ vector of independent errors associated with the trait. Our TisCoMM can be depicted as Figure 1, within which Figure 1A illustrates the TisCoMM method combing both the expression prediction model (1) and the corresponding association model (2) together with data input and output.

To pool expression regulatory information across relevant tissues, we assume the factorizable assumption [22, 23] for $\mathbf{B}_g = [\beta_{jt}]$, $j = 1, \ldots, M_g, t = 1, \ldots, T$. This assumption has been empirically validated for GTEx data in an imputation study [24] and Park et al. further used this assumption in a multi-tissue TWAS [25]. Here, we assume that the effect size of cis-SNP j in tissue t can be factorized by variant-dependent and tissue-dependent components: $\beta_{jt} = b_j w_{jt}$, where b_j (variant) is the eQTL effect of cis-SNP j shared in all the T tissues, and w_{jt} is the tissue-specific effect size. Thus, we have $\mathbf{B}_g = \text{diag}\{\mathbf{b}\}\mathbf{W}$. This factorization allows us to model the co-regulation of cis-SNPs shared across different tissues explicitly (Figure 1A, right). To make TisCoMM identifiable, we further assume that b_j independently follows

a normal distribution $\mathcal{N}(0, \sigma_b^2)$ due to polygenicity and by following the adaptive weighting strategy used in [24], the adaptive weight w_{jt} is estimated using the marginal regression of gene expression in tissue t on the j-th genetic variant.

The parameter of our interest in TisCoMM is the vector of effect size $\boldsymbol{\alpha}_g$. To prioritize candidate target genes, we conduct hypothesis testing for a joint null, $H_0 : \boldsymbol{\alpha}_g = 0$ (Figure 1B). To further explore the tissue-specific roles of candidate genes, we conduct hypothesis testing for each tissue, $H_0 : \alpha_{gt} = 0$, $t = 1, \ldots, T$ (Figure 1C). We refer to the two inference tasks as the TisCoMM joint test and TisCoMM tissue-specific test, respectively. We develop an expectation-maximization (EM) algorithm for parameter estimation by maximizing the complete-data likelihood. A parameter expansion technique is further adopted to accelerate computational efficiency (see details in Supplementary Text). In contrast to the existing two-step TWAS methods, we perform TisCoMM analysis in a unified model by treating $\mathbf{b}$ as a hidden random variable. Generally, the computational cost for the TisCoMM tissue-specific test is $\mathcal{O}(T)$ of that for the TisCoMM joint test. To enable computational efficiency, we only conduct the TisCoMM tissue-specific test for candidate genes detected in the joint test, rather than for all genes.

In a single-tissue analysis, it is difficult to explore the tissue-specific role of a candidate gene. The disease-associated genes will be identified in all the causal tissues as well as the tissues (possibly non-causal) highly correlated with the causal one, because there exist sharing patterns for expressions in multiple tissues. By conditioning on the trait-relevant tissues, our tissue-specific test could largely remove the spurious discoveries due to correlated expression across tissues.

[Figure 1 about here.]

Inferring TisCoMM results from GWAS summary statistics

To make our method widely applicable, we extend TisCoMM to use summary-level GWAS data, denoted as TisCoMM-S^2. The model details are given in Supplementary Text.

We observe high concordance between TisCoMM and TisCoMM-S^2 results. Figure 2 shows the comparison of TisCoMM and TisCoMM-S^2 test statistics for ten traits from the Northern Finland Birth Cohorts program 1966 (NFBC1966) data set [26] (see Methods section). The reference panel was 400 subsamples from the NFBC1966 data set. The high correlation between

TisCoMM and TisCoMM-S^2 suggests the goodness of detections for trait-associated genes using summary-level GWAS data.

[Figure 2 about here.]

To test the robustness of TisCoMM-S^2, we applied European subsamples from 1000 Genomes as the reference panel. Note that the NFBC1966 data set is Finns study, and it is well known that Finns have significant genetic differences with other Europeans [27]. Hence, the estimated LD did not well match that of the GWAS study. Supplementary Figure S1 shows the performance of TisCoMM-S^2 using European subsamples as a reference panel data set. Despite the high concordance between TisCoMM and TisCoMM-S^2 in the null region ($\Lambda > 34.67 = p$-values $> 5 \times 10^{-6}$), the test statistics of TisCoMM-S^2 in the non-null region are much more significant than TisCoMM.

Simulation

Methods for comparison To detect gene-trait association, we compared the performance of three methods in the main text: (1) our TisCoMM and TisCoMM-S^2 implemented in the R package *TisCoMM*; (2) MultiXcan and S-MultiXcan implemented in the MetaXcan package available at http://gene2pheno.org/; (3) UTMOST available at https://github.com/Joker-Jerome/UTMOST/. To detect the tissue-specific effect, we compared the performance of TisCoMM tissue-specific test with three single-tissue methods that include (1) CoMM available at https://github.com/gordonliu810822/CoMM; (2) PrediXcan available at http://gene2pheno.org/; (3) TWAS relies on the BSLMM [28] implemented in the GEMMA [28] software. All methods were used with default settings. We conducted comprehensive simulations to gauge the performance of each method better by performing gene-trait joint and tissue-specific tests across different tissues.

Simulation settings In detail, we considered the following simulation settings. We set $\{n_1, n_r, n_2\} = \{400; 400; 5,000\}$ as the sample size for eQTL data, GWAS data and reference panel data. We first generated the genotype data for $M_g = 400$ cis-SNPs from a multivariate normal distribution assuming an autoregressive correlation with parameter ρ. We then discretized each SNP to a trinary variable $\{0, 1, 2\}$ by assuming Hardy-Weinberg equilibrium and a minor allele frequency randomly selected from a uniform $[0.05, 0.5]$ distribution. The genotype correlation was varied at $\rho = \{0.2, 0.5, 0.8\}$. All three genotype matrices, $\mathbf{X}_{1g}, \mathbf{X}_{rg}$,

and $\mathbf{X}_{2g}$, for eQTL data, GWAS data and reference panel data, respectively, are generated in this manner.

To generate multi-tissue gene expressions, we considered different cellular-level heritability levels (h_c^2) and sparsity levels (s). These are key parameters to describe the genetic architecture of gene expression [29]. The cellular-level heritability represents the proportion of variance of the gene expression that can be explained by genotype, while sparsity represents the proportion of genetic variants that are associated with the gene expression. First, SNP effect size $\mathbf{B}_g = \text{diag}\{\mathbf{b}\}\mathbf{W}$ is generated. Specifically, we simulated SNP effect size $\mathbf{b}$ from a standard normal distribution, and randomly selected 10%, 50% or 100% of the SNPs to have non-zero tissue-specific effect $\mathbf{W}$ for gene expressions in all T tissues, while simulated their effects from a standard normal distribution. We then simulated errors $\mathbf{E}_g$ from a normal distribution, where their variances were chosen according to h_c^2, and the covariance structure was autoregressive with $\rho_e = 0.5$. Here we set $h_c^2 = 0.025, 0.05, 0.1, 0.2, 0.4$. Afterward, we simulated a multi-tissue eQTL data set assuming $\mathbf{Y}_g = \mathbf{X}_{1g}\mathbf{B}_g + \mathbf{E}_g$.

To simulate a quantitative trait, we generated nonzero entries of $\boldsymbol{\alpha}_g$ from a uniform distribution and $\mathbf{e}_z$ from a normal distribution. The variance σ^2 was chosen according to the tissue-level heritability $h_t^2 = \frac{\text{Var}(\mathbf{X}_{2g}\mathbf{B}_g\boldsymbol{\alpha}_g)}{\text{Var}(\mathbf{z})}$. Here we set $h_t^2 = 0$ for null simulations and type I error control examination and $h_t^2 = 0.01$ for non-null simulations and power comparisons.

Simulation I: Testing gene-trait associations We focus on the detection of trait-associated genes in the first set of simulations. Here, we compared TisCoMM and TisCoMM-S^2 with three different multi-tissue methods that include MultiXcan, S-MultiXcan, and UTMOST. We set $T = 10$, and all tissues are causal. For each scenario, we run 5,000 replicates. We first examined type I error control of different methods under the null. Results are shown in Supplementary Figures S2 – S6. By comparing the distribution of p-values with the expected uniform distribution, we observe that all methods provide well-controlled type I errors.

Next, we examined the power of different methods under the alternative hypothesis, as shown in Figure 3. We observe that the performance of all five methods improves with the increment of cellular heritability. In general, the summary-level methods (TisCoMM-S^2 and S-MultiXcan) perform similarly to their counterparts in individual-level data. Moreover, TisCoMM and TisCoMM-S^2 have better performance than other alternative methods when cellular heritability is relatively small ($h_c^2 = 0.025, 0.05, 0.1$), and comparable performance when cellular heritability

is large. Finally, we observe that although our model favors dense eQTLs, it was robust to the sparsity level s. Specifically, the power of TisCoMM and TisCoMM-S^2 in the setting where 10% of cis-SNPs have non-zero effects on gene expression are similar to the setting where all cis-SNPs have non-zero effects.

[Figure 3 about here.]

Simulation II: Testing tissue-specific effects We focus on the detection of tissue-specific effects in the second set of simulations. Here, we compared the TisCoMM tissue-specific test with the single-tissue methods including CoMM [10], PrediXcan [7], and TWAS[8] under the alternative hypothesis with fixed tissue heritability $h_t^2 = 0.01$ and fixed sparsity $s = 0.1$. We considered three tissues $T = 3$ and varied the number of causal tissues to simulate different levels of tissue specificity of a trait. Specifically, we considered settings with one ($\alpha_{g2} = \alpha_{g3} = 0$) and two causal tissues ($\alpha_{g3} = 0$), respectively. To allow correlated gene expression in the GWAS, the nonzero of tissue-specific effect $\mathbf{W}$ was generated with rows drawn from a multivariate normal distribution, with AR correlation parameter $\rho_W = 0.2, 0.5, 0.8$. A large value of ρ_W implies a higher correlation among columns of $\mathbf{X}_{2g}\mathbf{B}_g$. Other sittings are similar to Simulation I.

We repeated the whole process 1,000 times. We calculated statistical power and false positive rate (FPR) as the proportion of p-values reaching the significance level in causal tissues and non-causal tissues, respectively. Specifically, we set the significance level at 0.05/3 for all considered methods. Figure 4 shows simulation results for the case that one tissue is causal. We observe that in all settings, the TisCoMM tissue-specific test has comparable or slightly inferior power, as shown in Figure 4A, compared to the single-tissue methods, but much smaller FPR (Figure 4B).As expected, the statistical power of all methods increases with cellular heritability (h_c^2). However, the FPR of single-tissue methods substantially inflates while that of TisCoMM tissue-specific test remains at the same level. Furthermore, the FPR of TisCoMM tissue-specific test does not vary with correlations among expressions across multiple tissues (ρ_W) while that of single-tissue methods increase with ρ_W. The similar pattern could be observed for the case that two tissues are causal (Supplementary Figure S7). These results demonstrate the usefulness of TisCoMM tissue specific test in exploring the tissue-specific role of genes.

[Figure 4 about here.]

Real Data Applications

We performed multi-tissue TWAS analysis for summary-level GWAS data in 33 complex traits (see Supplementary Table S1 for details), including 15 traits from Gamazon et al. [19] and 18 traits from the UK Biobank. Hereafter we refer to as NG traits and UKB traits, respectively. These traits can be roughly divided into four categories, including metabolites (e.g., HDL-C, LDL-C and fasting glucose), autoimmune diseases (e.g., asthma, Crohn's disease and macular degeneration), psychiatric/neurodegenerative disorders (e.g., Alzheimer's disease, major depression disorder, and psychiatric disorder), and cardiovascular disorders (e.g., coronary artery disease and peripheral vascular disease). The Genotype-Tissue Expression (GTEx) Project [6] reported eQTL in 48 tissues, where the number of genes in each tissue ranges from 16,333 to 27,378. In the analysis, we extracted cis-SNP that are within either 500 kb upstream of the transcription start site or 500 kb downstream of the transcription end site.

In a single-tissue analysis, there are two different strategies to select a tissue for TWAS: one uses expressions from the most biologically related tissue while the other selects a tissue with the largest number of available individuals [9]. To select multiple tissues for TisCoMM-S^2, there exists a trade-off between biological relevance and its corresponding sample size for each tissue. In [19], it provides the most biologically related tissues and thus we used trait-relevant tissues for the NG traits from Supplementary Table 2 in [19]. In detail, for each trait, a set of tissues with significant enrichment p-values (after Bonferroni correction) was identified, and a subset with more than 100 overlapped samples [30] was chosen for further analysis in TisCoMM-S^2. On the other hand, although methods like LD score regression [17] can be used for the UKB traits, it is difficult to balance the tissue relevance and sample size for each tissue. To make efficient use of the GTEx data set, we used six tissues with the largest number of overlapped samples for the UKB traits.

The analysis for each trait based on its GWAS summary statistics together with the eQTL data from multiple tissues can be done around 100 min on a Linux platform with 2.6 GHz Intel Xeon CPU E5- 2690 with 30720 KB cache and 96 GB RAM (0nly 10$\sim$12 GB RAM used) on 24 cores.

TisCoMM-S^2 joint test provides statistically powerful results of disease relevant genes

To prioritize trait-associated genes, we compared TisCoMM-S^2 with other two multi-tissue TWAS methods, i.e., S-MultiXcan and UTMOST. Both alternative methods take advantage of prediction models to impute gene expressions. The prediction models used here were Elastic Net models trained on 48 GTEx tissues. See Table 1 and 2 for the summary of detections across different approaches for the 15 NG and 18 UKB traits, respectively. Generally, TisCoMM-S^2 identifies more genome-wide associations than S-MultiXcan and UTMOST in most traits. In detail, TisCoMM-S^2/S-MultiXcan/UTMOST identified 3,058/2,008/1,769, and 443/338/277 genome-wide significant genes in all the NG traits and UKB traits, respectively. Their qq-plots of p-values are shown in Supplementary Figures S8 – S11 and plots for their genomic inflation factors are shown in Supplementary Figure S12. As case study examples, we carefully examined the results for late-onset Alzheimer's disease (LOAD) and asthma.

LOAD results After Bonferroni correction, TisCoMM-S^2/S-MultiXcan/UTMOST identified 92/71/70 genome-wide significant genes, respectively, with 45 overlapping genes (17 of them are known LOAD GWAS genes). Here we define known LOAD GWAS gene as the ones reported in GWAS catalog. The qq-plots for associations in these three approaches are shown in Figure 5A. Among the 92 candidate target genes identified by TisCoMM-S^2, 24 of them are previously known LOAD GWAS genes, which are annotated in the Manhattan plot in Figure 5A. These include genes on the chromosome (CHR) 2 (*BIN1*), CHR 6 (*CD2AP*), CHR 7 (*EPHA1*), CHR 8 (*CLU*), CHR 11 (*PICALM, CCDC89, MS4A2, MS4A6A*), CHR 16 (*IL34*) , and CHR 19 (*STK11* and *APOE* region). Moreover, TisCoMM-S^2 also identified 35 genes that were not significant in neither S-MultiXcan nor UTMOST, and four of them are known LOAD GWAS genes, including *IL34* (p-value $=1 \times 10^{-6}$), *PTK2b* (p-value $=1.4 \times 10^{-9}$), *EPHX* (p-value $=4.7 \times 10^{-8}$) and *STK11* (p-value $= 7.2 \times 10^{-7}$).

Among all novel genes for LOAD identified by TisCoMM-S^2, some of them were identified to be LOAD-related genes based on other computational models (e.g., *MAP3K2*) while some of them have not been directly linked to LOAD yet, but have been proven to be important regulators in different regions of the neuron system (e.g., *STMN4*, *EED* and *APC2*). *MAP3K2* is 200kb downstream of *B1N1*, a reported LOAD risk gene [31] that was also genome-wide significant in our joint test (p-values for both *B1N1* and *MAP3K2* $< 10^{-10}$). *MAP3K2* belongs

to the serine/threonine protein kinase family and has been previously identified as a member of the Alzheimer's disease susceptibility network [32]. *STMN4* (p-value $< 10^{-10}$) encodes the known protein that exhibits microtubule-destabilizing activity. The expression levels of this gene in mouse neurons have been shown to change significantly after different exposure of cortical nerve cells to the Aβ peptide [33]. The expression of *STMN4* in zebrafish has also been shown to have an important role in regulating neurogenesis in the neural keel stage [34]. *EED* (p-value $=5.7 \times 10^{-7}$) encodes a Polycomb protein, which plays a starring role as an important modulator of hippocampal development [35]. *APC2* (p-value $= 1.3 \times 10^{-6}$) is preferentially expressed in postmitotic neurons and involved in brain development through its regulation of neuronal migration and axon guidance [36]. We annotate these four genes in red in Figure 5A.

Validation of these potential target genes requires further functional studies. The list of significant gene-trait associations of TisCoMM-S^2, S-MultiXcan, and UTMOST can be found in Supplementary Table S2. To replicate our findings in another independent data set, we used the summary statistics from the GWAS by proxy (GWAX [37], the sample size is 114,564). Our replication rate was high (Supplementary Table S3), where 31 out of 92 genes were successfully replicated under the Bonferroni-corrected significance threshold and the numbers of replicated genes raised to 44 under a relaxed p-value cutoff of 0.05.

Asthma results After Bonferroni correction, TisCoMM-S^2/S-MultiXcan/UTMOST identified 200/157/140 genome-wide significant genes, respectively, with 98 overlapping genes in all three methods (and 21 of them are known asthma GWAS genes). The qq-plots for associations in these three approaches are shown in Figure 5B. Among all 200 candidate target genes identified by TisCoMM-S^2, 31 of them are known asthma GWAS genes, which is annotated in the Manhattan plot in Figure 5B, including genes on CHR 2 (*IL1RL1/IL18R1*), CHR 5 (*TSLP/WDR36*, *RAD50*), CHR 6 (*HLA-DR/DQ* regions, *MAP3K7*), CHR 9 (*IL33*), CHR 11 (*C11orf30*, *LRRC32*), CHR 15 (*SMAD3*), and CHR 17 (genes from the 17q21 asthma locus). Also, TisCoMM-S^2 identified 56 genes that were not significant in neither S-MultiXcan nor UTMOST, and two of them are known asthma GWAS genes, which are *PSORS1C1* (p-value $=2.2 \times 10^{-7}$), and *MAP3K7* (p-value $=3 \times 10^{-7}$).

Among all novel loci for asthma identified by TisCoMM-S^2, *PDCD1LG2* was shown to have essential roles in modulating and polarizing T-cell functions in airway hyperreactivity [38]. Validating causal role of this gene in asthma requires further investigation. The list of

significant gene-trait associations of TisCoMM-S^2, S-MultiXcan, and UTMOST can be found in Supplementary Table S4. We annotate these two genes in red in Figure 5B.

To replicate our findings in another independent data set, we used the summary statistics from TAGC European-ancestry GWAS [39] (the sample size is 127,669). Our replication rate was high (Supplementary Table S5), where 179 out of 200 genes were successfully replicated under the Bonferroni-corrected significance threshold and the numbers of replicated genes raised to 189 under a relaxed p-value cutoff of 0.05.

TisCoMM-S^2 tissue-specific test infers gene effects in causal tissues

To demonstrate the utility of the TisCoMM-S^2 tissue-specific test, we applied the tissue-specific test to all identified 92 candidate genes of LOAD and 200 candidate genes of asthma by using the TisCoMM-S^2 joint test, and compared analysis results with those from CoMM [10, 11]. Table 3 shows the distributions of identified tissues with which candidate genes are associated in LOAD and asthma, respectively (see details in Supplementary Tables S6 and S7). Among all identified candidate genes respectively for both LOAD and asthma, 76.1% and 81.5% were significant in less than two tissues using TisCoMM-S^2 while 70.7% and 60% were significant in all six tissues using CoMM-S^2. The most plausible explanation is that compared to the multivariate perspective of our TisCoMM-S^2 tissue-specific test, single-tissue approaches, e.g., CoMM-S^2, tend to have larger tissue bias and more inflation in significant findings [9]. Suppose a gene is causal in tissue A but not in tissue B, and its expressions in tissues A and B are correlated. In a single-tissue test, the association can be spuriously significant for tissue B because of the similar gene expression pattern observed in both tissues. By performing a tissue-specific test for this gene in tissue B conditioned on tissue A, the significant spurious association will be largely excluded.

[Table 1 about here.]

To demonstrate the tissue-specific role of candidate genes inferred by TisCoM-S^2 tissue-specific test for LOAD and asthma, respectively, we plot the volcano plots in Supplementary Figure S13, where the x-axis is the effect size showing in log scale, the y-axis is $-log10$ of the p-value from tissue-specific test, and the size of points reflect the cellular-heritability in each tissue. Known GWAS genes are also annotated. Next, we explored the tissue-specific effects of

some well-replicated genes that are identified by the TisCoMM-S^2 joint test for LOAD and asthma, respectively.

LOAD results The well-replicated risk gene *APOE* [40] and its 50Kb downstream *CLPTM1* have been identified by the TisCoMM-S^2 joint test. Moreover, the TisCoMM-S^2 tissue-specific test identified *CLPTM1* to be significantly associated with LOAD in all four tissues (artery aorta, esophagus mucosa, nerve tibial, and skin sun-exposed lower leg with tissue-specific p-values $< 4.9 \times 10^{-7}$), but *APOE* to be only significantly associated with LOAD in artery aorta (tissue-specific p-value $=8.3 \times 10^{-9}$) and nerve tibial (tissue-specific p-value $=1.2 \times 10^{-8}$). On the other hand, CoMM-S^2 significantly identified both *APOE* and *CLPTM1* in all four tissues (p-values $\leq 10^{-10}$) but failed to identify the difference of tissue-specific role for these two genes. We further investigate the molecular functions of LOAD associated genes in each tissue. In each of tested tissues in LOAD, there are about 40 tissue-specific genes. It is difficult to carry out a proper pathway analysis with such limited gene sets. So we classified the genes into seven functional groups based on which molecular functions they belong to. As shown in Figure 6A and 6B, majority ($> 62\%$) of LOAD-associated genes belonged to binding and catalytic activity, and a small portion of significant LOAD genes were transcription factors suggesting that many regulation processes are going on at both protein and mRNA levels in different tissues.

According to our tissue selection strategy, above tissue-specific test for LOAD was conducted on four non-brain tissues (enriched tissues). To further investigate the gene expression changes in the well-studied disease tissues, three more brain regions (hippocampus, frontal cortex, and cerebellar hemisphere) were selected for another tissue-specific analysis for LOAD. Because it is known that hippocampus is one of the first brain regions to be affected by Alzheimer's disease and related to the memory lost [41], markers such as Aβ in frontal cortex can be used to predict future Alzheimer's disease [42], and cerebellum is affected in the final stage of the disease and related to cognitive decline [43]. The joint test conducted on brain regions revealed 105 LOAD associated genes, of which 73 were identified in the enriched tissues (Figure S14A), and the other 32 genes were uniquely identified in brain regions (Figure S14B). The most significant gene uniquely identified in brain regions is *KLC3* according to the joint test (p-value $< 10^{-10}$), which is within 50kb downstream of *APOE*. Moreover, it is significantly associated with LOAD in hippocampus region only, but not the other two brain regions according to the tissue-specific

test (Figure S14B). Thus, we propose *KLC3* as one of the potential novel targets for LOAD in hippocampus.

Asthma results We take identified genes ORMDL3 and GSDMB in the 17q21 asthma locus as an example, because these two genes have been mentioned as asthma susceptibility locus by many studies, a comprehensive review was written by Stein et al. [44]. The original finding of ORMDL3 was observed in one GWAS study, and have been further validated in a mouse model [45]. The TisCoMM-S^2 tissue-specific test identified both *ORMDL3* and *GSDMB* to be significantly associated with asthma only in lung tissue (see the volcano plot in Supplementary Figure S14B, tissue-specific p-values for these two genes are 1.7×10^{-3} and 7.1×10^{-7}, respectively). However, CoMM-S^2 identified both *ORMDL3* and *GSDMB* in all six tissues (p-values $\leq 10^{-10}$) but failed to identify the relevant tissues with which these two genes are causally related to asthma. We further conducted pathway analysis using DAVID [46] on six sets of asthma-associated genes in all six tissues (thyroid, lung, artery tibial, muscle skeletal, adipose subcutaneous, and skin sun-exposed lower leg), respectively. As listed in Figure 6B, all three significant pathways in thyroid tissue belonged to the immune system, and the only significant pathway in lung tissue was immune response. However, no significant pathways were detected in the other four tissues. Among asthma-associated genes in immune response (first row in Figure 6C and 6D), the majority of them were shared between thyroid and lung, and located in the MHC region on CHR 6 including several *HLA* genes and *LST1*. Our pathway analysis suggests that nearly the same set of immune genes in thyroid and lung are responsible for asthma development.

[Figure 5 about here.]

[Figure 6 about here.]

Discussion

Despite the substantial successes of TWAS and its variants, the existing multi-tissue methods have several limitations, e.g., incapability to identify the tissue-specific effect of a gene, ignorance of imputation uncertainty, and failure to efficiently use tissue-shared patterns in eQTLs. To overcome these limitations and provide additional perspectives over tissue-specific roles of identified genes, we have proposed a powerful multi-tissue TWAS model, together with a

computationally efficient inference method and software implementation in TisCoMM. Specifically, we have developed a joint test for prioritizing gene-trait associations and a tissue-specific test for identifying the tissue-specific role of candidate genes. Conditioned on the inclusion of trait-relevant tissues, the tissue-specific test in TisCoMM can mostly remove the spurious associations in a single-tissue test due to high correlations among gene expression across tissues. We have also developed a summary-statistic-based model, TisCoMM-S^2, extending the applicability of TisCoMM to publicly available GWAS summary data. Using both simulations and real data, we examined the relationship between TisCoMM and TisCoMM-S^2. Our results, as shown in Figure 2, show that the test statistics from TisCoMM and TisCoMM-S^2 are highly correlated ($R^2 > 0.95$). We further analyzed summary-level GWAS data from 33 traits with replication data for Alzheimer's disease and asthma. Overall, the findings from TisCoMM-S^2 are around 30% more than those from S-MultiXcan or UTMOST while qq-plots from these studies show that there are no apparent inflations. To replicate our findings for Alzheimer's disease and asthma, we applied TisCoMM-S^2 to independent data sets for each disease. Results show that replication rates for Alzheimer's disease and asthma are high.

We further inferred the tissue-specific effects of identified genes using the TisCoMM-S^2 tissue-specific test. By classifying these genes into seven functional groups, we observed that majority (62%) of LOAD-associated genes were related to binding and catalytic activity while a small portion was from transcription factors suggesting active regulation processes at both protein and mRNA level in different tissues. We also observed about 40 LOAD-associated genes in each non-brain tissues. The significance of these genes could be due to the exclusion of LOAD-relevant tissues, e.g., brain tissues. To fill this gap, we further conducted one more analysis on three brain regions, and identified 32 brain specific genes. For asthma, genes *ORMDL3* and *GSDMB* were identified to be significantly associated with asthma only in lung tissue using TisCoMM-S^2 tissue-specific test. However, single-tissue analysis (CoMM-S^2) identified both genes significant in all six tested tissues. Further pathway analysis shows that all three significant pathways for thyroid tissue belong to the immune system and the only significant pathway for lung tissue was immune response. The majority of shared genes between thyroid and lung tissues are located in the MHC region on CHR 6, including several *HLA* genes and *LST1*. The proteins encoded by *HLA* genes are known as antigens. In combination with antigen-presenting cells (e.g., macrophages and dendritic cells), they play an essential role in the

activation of immune cells as well as airway inflammation in response to asthma-related allergens [47, 48]. Based on our tissue-specific test, *TNF* that is a well-studied asthma gene [49, 50] was explicitly identified to be associated with asthma in lung tissue. The positive correlation between *TNF* expression and asthma in lung confirmed our previous understanding of *TNF* activation in asthma, promoting airway inflammation and airway hyperresponsiveness. On the other hand, *LTA* was specifically regulated in thyroid tissue. It is a cytokine produced by lymphocytes, and also known as a regulator of lipid metabolism [51]. Another immune gene regulated individually in thyroid tissue is *NCR3*, which mediates the crosstalk between natural killer cells and dendritic cells [52]. However, it remains unclear how the alteration of *LTA* and *NCR3* in thyroid could lead to asthma development.

Despite the utility of TisCoMM to perform gene-trait association analysis in a tissue-specific manner, it is primarily designed to test genes with direct effects from cis-eQTL. Recently, an omnigenic model was proposed to better understand the underlying mechanism of so-called polygenicity in complex traits [53]. Liu et al. [54] further provided a theoretical model to understand complex trait architecture by partitioning genetic contributions into direct effects from core genes and indirect effects from peripheral genes acting in trans. Most works from TWAS identify core genes with direct effects. How to effectively interrogate peripheral genes with indirect effects essentially remains an open question. As high-throughput data are continuously generating for a much larger sample size with more precision, TisCoMM sheds light on how to integrate useful data for the desired analysis effectively.

Methods

Model settings

Conventionally, both single-tissue and multi-tissue TWAS methods proceed by conducting a prediction model in Equation (1) followed by a subsequent association analysis in Equation (2), where a steady-state gene expression is imputed from $\mathbf{X}_{2g}\widehat{\mathbf{B}}_g$ and $\widehat{\mathbf{B}}_g$ is estimated in the first prediction model, e.g., PrediXcan, MultiXcan, S-MulitXcan, and UTMOST. However, this imputation strategy ignores the uncertainty in the process of expression imputation. Here, we describe the individual-level data version of TisCoMM by jointly analyzing models (1) and (2), and extensions to summary statistics will be discussed in the Supplementary Text. Assume $\mathcal{D}_1 = \{\mathbf{Y}_g, \mathbf{X}_{1g}\}$ denote the reference transcriptome data set of gene g for n_1 samples over

T tissues, where $\mathbf{Y}_g$ is the $n_1 \times T$ expression matrix for this gene over T tissues, and $\mathbf{X}_{1g}$ is the corresponding $n_1 \times M_g$ standardized genotype matrix for M_g cis-SNPs within this gene. Denote the GWAS data $\mathcal{D}_2 = \{\mathbf{z}, \mathbf{X}_{2g}\}$, where $\mathbf{z}$ is an $n_2 \times 1$ vector of phenotypic values, $\mathbf{X}_{2g}$ is the corresponding $n_2 \times M_g$ standardized genotype matrix for M_g cis-SNPs. Since we conduct hypothesis testing sequentially or parralelly for each gene, we will omit the subscript g in all the expression that has dependence on gene g to simplify notations. Our model becomes

$$\mathbf{Y} = \mathbf{X}_1\mathbf{B} + \mathbf{E}, \quad \mathbf{z} = \mathbf{X}_2\mathbf{B}\boldsymbol{\alpha} + \mathbf{e}_z, \tag{3}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^T$, $\mathbf{E} \sim \mathcal{MN}(0, \mathbf{I}_m, \mathbf{V}_e)$, and $\mathbf{e}_z \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$. Note that we assume $\mathcal{D}_1$ and $\mathcal{D}_2$ are centered and thus intercepts can be omitted.

To estimate the tissue-specific eQTL effects, we need to first estimate an $M \times T$ coefficient matrix $\mathbf{B}$. To reduce the number of parameters, we follow an adaptive weighting scheme [22, 23, 24]: we regress the gene expression in tissue type t on the jth eQTL and let the marginal eQTL effect be the adaptive weight, w_{jt}. Specifically, we assume the joint eQTL effect size β_{jt} can be decomposed into variant-dependent components b_j and tissue-specific components w_{jt}: $\beta_{jt} = b_j w_{jt}$. That is, $\mathbf{B} = \text{diag}\{\mathbf{b}\}\mathbf{W}$. Similar strategies have been applied to model tissue-shared patterns [24, 21]. Let $\mathbf{y}_i$, $\mathbf{x}_{1i}$ and $\mathbf{w}_j$ denote the ith row of $\mathbf{Y}$, $\mathbf{X}_1$ and $\mathbf{W}$, respectively. Our model can be written as

$$\begin{aligned}
\mathbf{y}_i|\mathbf{b} &\sim \mathcal{N}\big(\sum_j x_{1ij} b_j \mathbf{w}_j, \mathbf{V}_e\big), \\
z_i|\mathbf{b} &\sim \mathcal{N}\Big(\boldsymbol{\alpha}^\top(\sum_j x_{2ij} b_j \mathbf{w}_j), \sigma^2\Big), \\
b_j &\sim \mathcal{N}(0, \sigma_b^2).
\end{aligned}$$

Denote $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \sigma_b^2, \sigma^2, \mathbf{V}_e)^T$ the vector for all model parameters. We need to estimate parameters and maker inference for $\boldsymbol{\alpha}$. Both the TisCoMM joint test and tissue-specific test are based on likelihood ratio tests. The joint test for gene-trait associations can be formally set up as $H_0 : \boldsymbol{\alpha} = 0$ verses $H_1 : \boldsymbol{\alpha} \neq 0$. The corresponding likelihood ratio test statistic is given by

$$\Lambda = 2\left[\log \Pr(\mathbf{y}, \mathbf{z}|\mathbf{X}_1, \mathbf{X}_2; \hat{\boldsymbol{\theta}}) - \log \Pr(\mathbf{y}, \mathbf{z}|\mathbf{X}_1, \mathbf{X}_2; \hat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}=0})\right],$$

where $\hat{\boldsymbol{\theta}}$ is the vector of parameter estimates under the full model, and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\alpha}=0}$ is the vector of estimates under the constrain $\boldsymbol{\alpha} = 0$. Similarly, the tissue-specific test for the tissue-specific effect can be formally set up as $H_0 : \alpha_t = 0$ verses $H_1 : \alpha_t \neq 0$. The corresponding likelihood

ratio test statistic is given by

$$\Lambda_t = 2\left[\log \Pr(\mathbf{y}, \mathbf{z}|\mathbf{X}_1, \mathbf{X}_2; \hat{\boldsymbol{\theta}}) - \log \Pr(\mathbf{y}, \mathbf{z}|\mathbf{X}_1, \mathbf{X}_2; \hat{\boldsymbol{\theta}}_{\alpha_t=0})\right],$$

where $\hat{\boldsymbol{\theta}}_{\alpha_t=0}$ is the vector of parameter estimates under $\alpha_t = 0$.

For statistical inference, we developed an expectation-maximization (EM) algorithm accelerated by expanding parameters [55]. Details of updating equations for each parameter and the corresponding algorithm can be found in Supplementary Text.

GWAS data

The NFBC1966 data set

The NFBC1966 data set consists of ten traits and 364,590 SNPs from 5402 individuals [26], including total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C) and triglycerides (TG), inflammatory marker C-reactive protein, markers of glucose homeostasis (glucose and insulin), body mass index (BMI) and blood pressure (BP) measurements (systolic and diastolic BP). Quality control procedures are conducted following similar steps to Shi et al. [56]. Specifically, individuals with missing-ness in any of the traits and with genotype missing call-rates $> 5\%$ were excluded. We excluded SNPs with minor allele frequency (MAF) $< 1\%$, missing call-rates $> 1\%$, or failed Hardy-Weinberg equilibrium. After quality control filtering, 172,412 SNPs from 5123 individuals were available for downstream analysis.

The tissues used in TisCoMM and TisCoMM-S^2 were the same, and the six tissues with the largest number of overlapped individuals were used. The summary statistics for TisCoMM-S^2 were calculated using PLINK [57].

Summary-level GWAS data

We obtained summary statistics from GWASs for 33 traits, including 15 traits from [19] and 18 traits from the UK Biobank. Details of these traits can be found in Supplementary Table S1. In the main text, we discussed LOAD and asthma. Analyses results for other traits can be found in Supplementary Text.

GTEx eQTL Data

Th GTEx data including genotype and RNA-seq data are obtained from dbGaP with accession number phs000424.v7.p2. Processed gene-expression data are available on the GTEx portal (https://gtexportal.org/home/). In the eQTL data, we removed SNPs with ambiguous alleles or MAF less 0.01.

We used two different strategies to select tissues used in our real data analysis. For the 15 NG traits, we obtained the top enriched tissues for each trait according to Supplementary Table 2 in [19], and a subset of tissues with sample sizes larger than 100 was kept. For the UKB traits, we used the six tissues with the largest number of overlapped individuals.

Reference panel

Due to the absence of genotype data using summary statistics, we use reference samples to estimate the LD structures R among SNPs in the study samples. Since diseases and traits considered in our real data application are for European population cohorts, we choose to use European subsamples from the 1000 Genome Project as a reference panel.

Let $\mathbf{X}_r$ denote the genotype matrix for cis-SNPs in the reference panel. To estimate the LD matrix R, we adopt a simple shrinkage method as follows. We first calculate the empirical correlation matrix $\hat{R}^{\text{emp}} = [r_{jk}] \in \mathbb{R}^{M\times M}$ with $r_{jk} = \frac{X_{rj}^\top X_{rk}}{\sqrt{(X_{rj}^\top X_{rj})(X_{rk}^\top X_{rk})}}$, where X_{rj} the jth column of $\mathbf{X}_r$. To make the estimated correlation matrix positive definite, we apply a simple shrinkage estimator [58]: $\hat{R} = \tau R^{\text{emp}} + (1-\tau)\mathbf{I}_M$, where $\tau \in [0,1]$ is the shrinkage intensity. In real data application, we fixed the shrinkage intensity at 0.95 both for simplicity and computational stability.

Web Resources

TisCoMM is available at https://github.com/XingjieShi/TisCoMM/.

PrediXcan, MultiXcan and *S-MultiXcan* are available at http://gene2pheno.org/.

UTMOST is available at https://github.com/Joker-Jerome/UTMOST/.

CoMM is available at https://github.com/gordonliu810822/CoMM.

Known trait-associated genes are available at the NHGRI-EBI GWAS Catalog https://www.ebi.ac.uk/gwas/.

Summary statistics from UK Biobank is available at http://geneatlas.roslin.ed.ac.uk/.

URLs for summary statistics from Gamazon et al. [19] are summarized in Supplementary Table

S1.

Acknowledgements

This work was supported by grant R-913-200-098-263 from the Duke-NUS Medical School, AcRF Tier 2 (MOE2016-T2-2-029, MOE2018-T2-1-046 and MOE2018-T2-2-006) from the Ministry of Education, Singapore, grant No. 71501089, No. 11501579 and No. 71472023 from the National Natural Science Foundation of China; and grant Nos. 22302815, No. 12316116 and No. 12301417 from the Hong Kong Research Grant Council. The computational work for this article was partially performed using resources from the National Supercomputing Centre, Singapore (https://www.nscc.sg).

References

1. Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

2. Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

3. William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184, 2009.

4. Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

5. Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197, 2015.

574 6. John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad,
575 Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue
576 expression (gtex) project. *Nature genetics*, 45(6):580, 2013.

577 7. Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-
578 Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox,
579 et al. A gene-based association method for mapping traits using reference transcriptome
580 data. *Nature genetics*, 47(9):1091, 2015.

581 8. Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH
582 Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative
583 approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245,
584 2016.

585 9. Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A
586 Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al.
587 Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*,
588 51(4):592, 2019.

589 10. Can Yang, Xiang Wan, Xinyi Lin, Mengjie Chen, Xiang Zhou, and Jin Liu. Comm: a
590 collaborative mixed model to dissecting genetic contributions to complex traits by leveraging
591 regulatory information. *Bioinformatics*, 35(10):1644–1652, 2018.

592 11. Yi Yang, Xingjie Shi, Yuling Jiao, Jian Huang, Min Chen, Xiang Zhou, Lei Sun, Xinyi Lin,
593 Can Yang, and Jin Liu. Comm-s2: a collaborative mixed model using summary statistics
594 in transcriptome-wide association studies. *bioRxiv*, page 652263, 2019.

595 12. Sini Nagpal, Xiaoran Meng, Michael P Epstein, Lam C Tsoi, Matthew Patrick, Greg Gibson,
596 Philip L De Jager, David A Bennett, Aliza P Wingo, Thomas S Wingo, et al. Tigar:
597 An improved bayesian tool for transcriptomic data imputation enhances gene mapping of
598 complex traits. *The American Journal of Human Genetics*, 2019.

599 13. Yang I Li, Garrett Wong, Jack Humphrey, and Towfique Raj. Prioritizing parkinsons disease
600 genes using population-scale transcriptomic data. *Nature communications*, 10(1):994, 2019.

14. Alvaro N Barbeira, Milton D Pividori, Jiamao Zheng, Heather E Wheeler, Dan L Nicolae, and Hae Kyung Im. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS genetics*, 15(1):e1007889, 2019.

15. Yiming Hu, Mo Li, Qiongshi Lu, Haoyi Weng, Jiawei Wang, Seyedeh M Zekavat, Zhaolong Yu, Boyang Li, Jianlei Gu, Sydney Muchnik, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. Technical report, Nature Publishing Group, 2019.

16. Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. *Nature communications*, 9(1):1825, 2018.

17. Hilary K Finucane, Yakir A Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*, 50(4):621, 2018.

18. Mingxuan Cai, Lin Chen, Jin Liu, and Can Yang. Quantifying the impact of genetically regulated expression on complex traits and diseases. *bioRxiv*, page 546580, 2019.

19. Eric R Gamazon, Ayellet V Segrè, Martijn van de Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics*, 50(7):956, 2018.

20. GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

21. Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Technical report, Nature Publishing Group, 2018.

22. John W Tukey. One degree of freedom for non-additivity. *Biometrics*, 5(3):232–242, 1949.

23. Nilanjan Chatterjee, Zeynep Kalaylioglu, Roxana Moslehi, Ulrike Peters, and Sholom Wacholder. Powerful multilocus tests of genetic association in the presence of gene-gene and

gene-environment interactions. *The American Journal of Human Genetics*, 79(6):1002–1016, 2006.

24. Jiebiao Wang, Eric R Gamazon, Brandon L Pierce, Barbara E Stranger, Hae Kyung Im, Robert D Gibbons, Nancy J Cox, Dan L Nicolae, and Lin S Chen. Imputing gene expression in uncollected tissues within and beyond gtex. *The American Journal of Human Genetics*, 98(4):697–708, 2016.

25. Yongjin Park, Abhishek K Sarkar, Kunal Bhutani, and Manolis Kellis. Multi-tissue polygenic models for transcriptome-wide association studies. *bioRxiv*, page 107623, 2017.

26. Chiara Sabatti, Susan K Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, Marika Kaakinen, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature genetics*, 41(1):35, 2009.

27. Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H ODonnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285, 2016.

28. Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821, 2012.

29. Heather E Wheeler, Kaanan P Shah, Jonathon Brenner, Tzintzuni Garcia, Keston Aquino-Michaels, Nancy J Cox, Dan L Nicolae, Hae Kyung Im, GTEx Consortium, et al. Survey of the heritability and sparse architecture of gene expression traits across human tissues. *PLoS genetics*, 12(11):e1006423, 2016.

30. Nicholas Mancuso, Huwenbo Shi, Pagé Goddard, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics*, 100(3):473–487, 2017.

31. De Jager, Philip L, Srivastava Gyan, Lunnon Katie, Burgess Jeremy, Leonard C Schalkwyk, Yu Lei, Matthew L Eaton, Brendan T Keenan, Ernst Jason, and Mc Cabe Cristin. Alzheimer's disease: early alterations in brain dna methylation at ank1, bin1, rhbdf2 and other loci. *Nature Neuroscience*, 17(9):1156–1163, 2014.

32. Alexei Kurakin and Dale E. Bredesen. Dynamic self-guiding analysis of alzheimer's disease. *Oncotarget*, 6(16), may 2015.

33. Rita R Romito-Digiacomo, Menegay Harry, Samantha A Cicero, and Herrup Karl. Effects of alzheimer's disease on different cortical layers: the role of intrinsic differences in abeta susceptibility. *Journal of Neuroscience the Official Journal of the Society for Neuroscience*, 27(32):8496–504, 2007.

34. M. J. Lin and S. J. Lee. Stathmin-like 4 is critical for the maintenance of neural progenitor cells in dorsal midbrain of zebrafish larvae. *Scientific Reports*, 6:36188, 2016.

35. Pei-Pei Liu, Ya-Jie Xu, Shang-Kun Dai, Hong-Zhen Du, Ying-Ying Wang, Xing-Guo Li, Zhao-Qian Teng, and Chang-Mei Liu. Polycomb protein EED regulates neuronal differentiation through targeting SOX11 in hippocampal dentate gyrus. *Stem Cell Reports*, 13(1):115–131, jul 2019.

36. Mariam Almuriekhi, Takafumi Shintani, Somayyeh Fahiminiya, Akihiro Fujikawa, Kazuya Kuboyama, Yasushi Takeuchi, Zafar Nawaz, Javad Nadaf, Hussein Kamel, Abu Khadija Kitam, Zaineddin Samiha, Laila Mahmoud, Tawfeg Ben-Omran, Jacek Majewski, and Masaharu Noda. Loss-of-function mutation in APC2 causes sotos syndrome features. *Cell Reports*, 10(9):1585–1598, mar 2015.

37. Jimmy Z Liu, Yaniv Erlich, and Joseph K Pickrell. Case–control association mapping by proxy using family history of disease. *Nature genetics*, 49(3):325, 2017.

38. A. K. Singh, P. Stock, and O. Akbari. Role of PD-l1 and PD-l2 in allergic diseases and asthma. *Allergy*, 66(2):155–162, aug 2010.

39. Florence Demenais, Patricia Margaritte-Jeannin, Kathleen C Barnes, William OC Cookson, Janine Altmüller, Wei Ang, R Graham Barr, Terri H Beaty, Allan B Becker, John Beilby, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nature genetics*, 50(1):42, 2018.

40. Chang-En Yu, Howard Seltman, Elaine R Peskind, Nichole Galloway, Peter X Zhou, Elisabeth Rosenthal, Ellen M Wijsman, Debby W Tsuang, Bernie Devlin, and Gerard D Schellenberg. Comprehensive analysis of apoe and selected proximate markers for late-onset

alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. *Genomics*, 89(6):655–665, 2007.

41. Aleksandra Maruszak and Sandrine Thuret. Why looking at the whole hippocampus is not enougha critical role for anteroposterior axis, subfield and activation analyses to enhance predictive value of hippocampal changes for alzheimers disease diagnosis. *Frontiers in cellular neuroscience*, 8:95, 2014.

42. Ville Leinonen, Anne M Koivisto, Sakari Savolainen, Jaana Rummukainen, Juuso N Tamminen, Tomi Tillgren, Sannakaisa Vainikka, Okko T Pyykkö, Juhani Mölsä, Mikael Fraunberg, et al. Amyloid and tau proteins in cortical brain biopsy and alzheimer's disease. *Annals of neurology*, 68(4):446–453, 2010.

43. Heidi IL Jacobs, David A Hopkins, Helen C Mayrhofer, Emiliano Bruner, Fred W van Leeuwen, Wijnand Raaijmakers, and Jeremy D Schmahmann. The cerebellum in alzheimers disease: evaluating its role in cognitive decline. *Brain*, 141(1):37–47, 2017.

44. Michelle M Stein, Emma E Thompson, Nathan Schoettler, Britney A Helling, Kevin M Magnaye, Catherine Stanhope, Catherine Igartua, Andréanne Morin, Charles Washington III, Dan Nicolae, et al. A decade of research on the 17q12-21 asthma locus: piecing together the puzzle. *Journal of Allergy and Clinical Immunology*, 142(3):749–764, 2018.

45. Jun Chen, Marina Miller, Hirotoshi Unno, Peter Rosenthal, Michael J Sanderson, and David H Broide. Orosomucoid-like 3 (ormdl3) upregulates airway smooth muscle proliferation, contraction, and ca2+ oscillations in asthma. *Journal of Allergy and Clinical Immunology*, 142(1):207–218, 2018.

46. Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2008.

47. GG Anderson and JFJ Morrison. Molecular biology and genetics of allergy and asthma. *Archives of disease in childhood*, 78(5):488–496, 1998.

48. Namita A Gandhi, Brandy L Bennett, Neil MH Graham, Gianluca Pirozzi, Neil Stahl, and George D Yancopoulos. Targeting key proximal drivers of type 2 inflammation in disease. *Nature reviews Drug discovery*, 15(1):35, 2016.

715 49. Mike Berry, Christopher Brightling, Ian Pavord, and Andrew J Wardlaw. Tnf-α in asthma.
716 *Current opinion in pharmacology*, 7(3):279–282, 2007.

717 50. Christopher Brightling, Mike Berry, and Yassine Amrani. Targeting tnf-α: a novel ther-
718 apeutic approach for asthma. *Journal of Allergy and Clinical Immunology*, 121(1):5–10,
719 2008.

720 51. James C Lo, Yugang Wang, Alexei V Tumanov, Michelle Bamji, Zemin Yao, Catherine A
721 Reardon, Godfrey S Getz, and Yang-Xin Fu. Lymphotoxin ß receptor–dependent control
722 of lipid homeostasis. *Science*, 316(5822):285–288, 2007.

723 52. H Mulcahy, KP O'rourke, C Adams, MG Molloy, and F O'gara. Lst1 and ncr3 expres-
724 sion in autoimmune inflammation and in response to ifn-γ, lps and microbial infection.
725 *Immunogenetics*, 57(12):893–903, 2006.

726 53. Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits:
727 from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

728 54. Xuanyao Liu, Yang I Li, and Jonathan K Pritchard. Trans effects on gene expression can
729 drive omnigenic inheritance. *Cell*, 177(4):1022–1034, 2019.

730 55. Chuanhai Liu, Donald B Rubin, and Ying Nian Wu. Parameter expansion to accelerate
731 em: the px-em algorithm. *Biometrika*, 85(4):755–770, 1998.

732 56. Xingjie Shi, Yuling Jiao, Yi Yang, Ching-Yu Cheng, Can Yang, Xinyi Lin, and Jin Liu.
733 VIMCO: variational inference for multiple correlated outcomes in genome-wide association
734 studies. *Bioinformatics*, 03 2019. btz167.

735 57. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira,
736 David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al.
737 Plink: a tool set for whole-genome association and population-based linkage analyses. *The*
738 *American journal of human genetics*, 81(3):559–575, 2007.

739 58. Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance
740 matrix estimation and implications for functional genomics. *Statistical applications in*
741 *genetics and molecular biology*, 4(1), 2005.

Table 1: Numbers of significant gene-trait associations across 15 NG traits. The reference penal is European subsamples from 1000 Genome. The number in the parenthesis denoted genes reported on the GWAS catalog. The full names of traits can be found in Supplementary Table S1.

	TisCoMM-S^2	S-MultiXcan	UTMOST
2hrGlu	1(0)	0(0)	5(0)
LOAD	92(24)	71(20)	70(19)
BMI	82(30)	59(21)	68(25)
CAD	69(9)	28(5)	36(11)
CD	468(52)	339(50)	291(55)
FG	94(11)	66(11)	54(8)
FI	3(0)	1(0)	2(0)
HDL-C	464(14)	268(13)	237(12)
HOMAB	10(0)	7(0)	4(0)
HOMAIR	1(0)	1(0)	1(0)
LDL-C	498(5)	273(5)	228(5)
TC	603(88)	376(77)	330(86)
TG	360(58)	250(48)	192(40)
UC	301(30)	262(31)	243(41)
WHR	12(3)	7(1)	8(1)

Table 2: Numbers of significant gene-trait associations across 18 UKB traits. The reference penal data is European subsamples from 1000 Genome. The number in the parenthesis denoted genes reported on the GWAS catalog. The full names of traits can be found in Supplementary Table S1.

	TisCoMM-S^2	S-MultiXcan	UTMOST
ALLERGIC_RHINITIS	25(3)	23(2)	12(4)
ASTHMA	200(31)	157(29)	140(36)
CARD	2(0)	2(0)	4(0)
DEPRESS	2(0)	1(0)	0(0)
DYSLIPID	166(0)	120(0)	91(0)
HEMORRHOIDS	0	1(0)	0
HERNIA_ABDOMINOPELVIC	2(1)	2(1)	1(1)
INSOMNIA	0	0	0
IRON_DEFICIENCY	0	0	0
IRRITABLE_BOWEL	0	1(0)	0
MACDEGEN	0	0	0
OSTIOA	1(0)	2(0)	2(1)
OSTIOP	0	1(0)	0
PEPTIC_ULCERS	1(0)	1(0)	0
PSYCHIATRIC	1	1	3
PVD	2(0)	3(0)	3(0)
STRESS	1(0)	1(0)	1(0)
VARICOSE_VEINS	40(2)	22(2)	20(2)

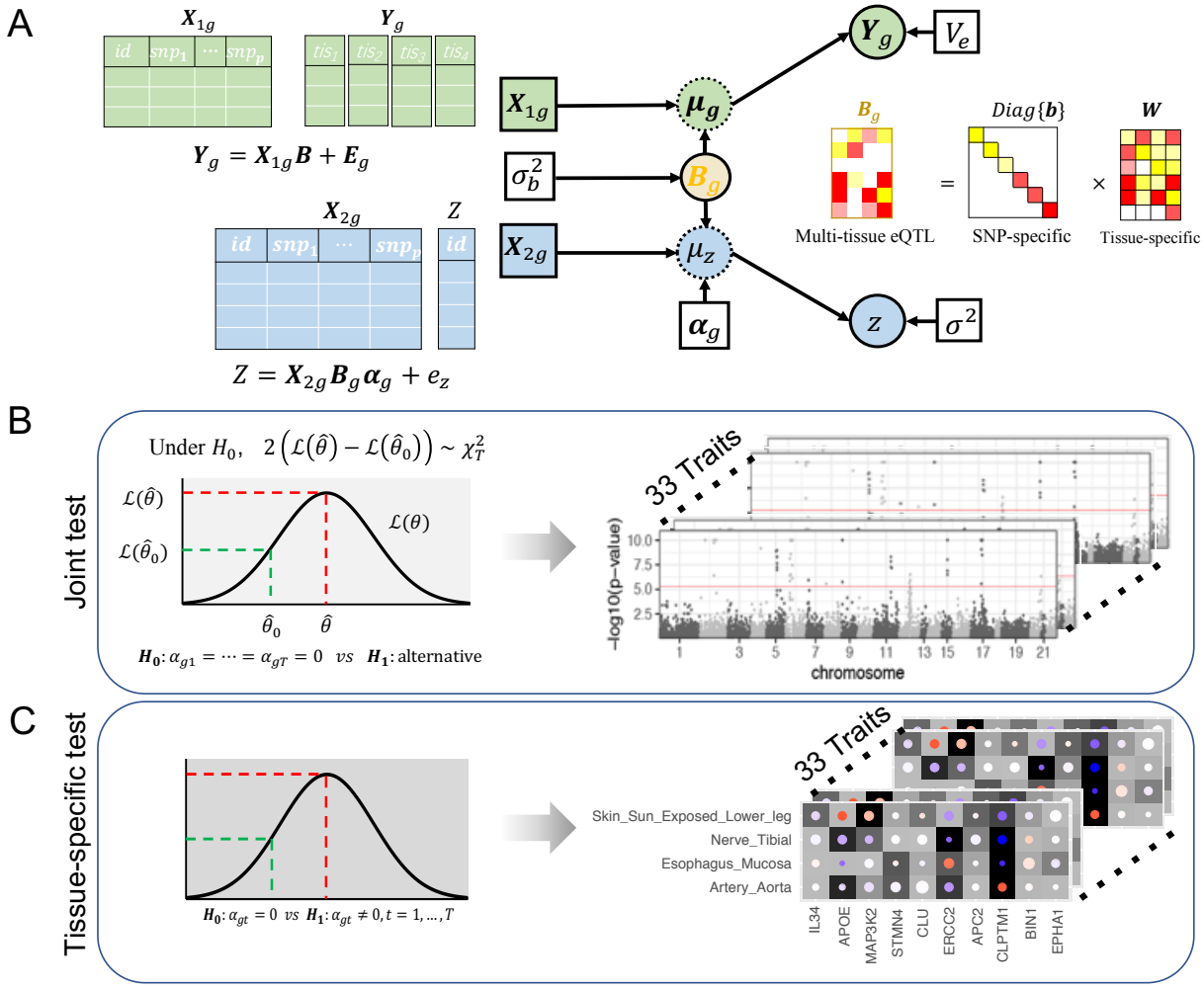


Figure 1: TisCoMM workflow. **A.** Two sets of TisCoMM input matrices are highlighted in green and blue separately (left). The probabilistic graphical model for TisCoMM is shown in the middle, which integrates gene expressions and models the co-regulation of cis-SNPs across different tissues explicitly. $\boldsymbol{\mu}_g$ and μ_z denote expectations of gene expression in eQTL and phenotype in GWAS, respectively. The decomposition of the **B** matrix is illustrated on the right-hand side of the figure. **B.** The TisCoMM joint test for all genes to prioritize candidate causal genes. See more details of $\mathcal{L}(\boldsymbol{\theta})$ in Methods section. The example outputs (right) are shown as Manhattan plots for 33 traits. **C.**The TisCoMM tissue-specific test for all candidate genes to explore the tissue-specific roles of candidate genes. The example outputs (right) are shown as heatmaps which summarize the tissue-specific effect of each gene. Significance level, effect size, and heritability are converted into background color, circle color, and circle size.

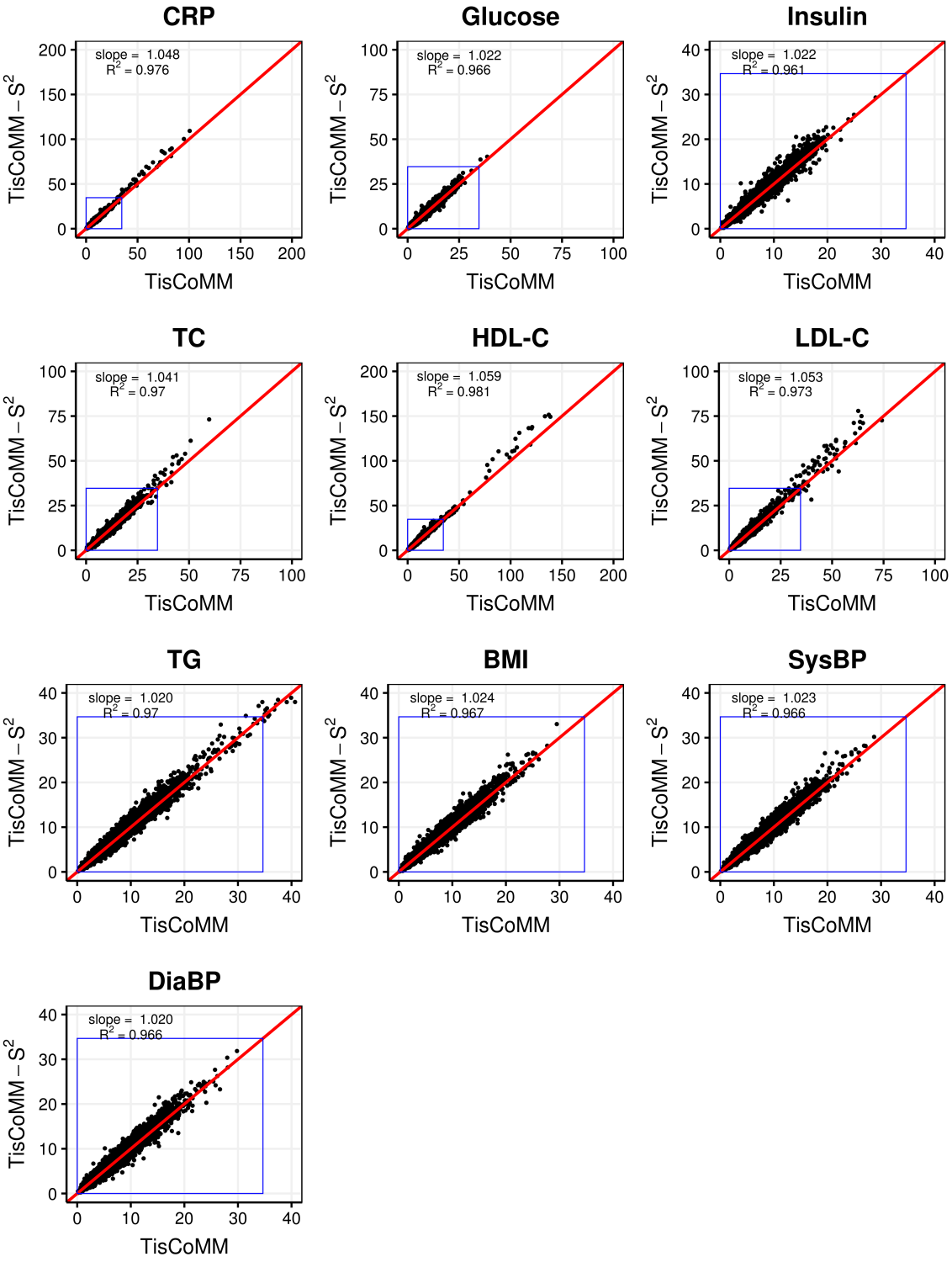


Figure 2: Comparison of TisCoMM and TisCoMM-S^2 results in NFBC1966 traits. The reference panel is subsamples from the NFBC1966 data set. The summary-based method shows similar results to the individual-based method. The blue rectangle indicates the null region.

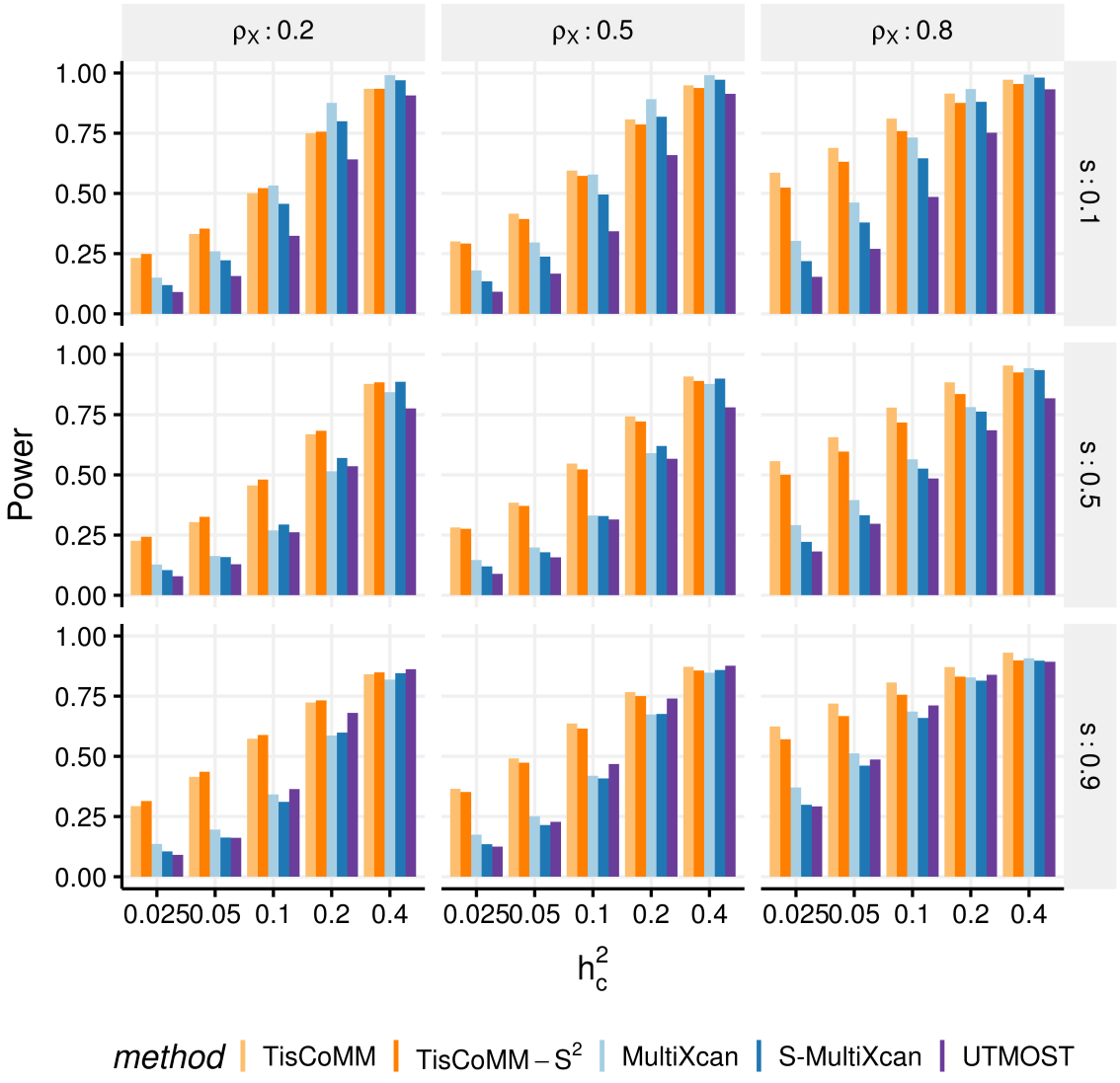
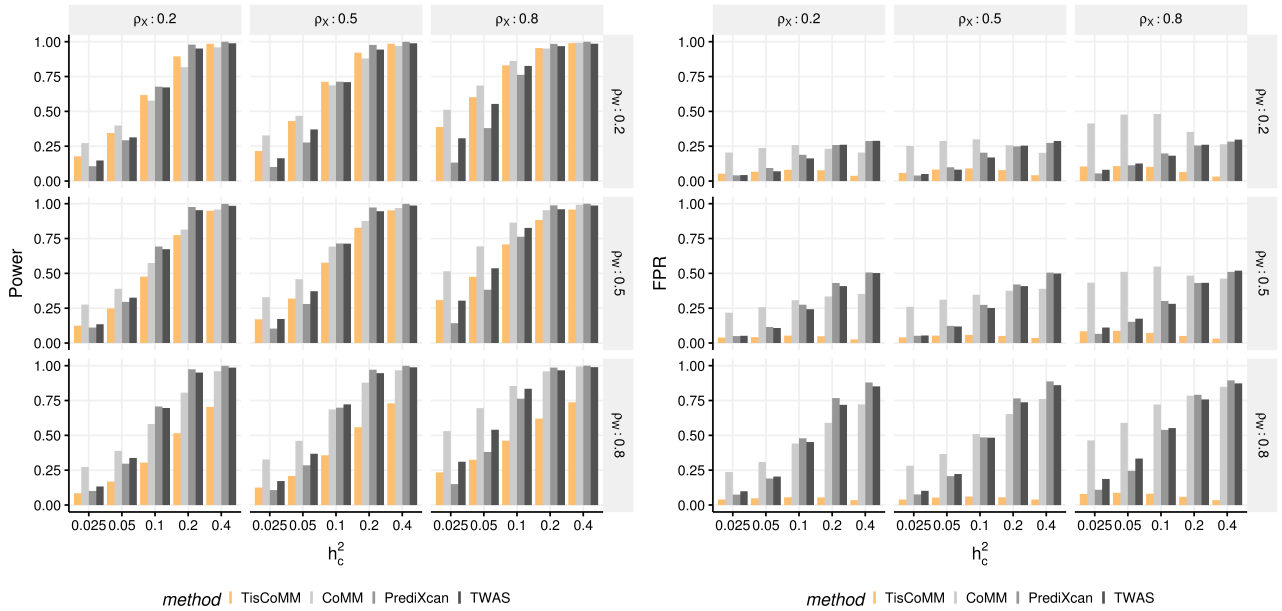


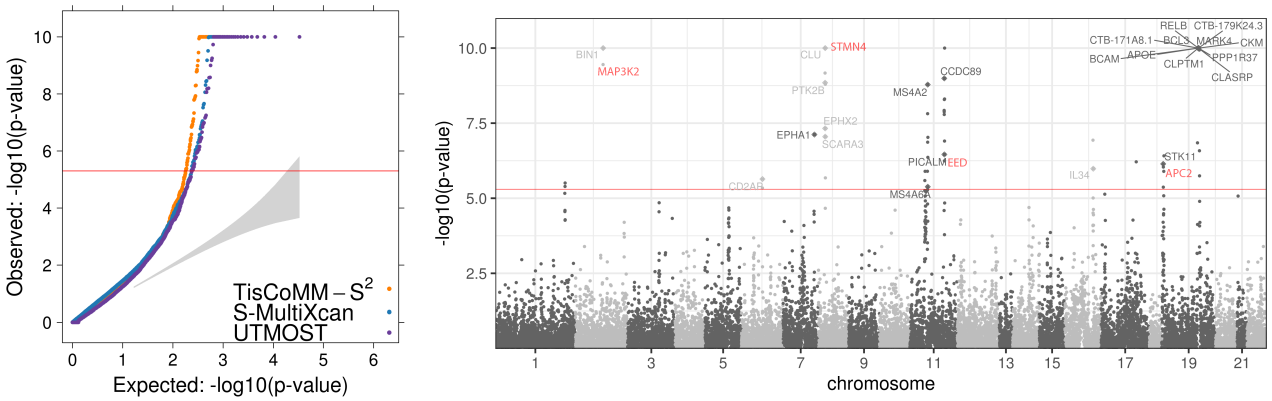
Figure 3: TisCoMM joint test outperforms the other multi-tissue methods. The number of replicates is 5,000. In each subplot, the x-axis stands for the SNP heritability level, and the y-axis stands for the proportion of significant genes within 5,000 replicates.



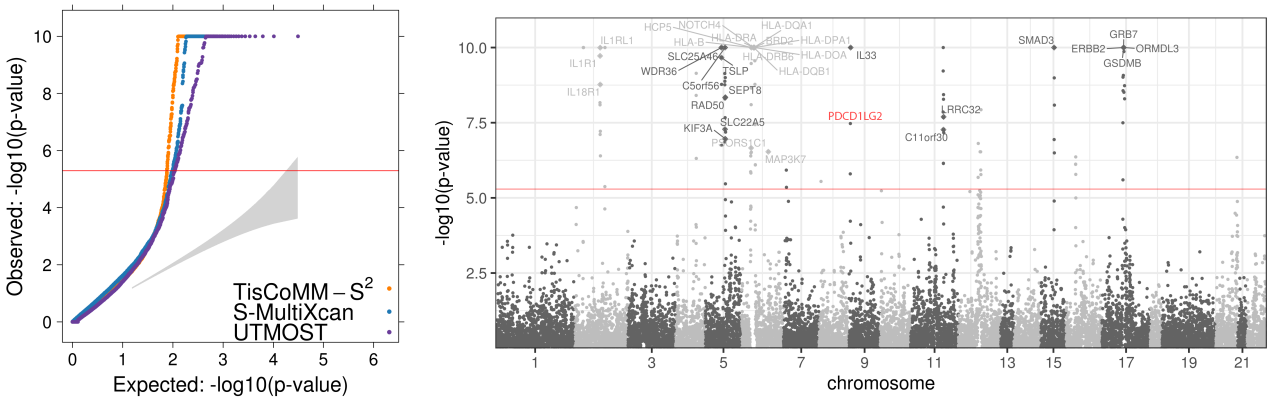
(A) Power

(B) False Positive Rate (FPR)

Figure 4: The comparison of the TisCoMM tissue-specific test and the single-tissue association tests under the alternative hypothesis with one causal tissue. **A.** The power of TisCoMM tissue-specific test and the single tissue methods with Bonferroni correction applied. **B.** The corresponding false positive rates under each setting.



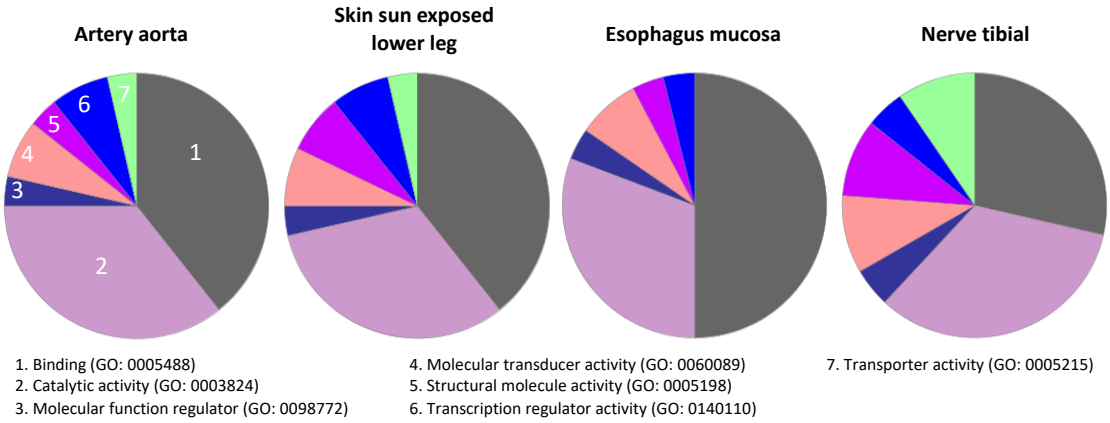
(A) LOAD



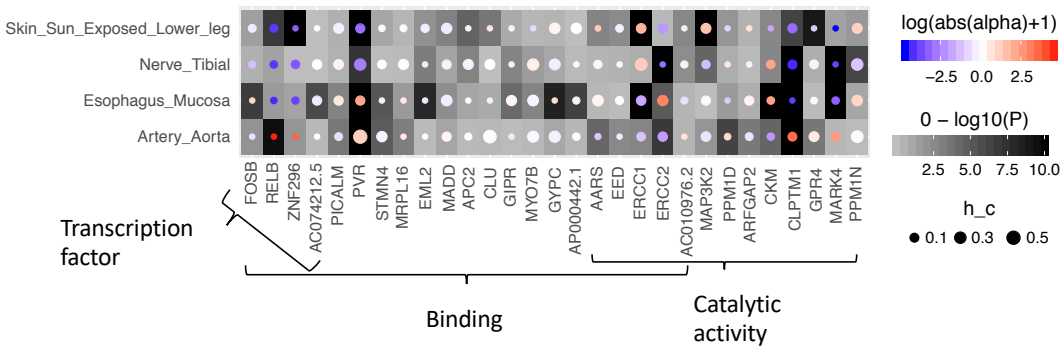
(B) Asthma

Figure 5: TisCoMM-S^2 results for LOAD and asthma. The reference panel is European subsamples from 1000 Genome. In each row, the two panels show the qq-plot (left) and Manhatton plot (right).

A



B



C

Pathway	Thyroid	Lung
1. Immune response	p (BH corrected) = 5.4E-6 (SMAD3,IL1RL1,IL1R1,LST1,LTA, HLA-DMA,HLA-DOB,HLA-DPB1, HLA-DQB1,HLA-DRA,NCR3)	p (BH corrected) = 0.0024 (IL1R1,IL18RAP,LST1,HLA-DOA HLA-DOB,HLA-DPB1,HLA-DRA, PDCD1LG2,TNF)
2. Antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	p (BH corrected) = 7.4E-6 (HLA-DMA,HLA-DOB,HLA-DPB1, HLA-DQB1,HLA-DRA)	< 5 genes
3. Antigen processing and presentation of exogenous peptide antigen via MHC class II	p (BH corrected) = 0.0051 (HLA-DMA,HLA-DOB,HLA-DPB1, HLA-DQB1,HLA-DRA)	< 5 genes

D

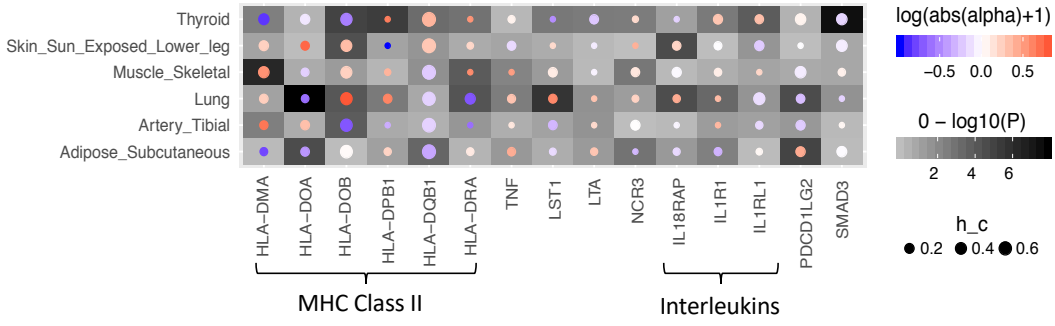


Figure 6: **A.** Each pie chart corresponding to a different tissue shows the percentage of LOAD-associated genes in each molecular function group (from gene ontology). **B.** The x-axis of the heatmap represents the union of LOAD-associated genes in 3 function groups (binding, catalytic activity, and transcription factor). The y-axis represents different tissue types. In each cell, the background color (shades of gray) indicates the significance level, the circle size indicates the heritability, and the color inside each circle indicates the effect size. **C.** Pathway analysis of asthma-associated genes in thyroid and lung. Pathway analysis was done using a web-based software DAVID, testing the enrichments of asthma-associated genes in biological processes (from gene ontology). Significant pathways were selected if gene count ≥ 5 and Benjamini-Hochberg (BH) corrected p-value ≤ 0.05. The asthma-associated genes are highlighted in blue. **D.** The x-axis of the heatmap represents the asthma-associated genes in the immune response pathway. And all the other settings are the same as the one used in part B.

Table 3: Distributions of tissues in which the candidate genes' associations arise in LOAD and asthma.

trait	#tissues	0	1	2	3	4	5	6
LOAD	TisCoMM-S^2	5	28	37	17	5	-	-
	CoMM-S	6	5	7	9	65	-	-
Asthma	TisCoMM-S^2	37	68	58	28	6	3	0
	CoMM-S	20	11	5	5	9	30	120