A statistical framework for cross-tissue transcriptome-wide association analysis

Yiming Hu[1,#], Mo Li[1,#], Qiongshi Lu[2,#], Haoyi Weng[3], Jiawei Wang[4], Seyedeh M. Zekavat[5], Zhaolong Yu[4], Boyang Li[1], Sydney Muchnik[6], Yu Shi[1], Brian W. Kunkle[7], Shubhabrata Mukherjee[8], Pradeep Natarajan[9,10,11,12], Adam Naj[14], Amanda Kuzma[14], Yi Zhao[14], Paul K. Crane[8], Alzheimer's Disease Genetics Consortium, Hongyu Zhao[1,4,6,13,*]

[1] Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA 06510

[2] Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA 53792

[3] Division of Biostatistics, The Jockey Club School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, NT, Hong Kong

[4] Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA 06510

[5] Yale School of Medicine, New Haven, CT, USA 06510

[6] Department of Genetics, Yale School of Medicine, New Haven, CT, USA 06510

[7] University of Miami Miller School of Medicine, Miami, FL, USA 33136

[8] Division of General Internal Medicine, Department of Medicine, University of Washington, Seattle, WA, USA 98195

[9] Department of Medicine, Harvard Medical School, Boston, MA, USA 02115

[10] Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA 02114

[11] Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA 02114

[12] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA 02142

[13] VA Cooperative Studies Program Coordinating Center, West Haven, CT, USA 06516

[14] School of medicine, University of Pennsylvania, Philadelphia, PA, USA 19104

[#] These authors contributed equally to this work

* To whom correspondence should be addressed:
Dr. Hongyu Zhao
Department of Biostatistics
Yale School of Public Health
60 College Street,
New Haven, CT, 06520, USA
hongyu.zhao@yale.edu

Key words: GWAS; gene-level association test; gene expression imputation; multi-tissue analysis; late-onset Alzheimer's disease

Abstract

Transcriptome-wide association analysis is a powerful approach to studying the genetic architecture of complex traits. A key component of this approach is to build a model to predict (impute) gene expression levels from genotypes from samples with matched genotypes and expression levels in a specific tissue. However, it is challenging to develop robust and accurate imputation models with limited sample sizes for any single tissue. Here, we first introduce a multi-task learning approach to jointly impute gene expression in 44 human tissues. Compared with single-tissue methods, our approach achieved an average 39% improvement in imputation accuracy and generated effective imputation models for an average 120% (range 13%-339%) more genes in each tissue. We then describe a summary statistic-based testing framework that combines multiple single-tissue associations into a single powerful metric to quantify overall gene-trait association at the organism level. When our method, called UTMOST, was applied to analyze genome wide association results for 50 complex traits (N_{total}=4.5 million), we were able to identify considerably more genes in tissues enriched for trait heritability, and cross-tissue analysis significantly outperformed single-tissue strategies (p=1.7e-8). Finally, we performed a cross-tissue genome-wide association study for late-onset Alzheimer's disease (LOAD) and replicated our findings in two independent datasets (N_{total}=175,776). In total, we identified 69 significant genes, many of which are novel, leading to novel insights on LOAD etiologies.

Introduction

Genome-wide association studies (GWASs) have successfully identified numerous single-nucleotide polymorphisms (SNPs) associated with complex human traits and diseases. Despite these successes, significant problems remain in statistical power and biological interpretation of GWAS results [1, 2]. In particular, the complex architecture of linkage disequilibrium (LD) and context-dependent regulatory machinery in the genome hinder our ability to accurately identify causal genes from GWAS, thereby raising challenges in downstream functional validation and therapeutic development. Recently, large-scale consortia, such as the Genotype-Tissue Expression (GTEx) project [3, 4], have generated matched genotype and expression data for various human tissues. These rich data sets have provided great insights into mechanisms of cross-tissue transcriptional regulation and accelerated discoveries for expression quantitative trait loci (eQTL) [4-7]. In addition, integrating eQTL information in genetic association analysis has become an effective way to bridge SNPs, genes, and complex traits. Many methods have been developed to co-localize eQTL with loci identified in GWAS to identify causal genes for complex traits [8-13]. Two recent studies addressed this issue through an innovative approach that is sometimes referred to as transcriptome-wide association analysis. First, based on an externally-trained imputation model, gene expression is imputed using genotype information in GWAS samples. Next, gene-level association is assessed between imputed gene expression and the trait of interest [14, 15]. These methods have become popular in the past two years due to their capability to effectively utilize signals from multiple eQTLs with moderate effects and to reduce the impact of reverse causality in expression-trait association analysis. The applications of these methods have led to novel insights into the genetic basis of many diseases and traits [16-18].

Despite these successes, the existing methods have several limitations. First, due to the tissue-dependent nature of transcription regulation, existing methods train separate imputation models for different tissues. This practice ignores similarity in transcription regulation across tissues, thereby limiting the effective sample sizes for tissues that are difficult to acquire. Second, a hypothesis-free search across genes

and tissues increases the burden of multiple testing and thus reduces statistical power. Pinpointing a subset of tissues based on prior knowledge may resolve this issue to some extent, but for many complex traits the biologically relevant tissues are either unknown or unavailable. Further, recent reports have shown that eQTL with large effects tend to regulate gene expression in multiple tissues [4]. Genetic correlation analysis has also suggested substantial sharing of local expression regulation across tissues [19]. This would inevitably result in statistically significant associations in tissues irrelevant to the complex trait, a phenomenon that has been extensively discussed recently [20]. Finally, the correlation between eQTL effects across tissues may be partially caused by sample overlap among tissues in the reference dataset, which further complicates the interpretation of association results. Jointly analyzing data from multiple genetically-correlated tissues has the potential to resolve these issues. It has been recently demonstrated that multi-trait analysis could improve accuracy of genetic risk prediction [21-23]. Multi-tissue modeling has also been shown to improve the statistical power in eQTL analysis [24-27] and gene network studies [28]. In this work, we demonstrate that a novel cross-tissue strategy could also improve transcriptome-wide association analysis.

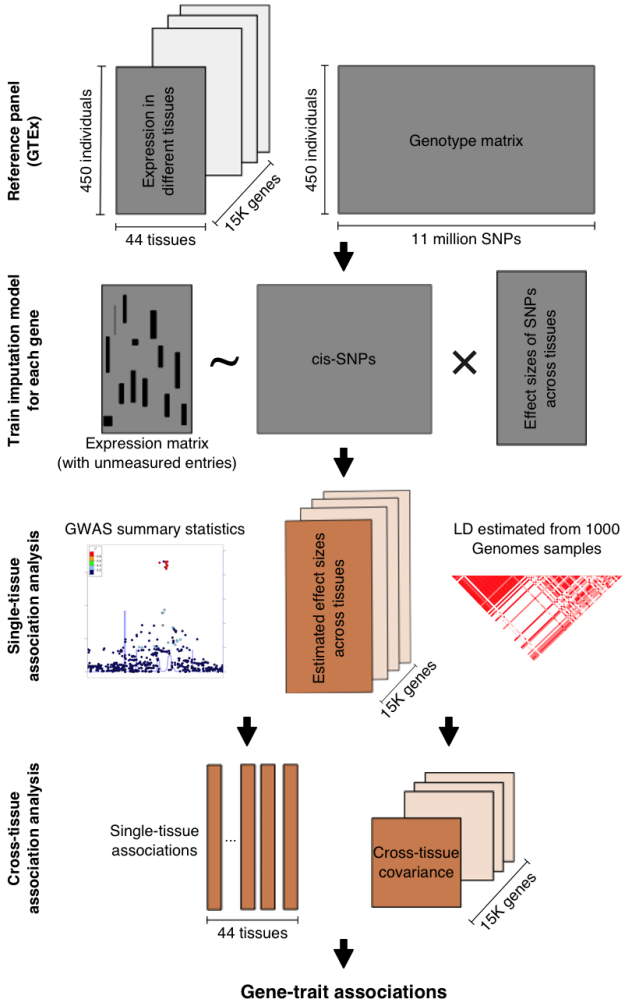
We introduce UTMOST (Unified Test for MOlecular SignaTures), a principled method to perform cross-tissue expression imputation and gene-level association analysis. We demonstrate its performance through internal and external imputation validation evaluations, numerical simulations, analyses of 50 complex traits, a case-study on low-density lipoprotein cholesterol (LDL-C), and a multi-stage association study for late-onset Alzheimer's disease (LOAD). We show that UTMOST substantially improves expression imputation accuracy in all tissues. For downstream analysis, UTMOST provides a powerful metric that summarizes gene-level associations across tissues and can be extended to integrate multiple molecular phenotypes.

Results

Model overview

The UTMOST framework consists of three main stages (**Figure 1**). First, for each gene in the genome, we train a cross-tissue expression imputation model using the genotype information and matched expression data from 44 tissues in GTEx. Next, we test associations between the trait of interest and imputed expression in each tissue. Lastly, a joint test is performed for each gene to summarize single-tissue association statistics into a powerful metric that quantifies the overall gene-trait association. Here, we briefly introduce the UTMOST framework. All the mathematical details are discussed in the **Methods** section.

Figure 1. UTMOST workflow. Gray and brown boxes denote input data and computed outcomes, respectively.



We formulate cross-tissue expression imputation as a penalized multivariate regression problem:

$$Y_{N\times P} = X_{N\times M} B_{M\times P} + \varepsilon_{N\times P},$$

where N, M, and P denote the sample size in the training data, the number of SNPs in the imputation model, and the total number of tissues, respectively. Since only a subset of tissues was collected from each individual, expression data in matrix Y were incomplete and sample sizes for different tissues were unbalanced. In order to effectively utilize information across tissues while addressing these issues, we estimate B by minimizing the squared loss function with a lasso penalty on the columns (within-tissue effects) and a group-lasso penalty on the rows (cross-tissue effects) via the following optimization problem (**Methods**).

$$\hat{B} = \underset{B}{\operatorname{argmin}} \sum_{i=1}^{P} \frac{1}{2N_i} \|Y_i - X_i B_{\cdot i}\|_2^2 + \lambda_1 \sum_{i=1}^{P} \frac{1}{N_i} \|B_{\cdot i}\|_1 + \lambda_2 \sum_{j=1}^{M} \|B_{j\cdot}\|_2$$

where Y_i, X_i, and N_i denote the observed expression, genotypes, and sample size of the *i*th tissue, respectively. Parameters λ_1 and λ_2 are tuned through cross-validation.

In the second stage, we test the associations between the trait of interest and imputed gene expression in each tissue. We denote imputed gene expression in the *i*th tissue as $E_i = X_i \hat{B}_{\cdot i}$ and test associations via a univariate regression model:

$$T = \alpha_i + E_i \gamma_i + \delta_i.$$

The z-scores for gene-trait associations in the *i*th tissue can be denoted as

$$Z_i = \frac{\hat{\gamma}_i}{se\ (\hat{\gamma}_i)} \approx \hat{B}_{\cdot i}^T \Gamma_i \tilde{Z}$$

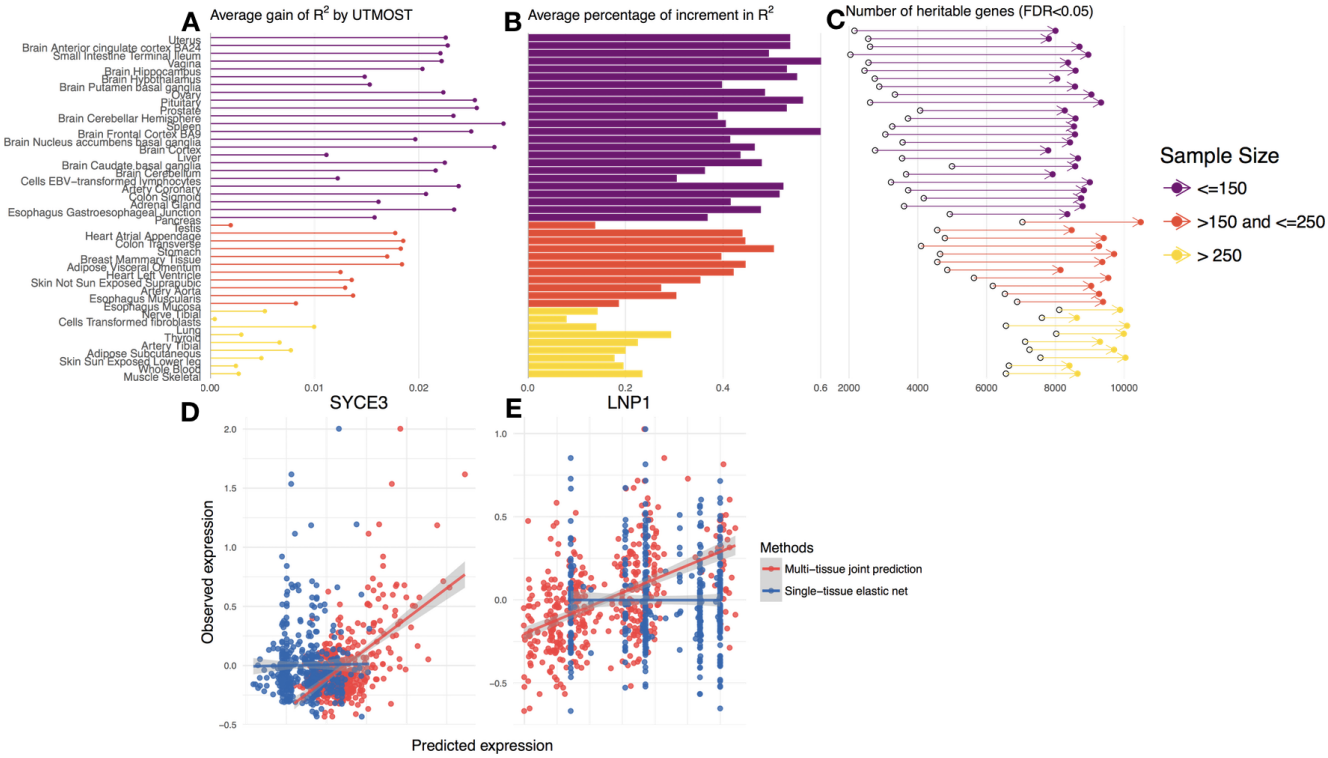
where $\tilde{Z}$ denotes the SNP-trait z-scores and Γ_i is a diagonal matrix whose *j*th diagonal element denotes the ratio between the standard deviation of the *j*th SNP and that of imputed expression in the *i*th tissue (**Methods**). When there is no SNP-trait association, $\tilde{Z}$ follows a multivariate normal distribution $N(0, D)$, where D is the LD matrix for SNPs. The covariance matrix of $Z = (Z_1, Z_2, \dots, Z_P)^T$ can be calculated as

$$\Sigma = cov\left(\Lambda^T \tilde{Z}\right) = \Lambda^T D \Lambda$$

where $\Lambda = (\hat{B}_{\cdot 1}\Gamma_1, \hat{B}_{\cdot 2}\Gamma_2, \dots, \hat{B}_{\cdot P}\Gamma_P)$.

Finally, we combine single-tissue gene-trait association results (i.e. Z) using a generalized Berk-Jones (GBJ) test, which takes the covariance among single-tissue test statistics (i.e. Σ) into account [29]. Details on the GBJ statistic and p-value calculation are discussed in the **Methods** section.

Figure 2. Improvement in gene expression imputation accuracy. Compared to single-tissue elastic net, UTMOST **(A-B)** showed substantially higher imputation accuracy and **(C)** identified more imputed genes, especially in tissues that have smaller sample sizes in GTEx. Panels **(D-E)** show the imputation improvement in two specific examples in whole blood tissue.



Cross-tissue analysis improves expression imputation accuracy

We first evaluated the accuracy of cross-tissue expression imputation through five-fold cross-validation on the training data. We used an elastic net model (i.e. the model used in PrediXcan [14]) trained in each tissue separately as the benchmark for prediction; PrediXcan does not leverage cross-tissue information. Cross-tissue imputation achieved higher imputation accuracy in all 44 tissues (**Figure 2A**). On average, imputation accuracy was improved by 38.6% across tissues (**Figure 2B**). The improvement was particularly high in tissues with low sample sizes in GTEx (N<150; 47.4% improvement). We further calculated the proportion of genes with increased imputation accuracy. Substantially more genes showed improved imputation performance in all 44 tissues (**Supplementary Table 1**). Using a false discovery rate (FDR) cutoff of 0.05 as the significance threshold, our cross-tissue method achieved 120% more significantly predicted genes across tissues. Among

tissues with low sample sizes, the improvement percentage rose even further to 175% (**Figure 2C**).

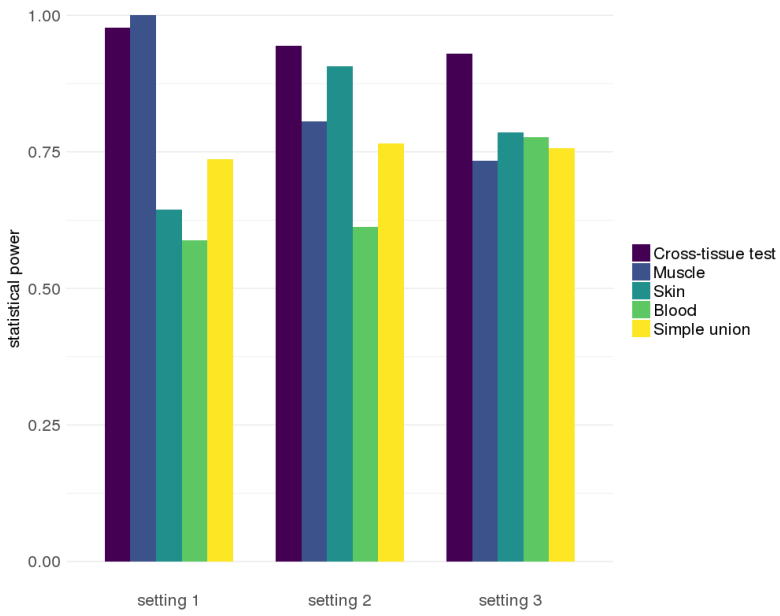
Next, we performed external validation using an independent dataset. We used our imputation model for whole blood in GTEx to predict gene expression levels in GEUVADIS LCLs (**Methods**; [30]). The imputation accuracy quantified as R^2 showed substantial departure from the expected distribution under the null (i.e. expression and SNPs are independent), which demonstrates the generalizability of cross-tissue imputation (**Supplementary Figure 1**). Compared to single-tissue elastic net, cross-tissue imputation achieved significantly higher prediction accuracy across genes (p=3.43e-7; Kolmogorov-Smirnov test), which is consistent with our findings from cross-validation. Two examples of well-predicted genes are illustrated in **Figures 2D-E**, showing improved concordance between observed (normalized gene expressions adjusted for potential confounding effects, **Methods**) and predicted expression values via cross-tissue imputation. These results demonstrate that the cross-tissue model could provide substantially and consistently more accurate expression imputation, especially in tissues with low sample sizes in the training dataset.

Multi-tissue association test provides statistically powerful inference results

Another key advancement in the UTMOST framework is a novel gene-level association test that combines statistical evidence across multiple tissues. We performed simulation studies using samples from the Genetic Epidemiology Research Study on Adult Health and Aging (GERA; N=12,637) to assess the association test's type-I error rate and statistical power in a variety of settings (**Methods**). We did not observe inflation in type-I error rate in two different simulation studies (**Supplementary Tables 2-3**). We observed substantial improvement in statistical power of the multi-tissue joint test when gene expression in multiple tissues were causally related to the trait (**Figure 3**). When the trait was determined by expression in only one tissue, statistical power of the joint test was comparable to a single-tissue test in the causal tissue. Compared to the naïve test that combines results across

tissues while applying an additional Bonferroni correction, our joint test was consistently more powerful (improvement ranged from 17.3% to 24.1%).

Figure 3. Cross-tissue analysis improves statistical power. We compared the statistical power of UTMOST, a single-tissue association test, and a simple union of findings from single-tissue analysis with Bonferroni correction applied. Muscle is the only causal tissue in setting 1. Both muscle and skin are causal tissues in setting 2. All three tissues are causal in setting 3.

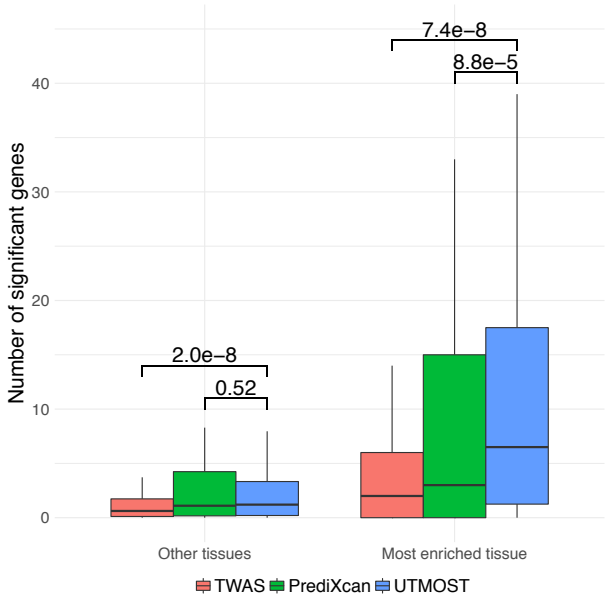


UTMOST identifies more associations in biologically relevant tissues for complex traits

To evaluate the performance of single-tissue association test based on cross-tissue expression imputation, we applied UTMOST to summary statistics from 50 GWAS (N_{total}≈4.5 million; **Supplementary Table 4**) and compared the results with those of PrediXcan [14] and TWAS [15]. To identify tissue types that are biologically relevant to these complex traits, we applied LD score regression [31] to these datasets and partitioned heritability by tissue-specific functional genome predicted by GenoSkyline-Plus annotations [32]. Tissue-trait relevance was ranked based on enrichment p-values (**Methods**). Compared to PrediXcan and TWAS, UTMOST identified substantially more associations in the most relevant tissue for each analyzed trait, showing 69.2% improvement compared to PrediXcan (p=8.79e-5;

paired Wilcoxon rank test) and 188% improvement compared to TWAS (p=7.39e-8). Such improvement was consistently observed across traits (**Supplementary Table 5**). On the contrary, among other tissues, UTMOST identified similar number of genes and showed no significant difference compared with PrediXcan (p=0.52). Comparing tissues that were most and least enriched for trait heritability, UTMOST identified significantly more associations in tissues strongly enriched for trait heritability than in tissues with the least enrichment (p=0.016) while the contrast was not significant based on PrediXcan (p=0.192) or TWAS (p=0.085). Finally, we applied the cross-tissue joint test to these traits and compared the number of significant genes with the combined results from 44 UTMOST single-tissue tests. UTMOST joint test identified more associations than single-tissue tests in 43 out of 50 traits (p=1.74e-8; binomial test; **Supplementary Figure 2**), showing improved statistical power in cross-tissue analysis.

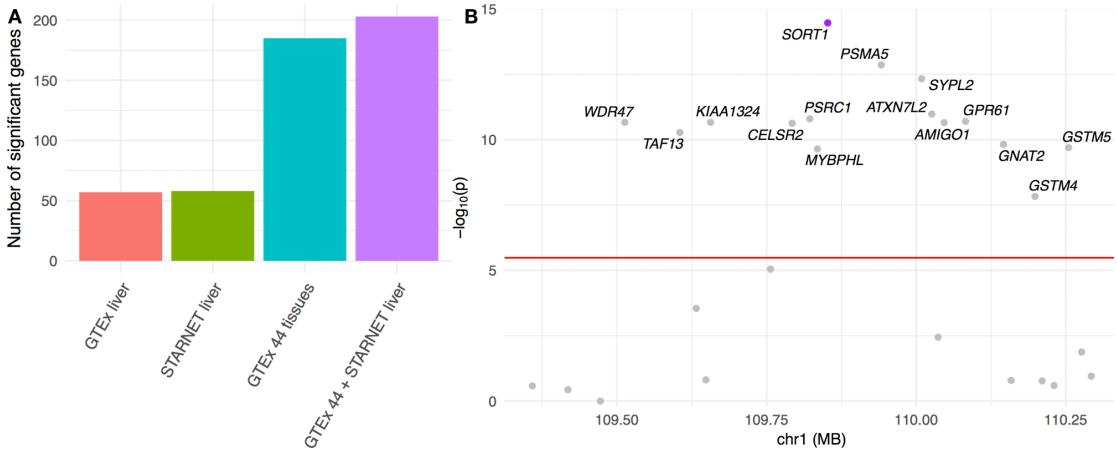
Figure 4. UTMOST identified more associations in biologically relevant tissues for 50 complex traits. Boxes on the right show the number of genes identified in the most relevant tissue for each trait. Boxes on the left show the number of genes identified in all other tissues. P-values were calculated via paired Wilcoxon rank tests.



Integrating external QTL resource further improves association performance

We applied UTMOST to the meta-analysis summary data of LDL cholesterol from the Global Lipids Genetics Consortium (N=173,082) [33]. Results based on four different analytical strategies, i.e. single-tissue test using liver tissue in GTEx (N=97), single-tissue test using liver eQTLs from STARNET (N=522; [34]), cross-tissue joint test combining 44 GTEx tissues, and cross-tissue joint test combining 44 GTEx tissues and the liver eQTLs from STARNET, were compared. We identified 57, 58, 185, and 203 significant genes in the four sets of analysis, respectively (**Figure 5A**).

Figure 5. Multi-tissue analysis identifies more associations for LDL cholesterol. (A) Number of significant genes identified in four sets of analysis. **(B)** Associations at the *SORT1* locus, values on the x-axis were based on the transcription start site of each gene. The horizontal line indicates the Bonferroni-corrected genome-wide significance threshold.



Among the identified genes in cross-tissue joint test of 44 GTEx tissues and STARNET-liver, *SORT1* had the most significant association (p=3.4e-15). *SORT1* is known to causally mediate LDL cholesterol levels, even though the GWAS association signal at this locus is clustered around the *CELSR2* gene [35, 36]. Of note, *SORT1* is not present in single tissue test with GTEx-liver due to its poor imputation quality (FDR > 0.05). In addition, its associations with LDL-C were not significant in tissues where the imputation was accurate (association p-values ranging from 0.01 in heart atrial appendage to 0.23 in tibial artery). As a result, UTMOST did not identify a strong association between *SORT1* and LDL-C even after combining results from 44 GTEx tissues. However, through integrating single-tissue associations in a large external liver dataset (STARNET), we successfully recovered the association of *SORT1*, which was the most significant at this locus and in the genome (**Figure 5B**).

These results suggest that integrative analysis of transciptomic data from multiple tissues can effectively increase statistical power in gene-level association mapping. Further, UTMOST is a flexible framework and is not limited to GTEx tissues only. Integrating relevant external QTL studies via UTMOST may further improve downstream association analysis.

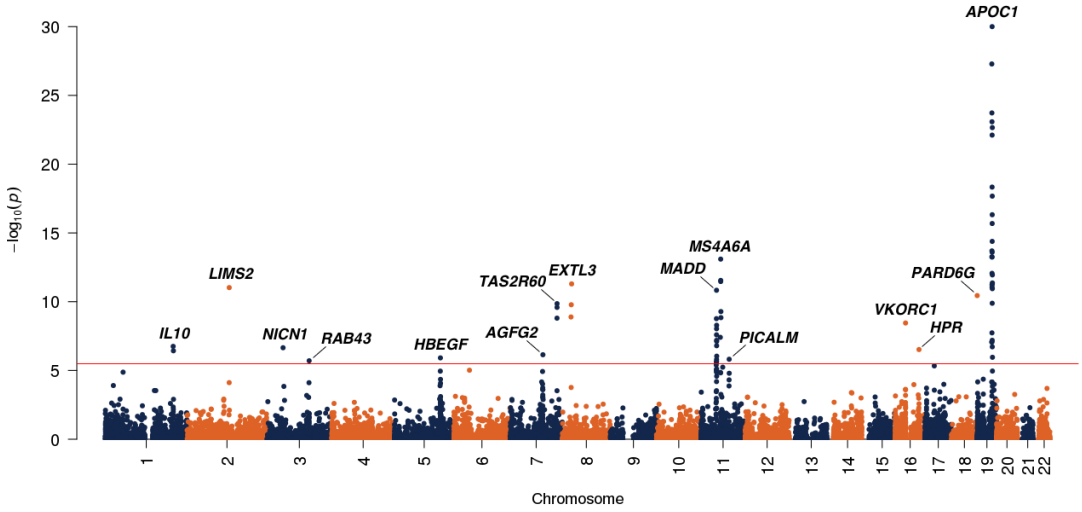
UTMOST identifies novel risk genes for Alzheimer's disease

Finally, to demonstrate UTMOST's effectiveness in real association studies, we performed a multi-stage gene-level association study for late-onset Alzheimer's disease (LOAD). In the discovery stage, we applied UTMOST to the stage-I GWAS summary statistics from the International Genomics of Alzheimer's Project (IGAP; N=54,162; [37]). Multiple recent studies have suggested that functional regions in the genome in liver and myeloid cells are strongly enriched for LOAD heritability [32, 38, 39]. It has also been suggested that alternative splicing may be a mechanism for many risk loci of LOAD [40]. Therefore, in addition to 44 tissues from GTEx, we also incorporated liver eQTLs from STARNET and both eQTL and sQTL data in three immune cell types (i.e. CD14+ monocytes, CD16+ neutrophils, and naive CD4+ T cells) from the BLUEPRINT consortium in our analysis (**Methods**). Single-tissue association tests were performed and then combined using the GBJ test. In total, our cross-tissue analysis identified 68 genome-wide significant genes in the discovery stage (**Supplementary Table 6** and **Supplementary Figure 3**).

We tried to replicate our findings in two independent datasets, i.e. GWAS summary statistics based on samples in the Alzheimer's Disease Genetics Consortium (ADGC) that were not used in the IGAP stage-I analysis (N=7,050), and summary statistics from the genome-wide association study by proxy (GWAX; N=114,564; [41]). Despite the moderate sample size in the ADGC dataset and the 'proxy' LOAD phenotype in GWAX analysis, replication rate was high (**Supplementary Table 6**). 17 and 15 out of 68 genes were successfully replicated under the Bonferroni-corrected significance threshold in ADGC and GWAX, respectively. The numbers of replicated genes rose to 41 and 30 under a relaxed p-value cutoff of 0.05. 22 out of 68 genes had p-values below 0.05 in both replication datasets. We then combined p-values from all three

analyses via Fisher's method. A total of 69 genes, including 12 genes that were not significant in the discovery stage, reached genome-wide significance in the meta-analysis (**Supplementary Tables 6-7** and **Figure 6**). These 69 genes were significantly enriched for seven gene ontology terms (**Supplementary Table 8**), with "very-low-density lipoprotein particle" being the most significant (adjusted p=5.8e-3).

Figure 6. Manhattan plot for LOAD meta-analysis. P-values are truncated at 1e-30 for visualization purpose. The horizontal line marks the genome-wide significance threshold. The most significant gene at each locus is labeled.



Most significant genes are from previously identified LOAD risk loci [37, 42-47]. These include *CR1* locus on chromosome 1, *BIN1* locus on chromosome 2, *HBEGF* locus on chromosome 5, *ZCWPW1* and *EPHA1* loci on chromosome 7, *CLU* locus on chromosome 8, *CELF1*, MS4A6A, and *PICALM* loci on chromosome 11, and the *APOE* region on chromosome 19. Among these loci, *AGFG2* rather than *ZCWPW1*, the previously-suggested index gene at this locus [37], was significant in the meta-analysis (p=7.19e-7). Similarly, *BIN1* was not statistically significant in our analysis. But *LIMS2*, a gene 500 kb upstream of *BIN1*, was significantly associated (p=9.43e-12). SNPs in the 3'UTR of *LIMS2* have been previously suggested to associate with cognitive decline [48]. GWAS index genes for the rest of the loci were all statistically significant in our analysis.

Further, new associations at known risk loci provide novel insights into LOAD etiology. We identified a novel gene *IL10* for LOAD risk (p=1.77e-7). *IL10* is 700 kb upstream of *CR1*, a well-replicated LOAD risk gene that is also significant in our analysis (p=3.71e-7) [37, 47, 49]. Although some SNPs near the promoter region of *IL10* were moderately associated with LOAD in all three datasets (**Supplementary Figure 4**), the *IL10*-LOAD association was mostly driven by SNPs near *CR1* (**Supplementary Table 9**). *IL10* is associated with multiple immune diseases [50-53] and is known to encode an anti-inflammatory cytokine that has therapeutic potential to improve neurodegeneration [54, 55]. Its protein product is also known to physically interact with Tau protein [56]. *CLU* is another well-replicated risk gene for LOAD. Two independent association peaks at this locus, one at *CLU* and the other at *PTK2B*, have previously been identified in GWAS (**Supplementary Figure 5**) [37, 47]. In our analysis, in addition to *CLU* (p=1.66e-10), we identified two more significant genes at this locus, i.e. *ADRA1A* (p=1.29e-9) and *EXTL3* (p=5.08e-12). *PTK2*B showed marginal association (1.72e-4) with LOAD but did not reach genome-wide significance. Interestingly, *EXTL3* expression is predicted by a SNP in the LOAD association peak at *CLU* while *ADRA1A* is regulated by SNPs at both *CLU* and *PTK2B* (**Supplementary Table 10**). *ADRA1A* has been implicated in gene-gene interaction analysis for LOAD [57]. Its protein product physically interacts with amyloid precursor protein (APP) [56] and an α_1-adrenoceptor antagonist has been shown to prevent memory deficits in APP23 transgenic mice [58]. *EXTL3* is a putative membrane receptor for regenerating islet-derived 1α (Reg-1α), whose overexpression and involvement in the early stages of Alzheimer's disease has been reported [59]. Further, the effect of Reg-1α on neurite outgrowth is mediated through *EXTL3*. Our results provide additional evidence that *IL10*, *ADRA1A*, and *EXTL3* may be involved in LOAD etiology.

Finally, we identified five novel loci for LOAD (**Supplementary Tables 6-7**), each represented by one significant gene: *NICN1* (p=2.23e-7), *RAB43* (p=1.98e-6), *VKORC1* (p=3.53e-9), *HPR* (p=3.02e-7), and *PARD6G* (p=3.60e-11). The Rab GTPases are central regulators of intracellular membrane trafficking [60]. Although *RAB43* has not been previously identified in LOAD GWAS, *USP6NL*, the gene that encodes a GTPase-activating protein for *RAB43*, has been identified to associate with

LOAD in two recent studies [41, 46]. *USP6NL* also showed suggestive association with LOAD in the discovery stage of our analysis (p=0.004). However, the associations at *RAB43* and *USP6NL* were not strongly supported by ADGC or GWAX datasets. Further, the *RAB43*-LOAD association was driven by SNPs near *RPN1*, a gene 400 kb downstream of *RAB43* (**Supplementary Figure 6** and **Supplementary Table 11**). This locus is associated with a variety of blood cell traits including monocyte count [61, 62]. *VKORC1* is a critical gene in vitamin K metabolism and is the target of warfarin [63], a commonly prescribed anticoagulant. It is known that the *APOE* ε4 allele affects the efficacy of warfarin [64]. *HPR* has been identified to strongly associate with multiple lipid traits [65] and interact with *APOE* [56]. *NICN1* is known to associate with inflammatory bowel disease [66] and cognitive function [67]. These results provide potential target genes for functional validations in the future.

Discussion

In this paper, we introduced UTMOST, a principled framework to perform cross-tissue expression imputation and gene-level association testing. Compared to single-tissue methods, cross-tissue imputation substantially improves imputation accuracy in all tissues in the training data. Through analysis of 50 complex traits, we demonstrated that single-tissue association analysis based on cross-tissue expression imputation could identify more associations in tissues relevant to trait etiology. In addition, we proposed a unified test that combines evidence of gene-trait association across multiple tissues. We demonstrated through both simulation and analysis of real GWAS data that our cross-tissue joint test is more powerful than single-tissue analysis, especially when a true association is present in multiple tissues. Further, UTMOST is capable of integrating external eQTL data to further enhance performance and can be expanded to incorporate any linearly imputed molecular phenotypes such as splicing or methylation QTL. Finally, we note that UTMOST only requires GWAS summary statistics as input.

To demonstrate the performance of UTMOST on real GWAS data, we performed a cross-tissue transcriptome-wide association meta-analysis for LOAD using three independent datasets. In total, we identified 69 genome-wide significant genes from 15 loci. A few novel genes are several hundred kb from known LOAD-associated SNPs and have not been previously identified as candidate risk genes for LOAD. In addition, we also identified five novel risk loci. Our results showcased UTMOST's effectiveness in interpreting GWAS associations and identifying novel genetic loci. More importantly, these findings provide insights into LOAD etiology and may guide future functional validations.

Despite the advancements, researchers need to be cautious when interpreting findings of UTMOST analyses. First, gene-level associations identified in UTMOST do not imply causality. It has been recently discussed that correlations among the imputed expression of multiple genes at the same locus may lead to apparent associations at non-causal genes [20], which is comparable to LD's impact on SNP-level associations in GWAS. Fine-mapping of causal genes in both single-tissue and cross-tissue analyses remains challenging. Second, we emphasize one of the principles in hypothesis testing – one should not conclude the null hypothesis when an association is not statistically significant. UTMOST is a general framework that involves many analytical steps, and technical issues might mask true gene-trait associations. For example, *SPI1* from the *CELF1* locus has been causally linked to LOAD risk [39]. We identified multiple significant associations at this locus but *SPI1* was not a significant gene in our analysis. Possible reasons for this include insufficient imputation quality based on the current model, non-availability of causal tissue in the training data, key eQTL missing from the GWAS summary statistics, causal mechanism (e.g. alternative splicing) not well-represented in our analysis, or insufficient sample sizes. In practice, these issues need to be carefully investigated before ruling out any candidate gene.

Nevertheless, UTMOST is a novel, powerful, and flexible framework to perform gene-level association analysis. It integrates biologically-informed weights with GWAS summary statistics via modern statistical techniques. Interpreted with caution, its findings may provide insights into disease and trait etiology, motivate downstream

functional validation efforts, and eventually benefit the development of novel therapeutics. It is also exciting that statistical and computational methodology in this field evolves at a fast pace. Several methods on mediation analysis and functional gene fine-mapping in the context of transcriptome-wide association study have been proposed recently [68, 69]. It has also been shown that data-adaptive SNP weights could effectively improve statistical power at the cost of clear interpretation of associations [70]. Extension of these methods into multi-tissue analysis is an interesting possible future direction. As high-throughput data continue to be generated for more individuals, cell types, and molecular phenotypes, UTMOST promises to show even better performance and provide greater insights for complex disease genetics in the future.

Methods

Penalized regression model for cross-tissue expression imputation

Given a gene, we use genotype information to predict its covariate-adjusted expression levels in P tissues. We use SNPs between 1 Mb upstream of the transcription start site and 1 Mb downstream of the transcription end site of the given gene as predictor variables in the model. This is denoted as an $N \times M$ matrix X where N is the total number of individuals and M denotes the number of SNPs. Throughout the paper, we assume each column of X to be centered but not standardized. Of note, expression data may not be available for all individuals since only a subset of tissues were collected from each individual. For the ith tissue, we use N_i to denote its sample size. We further use an N_i-dimensional vector Y_i to denote the observed expression data in the ith tissue, and use an $N_i \times M$ matrix X_i to denote the genotype information for the subset of individuals. Then, cross-tissue gene expression imputation can be formulated as the following regression problem.

$$Y_i = X_i B_{\cdot i} + \varepsilon_i\,, \;\; i = 1, \dots, P.$$

Here, the $M \times P$ matrix B summarizes SNPs' effects on the given gene with its ith column $B_{\cdot i}$ denoting the effect sizes of SNPs in the ith tissue and the jth row $B_{j\cdot}$ denoting the effect sizes of the jth SNP in all P tissues. To effectively select

biologically relevant and statistically predictive SNPs, accurately estimate their effects across tissues, and address technical issues including shared samples and incomplete data, we propose the following penalized least-squares estimator for genetic effects matrix B:

$$\hat{B} = \underset{B}{\operatorname{argmin}} \sum_{i=1}^{P} \frac{1}{2N_i} \|Y_i - X_i B_{\cdot i}\|_2^2 + \lambda_1 \sum_{i=1}^{P} \frac{1}{N_i} \|B_{\cdot i}\|_1 + \lambda_2 \sum_{j=1}^{M} \|B_{j\cdot}\|_2$$

Here, $\|.\|_1$ and $\|.\|_2$ denote the l_1 and l_2 norms, respectively (i.e. $\|x_{V\times 1}\|_1 = \sum_{v=1}^{V} |x_v|$ and $\|x_{V\times 1}\|_2 = \sqrt{\sum_{v=1}^{V} x_v^2}$). The first term in the loss function is the standard least-squares error. We use the l_1 penalty to select predictive variables and impose shrinkage in effect size estimation. The penalty on each tissue is set adaptively based on the sample sizes, which reflects the idea that models for tissues with a larger sample size are more robust to overfitting and therefore are penalized less. To integrate information across multiple tissues, we introduced the third term - a group-lasso penalty on the effect sizes of one SNP [71]. By imposing this joint penalty across tissues, we could effectively eliminate non-eQTL SNPs especially in tissues with small sample sizes and keep strong cross-tissue effects. Of note, the bias-variance tradeoff needs to be taken into consideration to achieve better prediction accuracy. Given the large number of tissues and the limited yet imbalanced sample size for each tissue in GTEx data, we chose to apply a cross-tissue penalty instead of explicitly modeling tissue-tissue correlations to avoid overfitting. Tuning parameters λ_1 and λ_2 control the within-tissue and cross-tissue sparsity, respectively. They are selected through cross-validation.

Optimization

To minimize the objective function, we developed a novel coordinate descent approach that extended the optimization procedure in [72] to deal with incomplete data in the outcome matrix. In each iteration, we update a row of B. When $B_{(-j)\cdot}$ (i.e. all the rows except for the jth) have been updated, the procedure is equivalent to minimizing

$$L(B_{j\cdot}) := L(B_{j\cdot}; B_{(-j)\cdot}, Y_1, \dots, Y_P, X_1, \dots, X_P)$$

$$= \sum_{i=1}^{P} \frac{1}{2N_i} \left\| \tilde{Y}_i - (X_i)_{\cdot j} B_{ji} \right\|_2^2 + \lambda_1 \sum_{i=1}^{P} \frac{1}{N_i} \left| B_{ji} \right| + \lambda_2 \sqrt{\sum_{i=1}^{P} B_{ji}^2}$$

where $\tilde{Y}_i = Y_i - (X_i)_{\cdot(-j)} B_{(-j)i}$ and $B_{j\cdot} = (B_{j1}, \dots, B_{jP})$. $(X_i)_{\cdot j}$ and $(X_i)_{\cdot(-j)}$denote the *j*th column and the matrix excluding the *j*th column of X_i $(N_i \times M)$, respectively. For a specific $i \in (1, \dots, P)$, when $B_{ji} > 0$, its partial derivative is

$$\frac{\partial L}{\partial B_{ji}} = \frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} B_{ji} - \frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i} + \frac{\lambda_1}{N_i} + \frac{\lambda_2}{\left\| B_{j\cdot} \right\|_2} B_{ji}$$

$$= \left(\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\left\| B_{j\cdot} \right\|_2} \right) B_{ji} - \left(\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i} - \frac{\lambda_1}{N_i} \right)$$

Therefore, when $B_{ji} > 0$, condition $\frac{\partial L}{\partial B_{ji}} > 0$ is equivalent to $B_{ji} > \tilde{B}_{ji}^+$, where $\tilde{B}_{ji}^+$ is defined as

$$\tilde{B}_{ji}^+ := \frac{\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i} - \frac{\lambda_1}{N_i}}{\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\left\| B_{j\cdot} \right\|_2}} = \frac{\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i}}{\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\left\| B_{j\cdot} \right\|_2}} \left(1 - \frac{\lambda_1}{(X_i)_{\cdot j}^T \tilde{Y}_i} \right)$$

And the minimum of $L(B_{ji}; B_{j(-i)})$ with $B_{ji} > 0$ and $B_{j(-i)}$ fixed is

$$\hat{B}_{ji,min}^+ := \begin{cases} \tilde{B}_{ji}^+, \ \ if \ \frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i} \left(1 - \frac{\lambda_1}{(X_i)_{\cdot j}^T \tilde{Y}_i} \right) > 0 \\ 0, \ \ if \ \frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i} \left(1 - \frac{\lambda_1}{(X_i)_{\cdot j}^T \tilde{Y}_i} \right) \leq 0 \end{cases}$$

Similarly, when $B_{ji} \leq 0$, we have

$$\frac{\partial L}{\partial B_{ji}} = \left(\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\left\| B_{j\cdot} \right\|_2} \right) B_{ji} - \left(\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i} + \frac{\lambda_1}{N_i} \right)$$

Condition $\frac{\partial L}{\partial B_{ji}} > 0$ is equivalent to $B_{ji} > \tilde{B}_{ji}^-$, where $\tilde{B}_{ji}^-$ is defined as

$$\tilde{B}_{ji}^- := \frac{\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i} + \frac{\lambda_1}{N_i}}{\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\left\| B_{j\cdot} \right\|_2}} = \frac{\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i}}{\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\left\| B_{j\cdot} \right\|_2}} \left(1 + \frac{\lambda_1}{(X_i)_{\cdot j}^T \tilde{Y}_i} \right)$$

The minimum of $L(B_{ji}; B_{j(-i)})$ with $B_{ji} \leq 0$ and $B_{j(-i)}$ fixed is

$$\hat{B}_{ji,min}^{-} := \begin{cases} \tilde{B}_{ji}^{-}, if \ \frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i}\left(1 + \frac{\lambda_1}{(X_i)_{\cdot j}^T \tilde{Y}_i}\right) < 0 \\ 0, if \ \frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i}\left(1 + \frac{\lambda_1}{(X_i)_{\cdot j}^T \tilde{Y}_i}\right) \geq 0 \end{cases}$$

Denote the minimum of $L(B_{ji}; B_{j(-i)})$ with $B_{j(-i)}$ fixed as $\hat{B}_{ji,min}$. If $(X_i)_{\cdot j}^T \tilde{Y}_i > 0$, and $\frac{\partial L}{\partial B_{ji}} < 0$ for $B_{ji} < 0$, then $B_{ji,min} \geq 0$ and

$$\hat{B}_{ji,min} = \hat{B}_{ji,min}^{+} = \frac{\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i}}{\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\|B_{j\cdot}\|_2}}\left(1 - \frac{\lambda_1}{\left|(X_i)_{\cdot j}^T \tilde{Y}_i\right|}\right)_+$$

Similarly, if $(X_i)_{\cdot j}^T \tilde{Y}_i \leq 0$ and $\frac{\partial L}{\partial B_{ji}} \geq 0$ for $B_{ji} \geq 0$, then $\hat{B}_{ji,min} \leq 0$ and

$$\hat{B}_{ji,min} = \hat{B}_{ji,min}^{-} = \frac{\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i}}{\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\|B_{j\cdot}\|_2}}\left(1 - \frac{\lambda_1}{\left|(X_i)_{\cdot j}^T \tilde{Y}_i\right|}\right)_+$$

which is the same formula as in the previous case. Denoting the minimum of $L(B_{j\cdot})$ as $\hat{B}_{j\cdot,min} := (\hat{B}_{j1,min}, \dots, \hat{B}_{jP,min})$, we have shown that if such a minimum exists, it should satisfy:

$$\hat{B}_{ji,min} = \frac{\frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{N_i}}{\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\|\hat{B}_{j\cdot,min}\|_2}}\left(1 - \frac{\lambda_1}{\left|(X_i)_{\cdot j}^T \tilde{Y}_i\right|}\right)_+ = \frac{\frac{X_{ij\cdot}^T X_{ij\cdot}}{N_i}}{\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i} + \frac{\lambda_2}{\|\hat{B}_{j\cdot,min}\|_2}} \hat{B}_{ji}^{lasso}$$

where $\hat{B}_{ji}^{lasso} = \frac{(X_i)_{\cdot j}^T \tilde{Y}_i}{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}\left(1 - \frac{\lambda_1}{\left|(X_i)_{\cdot j}^T \tilde{Y}_i\right|}\right)_+$ and $i = 1, \dots, P$. And minimizing $L(B_{j\cdot})$ is the equivalent to solving the above non-linear equation system. However, this system does not have closed-form solution and solving it numerically is very computationally expensive, especially considering we need to train over 15,000 models. To address this challenge, we proposed a efficient procedure to approximate $\hat{B}_{ji,min}$.

Note that $\frac{(X_i)_{\cdot j}^T (X_i)_{\cdot j}}{N_i}$ is the diagonal term of the LD matrix estimated by N_i individuals, we could thus approximate it with the LD matrix estimated with entire genotype matrix, i.e. $\frac{X_{\cdot j}^T X_{\cdot j}}{N}$, and we got a new non-linear equation systems, which is close to the original

one but has a closed-form solution.

$$\hat{B}_{ji,min} \approx \frac{\frac{X_{\cdot j}^T X_{\cdot j}}{N}}{\frac{X_{\cdot j}^T X_{\cdot j}}{N} + \frac{\lambda_2}{\left\|\hat{B}_{j\cdot,min}\right\|_2}} \hat{B}_{ji}^{lasso}, \quad i = 1, \dots, P$$

By squaring both sides of the approximation and summing over i, we obtain the following approximation:

$$\left\|\hat{B}_{j\cdot,min}\right\|_2 \approx \frac{\frac{X_{\cdot j}^T X_{\cdot j}}{N}}{\frac{X_{\cdot j}^T X_{\cdot j}}{N} + \frac{\lambda_2}{\left\|\hat{B}_{j\cdot,min}\right\|_2}} \left\|\hat{B}_{j\cdot}^{lasso}\right\|_2$$

where $\hat{B}_{j\cdot}^{lasso} = (\hat{B}_{j1}^{lasso}, \dots, \hat{B}_{jP}^{lasso})$. That is

$$\left\|\hat{B}_{j\cdot,min}\right\|_2 \approx \left\|\hat{B}_{j\cdot}^{lasso}\right\|_2 - \frac{N\lambda_2}{X_{\cdot j}^T X_{\cdot j}}$$

And given $B_{(-j)\cdot}$, $B_{j\cdot}$ is updated by

$$\hat{B}_{j\cdot} = \left(1 - \frac{\lambda_2}{\left\|\hat{B}_{j\cdot}^{lasso}\right\|_2 \frac{X_{\cdot j}^T X_{\cdot j}}{N}}\right) \hat{B}_{j\cdot}^{lasso}$$

We keep training the model until convergence in B or the mean squared error starts increasing in the validation set. Numerical experiments demonstrated that the proposed procedure provided reasonable approximation in each iteration and resulted in robust prediction models (**Figure 2** and **Supplementary Table 1**).

Model training and evaluation

We trained our cross-tissue gene expression imputation model using genotype and normalized gene expression data from 44 tissues in the GTEx project (version V6p) [3]. Sample sizes for different tissues ranged from 70 (uterus) to 361 (skeletal muscle). SNPs with ambiguous alleles or minor allele frequency (MAF) < 0.01 were removed. Normalized gene expressions were further adjusted to remove potential confounding effects from sex, sequencing platform, top three principal components of genotype data, and top probabilistic estimation of expression residuals (PEER) factors [73]. As previously recommended [17], we included 15 PEER factors for tissues with $N < 150$, 30 factors for tissues with $150 \leq N < 250$, and 35 factors for tissues with $N \geq 250$. All covariates were downloaded from the GTEx portal website

(https://www.gtexportal.org). We applied a 5-fold cross-validation for model tuning and evaluation. Specifically, we randomly divided individuals into five groups of equal size. Each time, we used three groups as the training set, one as the intermediate set for selecting tuning parameters, and the last one as the testing set for performance evaluation. Squared correlation between predicted and observed expression (i.e. R^2) was used to quantify imputation accuracy. We used a paired Wilcoxon rank test to compare the performance of different imputation methods across all genes in each tissue. For each model, we selected gene-tissue pairs with FDR < 0.05 for downstream testing. External validation of imputation accuracy was performed using whole-blood expression data from 421 samples in the 1000 Genomes Project (GEUVADIS consortium) [30].

Gene-level association test

We combined GWAS summary statistics with SNP effects estimated in the cross-tissue imputation model (i.e. $\hat{B}$) to quantify gene-trait associations in each tissue. For a given gene, we modeled its imputed expression in the ith tissue (i.e. $E_i = X_i \hat{B}_{\cdot i}$) and the phenotype T using a linear model

$$T = \alpha_i + E_i \gamma_i + \delta_i$$

Then, the association statistic for effect size in the ith tissue (i.e. γ_i) on the trait of interest is

$$Z_i = \frac{\hat{\gamma}_i}{se\,(\hat{\gamma}_i)}$$

where $\hat{\gamma}_i$ denotes the point estimate for effect size and $se\,(\hat{\gamma}_i)$ denotes its standard error. From the linear model, we have

$$\hat{\gamma}_i = \frac{cov(E_i, T)}{var(E_i)} = \frac{\hat{B}_{\cdot i}^T cov(X_i, T)}{\eta_i^2} = \hat{B}_{\cdot i}^T \Gamma_i^2 \tilde{\beta}$$

where Γ_i is an $M \times M$ diagonal matrix with the jth term equal to $\frac{\sigma_j}{\eta_i}$, where σ_j is the standard deviation of the jth SNP, and η_i is the standard deviation of imputed gene expression in the ith tissue. These parameters could be estimated using a reference panel. $\tilde{\beta}$ denotes the SNP-level effect size estimates acquired from GWAS summary statistics. Regarding the standard error of $\hat{\gamma}_i$, we have

$$\mathrm{se}(\hat{\gamma}_i) = \sqrt{\frac{var(\delta_i)}{N_{gwas}\eta_i^2}} \approx \frac{\sigma_Y}{\sqrt{N_{gwas}}\eta_i}$$

Here, σ_Y denotes the standard deviation of phenotype T and N_{gwas} is the sample size in GWAS. The approximation $var(\delta_i) \approx \sigma_Y^2$ is based on the empirical observation that each gene only explains a very small proportion of phenotypic variability [74]. The same argument can be extended to association statistics at the SNP level. For the jth SNP in the model, we have

$$\mathrm{se}(\tilde{\beta}_j) \approx \frac{\sigma_Y}{\sqrt{N_{gwas}}\sigma_j}.$$

Therefore, SNP-level z-scores can be denoted as

$$\tilde{Z}_j = \frac{\tilde{\beta}_j}{\mathrm{se}(\tilde{\beta}_j)} \approx \frac{\sqrt{N_{gwas}}\sigma_j\tilde{\beta}_j}{\sigma_Y}, \qquad j = 1, \dots, M$$

In matrix form, this is

$$\tilde{Z} \approx \frac{\sqrt{N_{gwas}}}{\sigma_Y}\begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_M \end{pmatrix}\tilde{\beta}$$

Combining the derivations above, we can denote the gene-level z-score as

$$Z_i = \frac{\hat{\gamma}_i}{se\ (\hat{\gamma}_i)} \approx \hat{B}_{\cdot i}^T \Gamma_i^2 \tilde{\beta} \times \frac{\sqrt{N_{gwas}}\eta_i}{\sigma_Y} = \frac{\sqrt{N_{gwas}}}{\sigma_Y}\hat{B}_{\cdot i}^T \Gamma_i \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_M \end{pmatrix}\tilde{\beta} \approx \hat{B}_{\cdot i}^T \Gamma_i \tilde{Z}$$

Under the null hypothesis (i.e. no SNP-trait association), $\tilde{Z}$ follows a multivariate normal distribution $\tilde{Z} \sim N(0, D)$, where D is the LD matrix for SNPs and could be estimated using an external reference panel. Denoting the cross-tissue gene-trait z-scores as $Z = (Z_1, Z_2, \dots, Z_P)^T$, the covariance matrix of Z could be calculated as

$$\Sigma = cov(\Lambda^T \tilde{Z}) = \Lambda^T D \Lambda,$$

where $\Lambda = (\hat{B}_{\cdot 1}\Gamma_1, \hat{B}_{\cdot 2}\Gamma_2, \dots, \hat{B}_{\cdot P}\Gamma_P)$.

In order to combine gene-trait associations across multiple tissues, we applied the generalized Berk-Jones (GBJ) test with single-tissue association statistics Z and their covariance matrix Σ as inputs. This approach provides powerful inference results while explicitly taking the correlation among single-tissue test statistics into account even under a sparse alternative (i.e. biologically meaningful associations are

only present in a small number tissues) [29]. The GBJ test statistic can be calculated as

$$G = \max_{1<i\leq P/2} log\left(\frac{Pr\ (S(|Z|_{(P-i+1)}) = i \mid E(Z) = \hat{\mu}_i,\ cov(Z) = \Sigma)}{Pr\ (S(|Z|_{(P-i+1)}) = i \mid E(Z) = 0,\ cov(Z) = \Sigma)}\right) \times I\left(2\overline{\Phi}(|Z|_{(P-i+1)}) < \frac{i}{P}\right)$$

where $|Z|_{(i)}$ denotes the ith order statistic of the absolute value of gene-trait z-scores in an increasing order; $S(t) = \sum_{i=1}^{P} 1(|Z_i| \geq t)$ denotes the number of gene-trait z-scores with absolute value greater than a threshold t; $\hat{\mu}_i$ denotes the corresponding value of $E(Z)$ that maximizes the probability of event $S(|Z|_{(P-i+1)}) = i$; and $\overline{\Phi}(t) = 1 - \Phi(t)$ is the survival function of the standard normal distribution. The GBJ test statistic can be interpreted as the maximum of a series of one-sided likelihood ratio test statistics on the mean of $S(t)$, where the denominator denotes the maximum likelihood when no gene-trait association exists in any tissue (all z-scores have zero mean) and the numerator denotes the unconstrained maximum likelihood. Of note, calculating the exact distribution of $S(t)$ is difficult when z-scores are correlated. As previously suggested, we calculate G by approximating the distribution of $S(t)$ with an extended beta-binomial (EBB) distribution. As a maximum-based global statistic, the p-value of GBJ test could be written as

$$pvalue = 1 - Pr(S(b_i) \leq (d - i), \forall i = 1,2, \dots, P \mid Z \sim MVN(0, \Sigma))$$

where $0 \leq b_1 \leq b_2 \leq \cdots \leq b_P$ are 'boundary points' derived from inversion of the test statistic, which depends on G, P and Σ. The last quantity in the equation can be calculated recursively with the EBB approximation [29].

Simulation settings

Genotype data from 12,547 individuals in the GERA dataset (dbGaP accession: phs000674), including 7,432 type-2 diabetes cases (phenotypic information not used) and 5,205 healthy controls, were used in the simulation studies. We removed SNPs with missing rate above 0.01 and individuals with genetic relatedness coefficients above 0.05. The genotype data were imputed to the 1000 Genomes Project Phase 1v3 European samples using the Michigan Imputation Server [75]. After imputation, we further removed SNPs with MAF < 0.05. After quality control, 5,932,546 SNPs remained in the dataset.

We performed two different simulation studies to evaluate the type-I error rate of our cross-tissue association test. First, we directly simulated quantitative traits from a standard normal distribution independent from the genotype data, and then performed single-tissue association tests for 44 tissues in GTEx and GBJ cross-tissue association test for all genes using the simulated data. In the second setting, we simulated genetically-regulated expression components and then simulated the GWAS trait based on gene expression values. For each gene, we simulated its expression in three tissues, namely skeletal muscle (N=361), skin from sun-exposed lower leg (N=302), and whole blood (N=338). Within the ith tissue, the cis-component of gene expression was generated as $E_i = X_i \hat{B}_{\cdot i}$. We used real effect sizes $\hat{B}_{\cdot i}$ estimated in our joint imputation model so that the genetic architecture of gene expression was preserved in the simulations. Next, the quantitative trait value was simulated as $Y = w_1 E_1 + w_2 E_2 + w_3 E_3 + \varepsilon$, where w_i is the effect of gene expression on the trait in the ith tissue. To evaluate type-I error, we set $w_1 = w_2 = w_3 = 0$, i.e. none of the three tissues are relevant to the trait.

To simulate data under the alternative hypothesis, we fixed the total variance explained by E_1, E_2, and E_3 (i.e. R^2) to be 0.01, and varied w_i to simulate different levels of tissue specificity of the trait. We generated traits using the following three settings:

Setting 1. $w_1 = 1$, $w_2 = w_3 = 0$. Only the first tissue contributes to the disease, the other two tissues are not relevant.

Setting 2. $w_1 = w_2 = \frac{1}{2}$, $w_3 = 0$. Both the first and the second tissue contribute equally to disease, the third tissue is irrelevant to the disease.

Setting 3. $w_1 = w_2 = w_3 = \frac{1}{3}$. All three tissues contribute equally to the disease.

Single-tissue and cross-tissue gene-trait associations were then estimated using the UTMOST framework. We repeated the whole procedure on 200 randomly selected genes. For each gene, we further replicated 5 times. Statistical power is calculated as the proportion of test p-values reaching the significance threshold, i.e. 0.05/15000 for

both single-tissue and cross-tissue tests and 0.05/45000 for single tissue tests while accounting for the number of tissues.

GWAS data analysis

We applied UTMOST to GWAS summary statistics for 50 complex diseases and traits. Details of these 50 studies are summarized in **Supplementary Table 4**. GWAS summary statistics for LDL cholesterol was downloaded from the Global Lipids Genetics Consortium website (http://lipidgenetics.org). Summary statistics from the IGAP stage-I analysis was downloaded from the IGAP website (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php). GWAX result for LOAD was downloaded from http://gwas-browser.nygenome.org/downloads/. ADGC phase 2 summary statistics were generated by first analyzing individual datasets using logistic regression adjusting for age, sex and the first three principal components in the program SNPTest v2 [76]. Meta-analysis of the individual dataset results was then performed using the inverse-variance weighted approach in METAL [77].

In the UTMOST analytical framework, multiple parameters need to be estimated using an external reference panel (e.g. LD). We used samples with European ancestry from the 1000 Genomes Project for this estimation [78]. When performing cross-tissue association tests, we combined single-tissue statistics from tissues that passed FDR < 0.05 criteria to reduce noise in the analysis. Genome-wide significance was defined as 3.3e-6 (i.e. Bonferroni correction based on 15,120 genes that passed the quality control steps). For heritability enrichment analysis, we applied LDSC to 27 GenoSkyline-Plus tissue-specific annotations that have matched tissue types in GTEx (**Supplementary Table 12**). The 53 LDSC baseline annotations were also included in the model as previously recommended [31]. The most and least relevant tissues were selected based on the enrichment test p-values. Gene ontology enrichment analysis was performed using DAVID [79]. Protein-protein interaction information was acquired from AlzData website (http://alzdata.org/index.html) [56]. Locus plots for SNP-level GWAS associations were generated using LocusZoom [80]. Manhattan plots were

generated using the qqman package in R [81].

Additional QTL data

Imputation model for liver tissue in the STARNET study (N=522) was downloaded from (https://github.com/Wainberg/Vulnerabilities_of_TWAS). Predictor effects were trained using an elastic-net model with variants within 500kb range of the transcription-starting site. Details on the quality control procedure has been previously reported [20]. We have also collected additional eQTL and sQTL data for three immune cell types (CD14+ monocytes, CD16+ neutrophils, and naive CD4+ T cells; 169-194 samples per tissue) from the BLUEPRINT consortium (ftp://ftp.ebi.ac.uk/pub/databases/blueprint/blueprint Epivar/qtl as/). eQTLs with FDR < 0.01 and sQTLs with FDR < 0.05 were used in the gene-level association analysis for LOAD.

Software availability

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

Acknowledgements

This study was supported in part by the National Institutes of Health grants R01 GM59507, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, and the Yale World Scholars Program sponsored by the China Scholarship Council. Dr. Lu's effort was supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. Dr. Crane's and Dr. Mukherjee's efforts were supported by grant R01 AG042437 and U01 AG006781. We thank Dr. Christopher Brown for his assistance in matching GTEx tissues to Roadmap cell types. This study makes use of summary statistics from many GWAS consortia. We thank the investigators in these GWAS consortia for generously sharing their data. We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the subjects and their families. The i-Select chips were funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2, and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant n° 503480), Alzheimer's Research UK (Grant n° 503176), the Wellcome Trust (Grant n° 082604/2/07/Z), and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant n° 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC–10–196728. We thank contributors who collected samples used in this study, as well as patients and their families, whose help and participation made this work possible; Data for this study were prepared,

archived, and distributed by the National Institute on Aging Alzheimer's Disease Data Storage Site (NIAGADS) at the University of Pennsylvania (U24-AG041689-01). We are also grateful for all the consortia and investigators that provided publicly accessible GWAS summary statistics.

Competing financial interests

The authors declare no competing financial interests

Author contribution

Y.H., M.L., Q.L., and H.Z. conceived the study and developed the statistical model.
Y.H., M.L., Q.L., H.W., J.W., S.M.Z., B.L., and Y.S. performed the statistical analyses.
S.M.Z. and P.N. assisted in LDL analysis.
Y.H., M.L., Z.Y., and Q.L. implemented the software.
B.K. prepared ADGC summary statistics.
A.N. A.K. and Y.Z. assisted in data preparation
S.M. and P.C. assisted in Alzheimer's disease data application, curation, and interpretation.
Y.H., M.L., Q.L., and H.Z. wrote the manuscript.
H.Z. advised on statistical and genetic issues.
All authors contributed in manuscript editing and approved the manuscript.

References

1. Visscher, P.M., et al., *10 years of GWAS discovery: biology, function, and translation.* The American Journal of Human Genetics, 2017. **101**(1): p. 5-22.
2. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.
3. Ardlie, K.G., et al., *The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-660.
4. Aguet, F., et al., *Genetic effects on gene expression across human tissues.* Nature, 2017. **550**: p. 204-213.
5. Yang, F., et al., *Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis.* Genome Research, 2017.
6. Saha, A., et al., *Co-expression networks reveal the tissue-specific regulation of transcription and splicing.* Genome Research, 2017.
7. Mohammadi, P., et al., *Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change.* Genome Research, 2017.
8. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.
9. Hou, L., et al., *Guilt by rewiring: gene prioritization through network rewiring in genome wide association studies.* Human molecular genetics, 2013. **23**(10): p. 2780-2790.
10. Giambartolomei, C., et al., *Bayesian test for colocalisation between pairs of genetic association studies using summary statistics.* PLoS Genet, 2014. **10**(5): p. e1004383.
11. Zhu, Z., et al., *Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.* Nat Genet, 2016. **48**(5): p. 481-7.
12. Hormozdiari, F., et al., *Colocalization of GWAS and eQTL signals detects target genes.* The American Journal of Human Genetics, 2016. **99**(6): p. 1245-1260.
13. Zhao, S.D., et al., *Sparse simultaneous signal detection for identifying*

genetically controlled disease genes. Journal of the American Statistical Association, 2016(just-accepted).

14. Gamazon, E.R., et al., *A gene-based association method for mapping traits using reference transcriptome data.* Nature genetics, 2015. **47**(9): p. 1091-1098.
15. Gusev, A., et al., *Integrative approaches for large-scale transcriptome-wide association studies.* Nature genetics, 2016. **48**(3): p. 245-252.
16. Mancuso, N., et al., *Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits.* The American Journal of Human Genetics, 2017. **100**(3): p. 473-487.
17. Barbeira, A.N., et al., *Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics.* bioRxiv, 2017: p. 045260.
18. Hoffman, J.D., et al., *Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk.* PLoS genetics, 2017. **13**(3): p. e1006690.
19. Liu, X., et al., *Functional architectures of local and distal regulation of gene expression in multiple human tissues.* The American Journal of Human Genetics, 2017. **100**(4): p. 605-616.
20. Wainberg, M., et al., *Vulnerabilities of transcriptome-wide association studies.* bioRxiv, 2017: p. 206961.
21. Li, C., et al., *Improving genetic risk prediction by leveraging pleiotropy.* Human genetics, 2014. **133**(5): p. 639-650.
22. Maier, R., et al., *Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder.* The American Journal of Human Genetics, 2015. **96**(2): p. 283-294.
23. Hu, Y., et al., *Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction.* PLoS genetics, 2017. **13**(6): p. e1006836.
24. Flutre, T., et al., *A statistical framework for joint eQTL analysis in multiple tissues.* PLoS genetics, 2013. **9**(5): p. e1003486.
25. Sul, J.H., et al., *Effectively identifying eQTLs from multiple tissues by*

combining mixed model and meta-analytic approaches. PLoS genetics, 2013. **9**(6): p. e1003491.

26. Duong, D., et al., *Applying meta-analysis to genotype-tissue expression data from multiple tissues to identify eQTLs and increase the number of eGenes.* Bioinformatics, 2017. **33**(14): p. i67-i74.
27. Li, G., et al., *HT-eQTL: Integrative eQTL Analysis in a Large Number of Human Tissues.* arXiv preprint arXiv:1701.05426, 2017.
28. Hore, V., et al., *Tensor decomposition for multiple-tissue gene expression experiments.* Nature genetics, 2016. **48**(9): p. 1094-1100.
29. Sun, R. and X. Lin, *Set-Based Tests for Genetic Association Using the Generalized Berk-Jones Statistic.* arXiv preprint arXiv:1710.02469, 2017.
30. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506.
31. Finucane, H.K., et al., *Partitioning heritability by functional annotation using genome-wide association summary statistics.* Nature Genetics, 2015.
32. Lu, Q., et al., *Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease.* PLOS Genetics, 2017. **13**(7): p. e1006933.
33. Global Lipids Genetics, C., *Discovery and refinement of loci associated with lipid levels.* Nature genetics, 2013. **45**(11): p. 1274-1283 %@ 1061-4036.
34. Franzén, O., et al., *Cardiometabolic risk loci share downstream cis-and trans-gene regulation across tissues and diseases.* Science, 2016. **353**(6301): p. 827-830.
35. Musunuru, K., et al., *From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus.* Nature, 2010. **466**(7307): p. 714-719 %@ 0028-0836.
36. Strong, A., et al., *Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism.* The Journal of clinical investigation, 2012. **122**(8): p. 2807.
37. Lambert, J.-C., et al., *Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease.* Nature genetics, 2013. **45**(12): p. 1452-1458.
38. Gagliano, S.A., et al., *Genomics implicates adaptive and innate immunity in*

Alzheimer's and Parkinson's diseases. Annals of Clinical and Translational Neurology, 2016. **3**(12): p. 924-933.

39. Huang, K.-l., et al., *A common haplotype lowers PU. 1 expression in myeloid cells and delays onset of Alzheimer's disease.* Nature neuroscience, 2017. **20**: p. 1052.
40. Raj, T., et al., *Integrative analyses of splicing in the aging brain: role in susceptibility to Alzheimer's Disease.* bioRxiv, 2017: p. 174565.
41. Liu, J.Z., Y. Erlich, and J.K. Pickrell, *Case-control association mapping by proxy using family history of disease.* Nature genetics, 2017. **49**(3): p. 325-331.
42. Hollingworth, P., et al., *Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease.* Nature genetics, 2011. **43**(5): p. 429-435.
43. Harold, D., et al., *Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease.* Nature genetics, 2009. **41**(10): p. 1088-1093.
44. Naj, A.C., et al., *Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease.* Nature genetics, 2011. **43**(5): p. 436-441.
45. Seshadri, S., et al., *Genome-wide analysis of genetic loci associated with Alzheimer disease.* Jama, 2010. **303**(18): p. 1832-1840.
46. Jun, G.R., et al., *Transethnic genome-wide scan identifies novel Alzheimer's disease loci.* Alzheimer's & Dementia, 2017.
47. Lambert, J.C., et al., *Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease.* Nat Genet, 2009. **41**(10): p. 1094-9.
48. Sherva, R., et al., *Genome-wide association study of the rate of cognitive decline in Alzheimer's disease.* Alzheimer's & Dementia, 2014. **10**(1): p. 45-52.
49. Crehan, H., et al., *Complement receptor 1 (CR1) and Alzheimer's disease.* Immunobiology, 2012. **217**(2): p. 244-250.
50. Liu, J.Z., et al., *Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across*

populations. Nature genetics, 2015. **47**(9): p. 979-986.

51. Remmers, E.F., et al., *Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behcet's disease.* Nature genetics, 2010. **42**(8): p. 698-702.
52. Plagnol, V., et al., *Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases.* PLoS genetics, 2011. **7**(8): p. e1002216.
53. Bentham, J., et al., *Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus.* Nature genetics, 2015.
54. Kiyota, T., et al., *AAV serotype 2/1-mediated gene delivery of anti-inflammatory interleukin-10 enhances neurogenesis and cognitive function in APP+ PS1 mice.* Gene therapy, 2012. **19**(7): p. 724-733.
55. Chakrabarty, P., et al., *IL-10 alters immunoproteostasis in APP mice, increasing plaque burden and worsening cognitive behavior.* Neuron, 2015. **85**(3): p. 519-533.
56. Xu, M., et al., *A systematic integrated analysis of brain expression profiles reveals YAP1 and other prioritized hub genes as important upstream regulators in Alzheimer's disease.* Alzheimer's & Dementia, 2017.
57. Hohman, T.J., et al., *Discovery of gene-gene interactions across multiple independent data sets of late onset Alzheimer disease from the Alzheimer Disease Genetics Consortium.* Neurobiology of aging, 2016. **38**: p. 141-150.
58. Katsouri, L., et al., *Prazosin, an α 1-adrenoceptor antagonist, prevents memory deterioration in the APP23 transgenic mouse model of Alzheimer's disease.* Neurobiology of aging, 2013. **34**(4): p. 1105-1115.
59. Duplan, L., et al., *Lithostathine and pancreatitis-associated protein are involved in the very early stages of Alzheimer's disease.* Neurobiology of aging, 2001. **22**(1): p. 79-88.
60. Stenmark, H. and V.M. Olkkonen, *The rab gtpase family.* Genome biology, 2001. **2**(5): p. reviews3007. 1.
61. Lin, B.D., et al., *Heritability and GWAS Studies for Monocyte–Lymphocyte Ratio.* Twin Research and Human Genetics, 2017. **20**(2): p. 97-107.
62. Astle, W.J., et al., *The allelic landscape of human blood cell trait variation and links to common complex disease.* Cell, 2016. **167**(5): p. 1415-1429.

e19.

63. Li, T., et al., *Identification of the gene for vitamin K epoxide reductase.* Nature, 2004. **427**(6974): p. 541-544.

64. Kohnke, H., et al., *Warfarin dose related to apolipoprotein E (APOE) genotype.* European journal of clinical pharmacology, 2005. **61**(5-6): p. 381-388.

65. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-713.

66. de Lange, K.M., et al., *Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease.* Nature genetics, 2017. **49**(2): p. 256-261.

67. Davies, G., et al., *Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N= 53 949).* Molecular psychiatry, 2015. **20**(2): p. 183.

68. Park, Y., et al., *Causal gene inference by multivariate mediation analysis in Alzheimer's disease.* bioRxiv, 2017: p. 219428.

69. Mancuso, N., et al., *Probabilistic fine-mapping of transcriptome-wide association studies.* bioRxiv, 2017.

70. Xu, Z., et al., *A Powerful Framework for Integrating eQTL and GWAS Summary Data.* Genetics, 2017.

71. Yuan, M. and Y. Lin, *Model selection and estimation in regression with grouped variables.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006. **68**(1): p. 49-67.

72. Li, Y., B. Nan, and J. Zhu, *Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure.* Biometrics, 2015. **71**(2): p. 354-363 %@ 1541-0420.

73. Stegle, O., et al., *A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies.* PLoS computational biology, 2010. **6**(5): p. e1000770.

74. O'Connor, L.J., et al., *Estimating the proportion of disease heritability mediated by gene expression levels.* bioRxiv, 2017: p. 118018.

75. Das, S., et al., *Next-generation genotype imputation service and methods.*

Nature genetics, 2016. **48**(10): p. 1284.

76. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nature genetics, 2007. **39**(7): p. 906-913 %@ 1061-4036.

77. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-2191 %@ 1460-2059.

78. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

79. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature protocols, 2009. **4**(1): p. 44-57.

80. Pruim, R.J., et al., *LocusZoom: regional visualization of genome-wide association scan results.* Bioinformatics, 2010. **26**(18): p. 2336-2337.

81. Turner, S.D., *qqman: an R package for visualizing GWAS results using QQ and manhattan plots.* BioRxiv, 2014: p. 005165.