Partitioning and aggregating cross-tissue and tissue-specific genetic effects in identifying gene-trait associations

Jun Liu
jliu@stat.harvard.edu

Harvard University https://orcid.org/0000-0002-4450-7239

Shuang Song
Tsinghua University https://orcid.org/0000-0002-1903-090X

Lijun Wang
Yale University https://orcid.org/0000-0001-5222-4506

Lin Hou
Tsinghua University https://orcid.org/0000-0002-4283-8501

Article

Keywords:

Posted Date: December 12th, 2023

DOI: https://doi.org/10.21203/rs.rs-3674156/v1

License: ☝️ This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is NO Competing Interest.

Version of Record: A version of this preprint was published at Nature Communications on July 9th, 2024. See the published version at https://doi.org/10.1038/s41467-024-49924-4.
Partitioning and aggregating cross-tissue and tissue-specific genetic effects in identifying gene-trait associations

Shuang Song\textsuperscript{1}, Lijun Wang\textsuperscript{2}, Lin Hou\textsuperscript{1,3,∗}, and Jun S. Liu\textsuperscript{4,∗}

\textsuperscript{1}Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing 100084, China
\textsuperscript{2}Department of Biostatistics, Yale School of Public Health, New Haven, CT 06510, USA
\textsuperscript{3}MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China
\textsuperscript{4}Department of Statistics, Harvard University, Cambridge, MA 02138, USA

∗To whom correspondence should be addressed. e-mail: houl@tsinghua.edu.cn, jliu@stat.harvard.edu

ABSTRACT

Transcriptome-wide association studies (TWAS) have shown great promises in extending GWAS loci to a functional understanding of disease mechanisms. In an effort to fully unleash the TWAS and GWAS information, we propose MTWAS, a statistical framework that partitions and aggregates cross-tissue and tissue-specific genetic effects in identifying gene-trait associations. Different from previous methods, we introduce a non-parametric imputation strategy to augment the inaccessible tissues, which allows for barren conditions, such as complex interactions and non-linear expression data structure across tissues. We further classify eQTLs into cross-tissue eQTLs (ct-eQTLs) and tissue-specific eQTLs (ts-eQTLs) via a step-wise procedure based on the extended Bayesian information criterion, which was consistent under high-dimensional settings. We have shown that MTWAS significantly improves the imputation accuracy across all 47 GTEx tissues compared with other single-tissue and multi-tissue methods, such as PrediXcan and UTMOST. MTWAS also identifies more predictable genes that can be replicated with independent studies. Applications to 84 UKBB GWAS studies have provided novel insights into disease etiology. The R package implementing MTWAS is available at https://github.com/szcf-weiya/MTWAS.

Introduction

Genome-wide association studies (GWAS) have identified tens of thousands of susceptibility loci for complex diseases, bringing novel insights into disease etiologies \cite{1}. However, it remains a challenge to understand and interpret the GWAS findings \cite{2, 3}, especially those significant hits in non-coding regions. One hypothesis is that some non-coding variants influence traits through the regulation of gene expression \cite{4}. To identify the path from genomic variants to gene expression is challenging. A simple strategy is to assign the associated variant a causal link to its nearest gene, but a shorter physical distance does not necessarily indicate a closer functional connection. A counter-example is that the obesity-associated single nucleotide polymorphisms (SNPs) within $FTO$ form long-range...
functional connections with \textit{IRX3} [2, 5]. In recent years, some large-scale consortium have provided rich resources to identify expression quantitative trait loci (eQTLs) across various human tissues. For example, version 8 of the Genotype-Tissue Expression (GTEx v8) project collects a comprehensive set of 54 tissues from hundreds of donors [6]. Integrating eQTLs with GWAS studies reveals that a large proportion of phenotype variability in disease risk can be explained by variants that regulate the expression levels of genes [7].

Recently, transcriptome-wide association studies (TWAS) have provided a successful path to bridge SNPs, gene expressions, and complex phenotypes. TWAS includes two stages: the first stage builds a linear model to predict genetically regulated gene expression, and the second stage associates complex diseases with the predicted gene expression. For simplicity, we refer to the two stages as the "prediction stage" and "association stage" throughout the paper. The widely used TWAS methods include PrediXcan [2], which trains an elastic net in the prediction stage, and FUSION [8], which includes more models such as top eQTL, LASSO, and Bayesian sparse linear model. The gene-trait associations identified by TWAS shed light on the genetic basis of complex diseases. However, the limited sample sizes of eQTL studies have become a bottleneck in TWAS, resulting in low power in both the prediction and association stages. In the GTEx v8 dataset, for example, the sample sizes for 21 out of the 54 tissues with genotypic information are less than 200.

Considering the tissue-dependent nature of transcription regulation and the shared identified eQTL across tissues, jointly modeling multiple tissues has the potential to improve the performance of TWAS. There have been some methods for improving statistical power in the association stage, including MultiXcan, which tests the joint effects of gene expression variation from different tissues [9]; and the sparse-canonical-correlation-analysis-TWAS (sCCA-TWAS), which combines the imputed expression of multiple tissues in the TWAS [10]. As for the prediction stage, the recently developed UTMOST method formulates cross-tissue expression imputation as a penalized multivariate regression problem, and introduces a group-lasso penalty on the cross-tissue effects and encourages the presence of eQTLs shared across tissues. Despite the improved imputation accuracy compared with single-tissue imputation, we note that the penalty strengthens the estimation of the effects of eQTLs shared across all tissues, but does not distinguish between biologically related or irrelevant tissues. In other words, the method is likely to identify the shared eQTLs across all tissues. However, the genotype effects can show stronger correlated effects in either subsets of tissues (e.g., brain-related tissues), or only one tissue (e.g., testis) [11]. Furthermore, the tissue-specific eQTLs can sometimes provide a more targeted mechanistic interpretation for GWAS associations, compared to the eQTLs shared across all tissues [12, 13].

In this manuscript, we develop a flexible TWAS framework aggregating multiple tissues, namely MTWAS, which partitions and aggregates both cross-tissue and tissue-specific genetic effects. Based on the intrinsic property of the genetic data where the transcriptome of one tissue can be highly correlated with the others, we leverage a non-parametric missing value imputation method for the
inaccessible tissues, and select eQTLs via a stepwise extended Bayesian information criteria (EBIC) procedure, which was shown to be consistent under high-dimensional settings [14, 15]. Compared with UTMOST, we consider not only the shared eQTLs across all tissues, but also eQTLs that regulate genes in specific tissues. In addition, we allow barren conditions including complex interactions and non-linear data structures. We derive an explicit form of the MTWAS statistics with imputation weights, GWAS summary statistics, and an LD reference panel.

Results

Model Overview

We divided the prediction stage of MTWAS into three steps (Figure 1). In the first step, instead of directly using the gene expression matrix with unmeasured entries, as implemented by previous methods, we introduce a non-parametric imputation step, where we predict the inaccessible entries of the gene expression matrix with other relevant tissues, using a machine learning missing value imputation method [16]. Then, we impute the missing entries of the expression matrix with the predicted values, and the predicted expression matrix is subsequently used as the response variable in the following step. We note that this imputation step of the expression matrices do not involve the information of genotype data.

The second step is to detect the cross-tissue eQTLs (ct-eQTLs). Intuitively, we select the eQTLs that can explain the variance of multiple tissues. In practice, we project the imputed expression matrix onto its principal components (PCs), which are further used for variable selection. The number of PCs is decided by the eigenvalues of the correlation matrix. We use 2 as a default cutoff for the eigenvalues, which results in 5 PCs on average. We also show that the performance of our method is robust to the number of PCs used for ct-eQTLs identification (Figure S1). We fit a linear regression model and select ct-eQTLs via EBIC (Methods and Materials).

At the third step, we fix the ct-eQTLs and perform tissue-wise analysis to identify the tissue-specific eQTLs (ts-eQTLs) by minimizing EBIC. The effects of ct-eQTLs and ts-eQTLs are further estimated with a computationally efficient stepwise linear regression method (Methods and Materials).

We further utilize the effects of both ct-eQTLs and ts-eQTLs trained in the prediction stage to infer the gene-trait associations. For each phenotype, we perform an association test in a tissue-specific manner to retain the tissue specificity of the trait-gene associations. We derive an explicit form of the MTWAS statistics with imputation weights, GWAS summary statistics, and LD reference panel (Methods and Materials). We demonstrate that MTWAS achieves the highest imputation accuracy compared with the state-of-the-art imputation methods, including PrediXcan and UTMOST, with both cross-validation (CV) and independent replication studies.
Figure 1. **Schematic diagram of the prediction stage of MTWAS.** The prediction stage of MTWAS are divided into three steps. The Step 1 is a imputation step, where the missing entries of the observed expression matrix were imputed with other relevant tissues with MissForest algorithm. The Step 2 is to identify the cross-tissue eQTLs, and the Step 3 is to identify the tissue-specific eQTLs, as well as effect sizes estimation.
Imputation accuracy

GTEx study

We evaluate the imputation accuracy of our method with the GTEx v8 dataset. After quality control, 47 tissues were retained with sample sizes larger than 100 (Methods and Materials). We avoided over-fitting issues by imputing the expression matrix with training samples in each fold separately. Compared with the state-of-the-art TWAS methods PrediXcan and UTMOST, MTWAS substantially improves the prediction accuracy across all tissues. Compared with PrediXcan, MTWAS achieved an average improvement of prediction $R^2$ by 0.02 across 47 tissues and an improved percentage of 47.3% (SD=17.4%) (Figure 2). Compared with UTMOST, MTWAS achieved significant improvement with a proportion of 9.1% (SD=2.2%) on average. Since PrediXcan performed worse than UTMOST and MTWAS among all methods in consideration, Figure 2 plots the improvements of the prediction $R^2$ over that of PrediXcan for all methods (0 for no improvement). We note that the improvement of our method is more significant in tissues with smaller sample sizes. For example, MTWAS improves the prediction $R^2$ by an average of 60.7% over that of PrediXcan for tissues with sample sizes less than 200; and 42.2% for tissues with sample sizes between 200 and 400, and 27.0% for tissues with sample sizes larger than 400 (blue dots in Figure 2). UTMOST (grey dots) also outperformed PrediXcan significantly, but was inferior to MTWAS uniformly.

The improvement in prediction accuracy can be attributed to three aspects: i) The imputation step for imputing expression matrices utilizes the cross-tissue information and increases the effective sample sizes of the training dataset; ii) While the cross-tissue effects aggregate information from various tissues, tissue-specific effects retain their inherent tissue-specific characteristics, and both contribute to the prediction; iii) Although computationally more demanding, EBIC for linear regression models under high-dimensional settings appears to be superior to other variable selection criterion such as those used in Lasso and elastic net.

We performed more experiments to further demonstrate each of the aforementioned points. For point (i), we first performed PrediXcan on the imputed expression matrix (denoted as PrediXcan-imp), and compared the prediction accuracy to the original PrediXcan. The PrediXcan-imp achieved higher prediction $R^2$ than PrediXcan on all 47 tissues, ranging from 5.32% on muscle skeletal ($N = 607$) to 32.7% on uterus ($N = 107$) (red dots in Figure 2). This reflects the benefits of the imputation step. Notably, this step does not involve genotype data, thus is very flexible when RNA-seq data with larger sample sizes are available without privacy concerns.

For point (ii), instead of partitioning the ct-eQTLs and ts-eQTLs, we directly performed variable selection with all cis-eQTLs on each tissue using EBIC, which we denoted as MTWAS-tissue. As shown in Figure 2, MTWAS (blue dots) achieved higher prediction $R^2$ compared with MTWAS-tissue (yellow dots) for all 47 tissues, suggesting that our partition and aggregation of ct-eQTLs and ts-eQTLs indeed helped. For point (iii), we compared the performances of MTWAS-tissue (corresponding to using EBIC) and PrediXcan-imp EBIC (corresponding to using elastic net), and observed that the
former outperformed the latter across all tissues (yellow versus red dots).

Figure 2. The improvement of prediction $R^2$ over PrediXcan, of MTWAS, MTWAS-tissue, UTMOST, and PrediXcan-imp, evaluated on 47 tissues of GTEx datasets. The tissues are arranged in descending order based on their respective sample sizes, from highest to lowest. The prediction $R^2$ is based on 5-fold CV.
The pattern of the improvement in the number of identified predictable genes is similar to that for the prediction $R^2$ (Table 1). We consider both the common threshold (Prediction $R^2 < 0.01$) and a more stringent threshold for the identification of a predictable gene (Materials and Methods). In general, the numbers of predictable genes identified by MTWAS range from 10,444 (for whole blood) to 17,076 (for uterus) under the common threshold; and range from 1,247 (for brain amygdala) to 5,988 (for nerve tibial) under the stringent threshold. MTWAS identified significantly more imputable genes than both PrediXcan and UTMOST across all 47 tissues, with an overall improvement of 41.7% (SD=10.0%) and 2.7% (SD=1.2%) over PrediXcan and UTMOST, respectively, under the common threshold (Figures 3 and S2); and 56.1% (SD=24.2%) and 15.7% (SD=3.1%) over PrediXcan and UTMOST, respectively, under the stringent threshold.

![Figure 3](image-url)

**Figure 3.** A comparison between the prediction accuracy of MTWAS with that of PrediXcan on the GTEx datasets. 

- **a.** Prediction $R^2$ of gene expressions of the muscle skeletal, which has the largest sample size among the GTEx datasets. 
- **b.** Prediction $R^2$ of gene expressions of the uterus, which has the smallest sample size. The x-axis and y-axis represent the prediction $R^2$ derived from a 5-fold CV. Purple and green dots represent genes that are imputable only for MTWAS and for PrediXcan, respectively. Darker and shallower grey dots represent genes that are consistently imputable and not imputable, respectively, using both methods.

**Table 1.** The average Prediction $R^2$ and the number of predictable genes across 47 tissues with MTWAS, PrediXcan, and UTMOST. The highest Prediction $R^2$ and the largest number of predictable genes across three methods are highlighted in boldface.
Replication on the GEUVADIS cohort

In addition to the GTEx analysis, we also performed an independent replication study by applying the weights trained with GTEx samples for Epstein-Barr virus (EBV) transformed lymphocytes to the expression levels of 344 European individuals from GEUVADIS lymphoblastoid cell lines (LCLs) [17]. The improvement in the independent replication studies was consistent with the performance in GTEx studies (Table 2). Compared with PrediXcan and UTMOST, the overall prediction $R^2$ was improved by 43.5% and 3.1%, respectively. MTWAS achieved significantly higher prediction accuracy than both methods in difference quantiles ($p$-value $< 2.2 \times 10^{-16}$; Kolmogorov-Smirnov test, one-sided). MTWAS also identified 5,132 genes with prediction $R^2 > 0.01$, which was 2,291 more than the predictable genes identified by PrediXcan and 238 more than that identified by UTMOST.

Table 2. Replication study on the GEUVADIS cohort for lymphoblastoid cell lines. The training weights are based on the Epstein-Barr virus transformed lymphocytes in the GTEx datasets. The highest Prediction $R^2$ and the largest number of predictable genes across three methods are highlighted in boldface.

<table>
<thead>
<tr>
<th></th>
<th>MTWAS</th>
<th>PrediXcan</th>
<th>UTMOST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average prediction $R^2$</td>
<td>0.033</td>
<td>0.023</td>
<td>0.032</td>
</tr>
<tr>
<td># predictable genes (common)</td>
<td>5,132</td>
<td>2,841</td>
<td>4,894</td>
</tr>
<tr>
<td># predictable genes (stringent)</td>
<td>646</td>
<td>546</td>
<td>638</td>
</tr>
</tbody>
</table>

Investigation of cross-tissue and tissue-specific effects

MTWAS partitioned the eQTLs into ct-eQTLs and ts-eQTLs, and aggregated the effects of both types of eQTLs in prediction. We observed that both ct-eQTLs and ts-eQTLs are significantly enriched in promoter regions, while ct-eQTLs ($\times 5.10$) are more enriched than ts-eQTLs ($\times 3.36$) (Figure 4a). A potential explanation for this difference in enrichment is that genes whose regulation is mostly tissue specific are under higher selection pressures [18]. Both are more enriched in promoter regions than PrediXcan ($\times 1.56$) and UTMOST ($\times 2.01$). In addition, the CADD scores of ct-eQTLs are also significantly higher than ts-eQTLs ($p = 9.4 \times 10^{-11}$, Figure 4b). Higher CADD scores indicate severer deleteriousness of the eQTLs shared across multiple tissues. The CADD scores of both ct-eQTLs and ts-eQTLs are higher than PrediXcan and UTMOST eQTLs.

We further consider two subsets of the predictable genes, according to whether the genes are regulated only by ct-eQTLs or ts-eQTLs, which we refer to as cross-tissue genes (ct-genes) and tissue-specific genes (ts-genes), correspondingly. Different characteristics were observed across the two sets. KEGG enrichment analysis [19] demonstrated that ct-genes were enriched for lysosome ($p = 5.6 \times 10^{-4}$), peroxisome ($p = 5.6 \times 10^{-4}$), phagosome ($p = 8.8 \times 10^{-4}$), and metabolism pathways that are abundant in all tissues and cell types (Figures S3-S4). In contrast, ts-genes were enriched in pathways that are critical to differentiate between cell types, as also observed in a recent study [18].
Figure 4. Characterization of the eQTLs identified by MTWAS, PrediXcan, and UTMOST. The eQTLs identified by MTWAS are partitioned into cross-tissue eQTLs and tissue-specific eQTLs. a, Enrichment in promoters of the eQTLs identified by TWAS methods. b, CADD scores of the identified eQTLs.

For example, the ts-genes in the brain cerebellar were enriched in the neuroactive ligand-receptor interaction pathway \( (p = 5.4 \times 10^{-5}) \) and the taste transduction pathway \( (p = 1.7 \times 10^{-4}) \). The ts-genes of esophagus mucosa were enriched in the gastric acid secretion pathway \( (p = 2.1 \times 10^{-4}) \). In addition, the esophagus mucosa ts-genes were also enriched in the cAMP signaling pathway \( (p = 3.8 \times 10^{-5}) \). It has been shown that esophageal mucosa expresses predominantly EP2 receptors and esophageal ulceration increases the expression of the EP2 receptor, activation of CREB, which is the downstream target of the cAMP signaling. The ts-genes in the pancreas were also enriched in multiple pathways, such as circadian entrainment \( (p = 4.8 \times 10^{-5}) \), protein digestion and absorption \( (p = 6.8 \times 10^{-5}) \), dopaminergic synapse \( (p = 2.7 \times 10^{-4}) \), insulin secretion \( (p = 2.9 \times 10^{-4}) \), morphine addiction \( (p = 3.7 \times 10^{-4}) \), and pancreatic secretion \( (p = 6.3 \times 10^{-4}) \), etc.

Furthermore, we found that the ts-genes were more intolerant to protein-loss-of-function compared with the ct-genes, as evaluated by LOEUF [20] and pLI [21]. Specifically, the ts-genes had significantly lower LOEUF \( (p = 3.2 \times 10^{-11}) \) and higher pLI \( (p = 4.1 \times 10^{-16}) \) than the ct-genes, indicating that the ts-genes are more variation intolerant compared with ct-genes. This is consistent with a previous finding that ts-genes have stronger selective pressures compared with ct-genes [18]. This has important implications for understanding the genetic basis of disease, as mutations in tissue-specific genes may be more likely to lead to specific diseases affecting particular tissues or organs.
We applied MTWAS, UTMOST, and PrediXcan to 84 UKBB self-reported cancer and non-cancer illness phenotypes with effective sample sizes larger than 5,000. A detailed summary of the phenotypes is provided in Table S1. We applied the Bonferroni correction to account for multiple testing with a $p$ value cutoff $6.06 \times 10^{-8} = 0.05/(47 \text{ tissues} \times 17,566 \text{ genes})$. Across the 84 traits, MTWAS, UTMOST, and PrediXcan identified significant genes in 44, 34, and 39 phenotypes, respectively. Specifically, for 8 phenotypes including heart valve problems, gastric/stomach ulcers, irritable bowel syndrome, peritonitis, muscle/soft tissue problems, iron deficiency anemia, other renal/kidney problems, and chronic fatigue syndrome, MTWAS identified genes whose expression are significantly associated, while UTMOST and PrediXcan did not find any associated genes. In addition, MTWAS identified a greater number of genes than both UTMOST and PrediXcan across 20 different phenotypes, whereas UTMOST and PrediXcan identified more genes than MTWAS in 2 and 19 phenotypes, respectively. Regarding the significant association between gene-tissue pairs and the phenotypes, UTMOST identified the highest number of gene-tissue pairs in 25 phenotypes. MTWAS followed closely, identifying the majority of gene-tissue pairs in 15 phenotypes. In contrast, PrediXcan only managed to identify the most gene-tissue pairs in 4 phenotypes.

Here we elaborate on some novel genes identified only by MTWAS. MTWAS identified that the expression of \textit{ANKRD55} ($p = 4.69 \times 10^{-8}$) on colon transverse is significantly associated with rheumatoid arthritis (RA). \textit{ANKRD55} encodes a protein called ankyrin repeat domain 55, which is well-established to be a potential risk factor for autoimmune diseases [22]. MTWAS found that the expression of \textit{RNF186} in the ovary is significantly associated with ulcerative colitis (UC) ($p = 7.83 \times 10^{-9}$). A recent study has revealed that a protein-truncating R179X variant in \textit{RNF186} has a protective effect against UC [23]. MTWAS also identified some novel genes on malignant melanoma, such as \textit{NOX4} in the thyroid ($p = 8.37 \times 10^{-12}$). Emerging evidence suggests a potential association between \textit{NOX4} gene expression in thyroid tissues and malignant melanoma. As a regulator of glycolysis within thyroid cells, \textit{NOX4} facilitates the proliferation of cancerous thyroid cells through the generation of mitochondrial reactive oxygen species [24]. In addition, in papillary thyroid carcinomas, the down-regulation of the sodium/iode symporter induced by the BRAFV600E mutation is mediated by \textit{NOX4} [25]. Remarkably, \textit{NOX4} inhibitors have shown promises as adjuncts to current therapies for melanoma patients with BRAF mutations [26]. Clinical observations have also highlighted a link between melanoma and thyroid disease, with reports of increased thyroid dysfunction among individuals with a history of melanoma [27]. These findings highlight the comparative advantages of MTWAS in gene and gene-tissue pair identification across multiple phenotypes, emphasizing its potential as a valuable tool in genetic research and analysis.
Despite that TWAS has supplemented GWAS loci with valuable functional understandings of disease mechanisms, the performance of gene expression prediction remains moderate. TWAS mainly includes two key steps: gene expression imputation and the association of gene expression with disease traits. Previous multi-tissue TWAS methods primarily focused on combining single-tissue statistics during the association step, without effectively modeling shared information during the imputation step. Recently, UTMOST was proposed to aggregate the shared information across multiple tissues in the imputation step by encouraging the presence of eQTLs across all tissues. However, this approach may not be optimal as shared information might be distributed across subsets of tissues or be tissue-specific, rendering it inappropriate to encourage cross-tissue effects for all genes. Moreover, these existing methods lack the capability to account for non-linear structures across tissues.

To address these limitations, we propose a novel statistical framework, MTWAS, which strategically partitions and aggregates cross-tissue and tissue-specific genetic effects in identifying gene-trait associations. We introduce a non-parametric imputation step in MTWAS and employ the EBIC criterion to select ct-eQTLs and ts-eQTLs. Our method significantly enhances the effective sample sizes, even under challenging scenarios with limited or sparse expression data. We have demonstrated that the proposed method outperforms existing methods in achieving a higher imputation accuracy of gene expression across all tissues, while also exhibiting superior performance in identifying gene-trait associations. The classification of ct-eQTLs and ts-eQTLs not only improves our ability to capture complex regulatory mechanisms underlying gene expression but also provides novel insights into the intricate interplay between genetic factors and disease traits.

It is crucial to acknowledge that, like other TWAS methods, the gene-trait associations identified by MTWAS do not necessarily indicate causality. Instead, they serve as valuable pointers toward potential functional links between genes and diseases, warranting further investigations through experimental validations and functional studies.

In conclusion, our proposed MTWAS framework represents a substantial step forward in TWAS methodology, offering both improved gene expression imputation and enhanced identification of gene-trait associations. We believe that MTWAS has the potential to become a valuable tool in deciphering the genetic basis of complex diseases and facilitating personalized treatment strategies.

**Materials and methods**

**Imputation of gene expression data**

The first step of our model is to impute expression data without genotype information. Consider an expression matrix containing $K$ tissues and $N$ samples. The collected samples vary across different tissues, resulting in missing values in the matrix. We use a non-parametric missing value imputation method "missForest" [16] for the imputation of inaccessible entries of the expression
matrix. Specifically, we first sorted the tissues according to the ascending number of missing samples, and used an initial imputation based on the mean of other available samples. Then, for each tissue, we trained a random forest model to predict the missing values. For example, for the \( k \)-th tissue, we trained the model on samples that have the non-missing entries of the \( k \)-th column, with the \( k \)-th column as the response and other columns as predictors. The missing entries of the \( k \)-th column can be predicted with other columns with the trained model. We iteratively trained the model until convergence. The final imputed matrix is used for the downstream TWAS analysis.

**Identification of cross-tissue and tissue-specific eQTLs**

We denote the final imputed expression matrix resulting from the aforementioned iterative imputation procedure as \( \tilde{E} \), which will be used as response variables in the following steps. For the identification of the ct-eQTLs, we first perform the principal component analysis (PCA) on \( \tilde{E} \). Then we use a stepwise strategy for variable selection. For each of the first \( Q \) principal components (PCs), we treat the PC as the response variable and fit a linear regression model between the PC and the genotypes of cis-SNPs. The predictors selected by EBIC are then fixed in analyzing the next PC. Figure S1(a) shows the number of ct-eQTLs identified in GTEx cohorts with increasing numbers of PCs \( (Q) \). To decide \( Q \), we evaluated the performance of the identified ct-eQTLs on predicting the gene expression on the whole blood tissue of GTEx data with a 5-fold CV. In general, both the prediction accuracy and the number of predictable genes increased with the increasing \( Q \) (Figure S1(b) and (c)), while the performance of prediction became robust when \( Q \geq 4 \). Therefore, we set \( Q \) to be 5 in the identification of ct-eQTLs as a balance of prediction accuracy and computational efficiency.

We denote the set of identified ct-eQTLs as \( S_0 \). If no ct-eQTLs are selected, we have \( S_0 = \emptyset \). For each tissue, we use its own expression as the response variable, and identify the ts-eQTLs with the EBIC criterion. The ct-eQTLs are fixed in this procedure. In other words, we find a ts-eQTLs set \( S \) that maximize

\[
EBIC_\gamma(S) = -2l_n(\hat{\beta}_S) + |S| \log N + 2\gamma |S| \log M, \quad S \supseteq S_0, \tag{1}
\]

where \(|S|\) is the size of set \( S \), \( M \) is the total number of cis-SNPs, and \( \gamma \) is the tuning parameter. We utilize a stepwise method for efficient computation, which follows the SODA procedure proposed in Li and Liu (2019) without considering interaction terms [15]. We adapt the algorithm by allowing customized fixed main effect terms and perform an exhaustive combination search to find the set achieving the smallest EBIC. The procedure is summarized in Algorithm 1.

**Model training and evaluation**

We trained MTWAS on 47 tissues with sample sizes larger than 100 in GTEx, using 5-fold CV for performance evaluation. Specifically, we first performed imputation on four randomly partitioned subsets of the data, which were used as the training data. Then, we identified ct-eQTLs and ts-eQTLs with Algorithm 1. Leveraging the selected eQTLs, we built linear models to predict gene expression.
Algorithm 1: The algorithm for identifying ct-eQTLs and ts-eQTLs with EBIC criterion.

**Require:** Fixed term set $S_0$, which can be $\emptyset$.

**Require:** Minimum number of non-fixed terms $n_0$.

1: Forward procedure for selecting main effects. Let $\mathcal{M}_t$ denote the selected set of main effects at step $t$. Start with $\mathcal{M}_1 = S_0$.

2: while not terminated do

3: for each $j \not\in \mathcal{M}_t$ do

4: create a new candidate set $\mathcal{M}_{t,j} = \mathcal{M}_t \cup \{j\}$, and evaluate its EBIC

5: end for

6: select predictor $j^*$ with the lowest EBIC: $j^* = \arg\min_j\text{EBIC}(\mathcal{M}_{t,j})$

7: if $\text{EBIC}(\mathcal{M}_{t,j^*}) < \text{EBIC}(\mathcal{M}_t)$ then

8: continue with $\mathcal{M}_{t+1} = \mathcal{M}_{t,j^*}$.

9: else

10: terminate and obtain set $\tilde{\mathcal{M}} = \mathcal{M}_t$.

11: end if

12: end while

13: Backward procedure for eliminating unimportant terms. Let $\mathcal{S}_t$ denote the selected set at step $t$ of the backward stage. Start with $\mathcal{S}_1 = \tilde{\mathcal{M}}$.

14: while not terminated do

15: for each $j \in \mathcal{S}_t \setminus S_0$ do

16: create a candidate set $\mathcal{S}_{t,j} = \mathcal{S}_t \setminus \{j\}$, and evaluate its EBIC

17: end for

18: find term $j$ with lowest EBIC: $j^* = \arg\min_j\text{EBIC}(\mathcal{S}_{t,j})$

19: if $\text{EBIC}(\mathcal{S}_{t,j^*}) < \text{EBIC}(\mathcal{S}_t)$ and $|\mathcal{S}_t \setminus S_0| > n_0$ then

20: remove term $j^*$

21: else

22: terminate and retain set $\tilde{\mathcal{S}} = \mathcal{S}_t$.

23: end if

24: end while

25: Enumerate all possible combinations of non-fixed terms and find the subset that reaches the smallest EBIC

$$\hat{\mathcal{A}} = \arg\min_{\mathcal{A} \subset \tilde{\mathcal{S}} \setminus S_0} \text{EBIC}(\mathcal{A} \cup S_0)$$
levels and tested the models with samples in the remaining fold with accessible expression data of the tissue. The performance of the imputation model was measured by its prediction $R^2$.

In addition to the internal CV in GTEx datasets, we conducted an external validation study using the GEUVADIS dataset, which comprises lymphoblastoid cell lines of 344 individuals with European ancestry. We trained the model on Epstein-Barr virus (EBV) cell lines of GTEx dataset. Initially, the transcriptomic information of 116 samples was available. We performed imputations on 715 GTEx samples, which were then utilized for predictive analysis in the GEUVADIS dataset.

Gene-trait association studies
We compute the gene-level summary statistics based on the training weights for the selected eQTLs. Let the sample size be $N$ and the number of genes be $G$. For each gene on each tissue, we assume a linear model between the complex phenotype ($Y$) and the gene expression ($E$):

$$Y = E\gamma + \eta,$$

where $Y$ is the $N \times 1$ phenotype vector, $E$ is an $N \times 1$ expression vector, $\gamma$ is the gene-level effect size. We assume both $Y$ and $E$ have been standardized, and the error $\eta$ follows the normal distribution with mean 0. The MTWAS $Z$-score vector is:

$$Z = \frac{\hat{\gamma}}{se(\hat{\gamma})}.$$ \hspace{1cm} (3)

If the individual-level genotype data are accessible, we could directly compute the expression $E$ with $\sum_{j \in A} \hat{\beta}_j X_j$, where $X_j$ is the genotype of the $j$-th SNP; $\hat{\beta}_j$ is the estimated weights on gene expression of the $j$-th SNP; and $A$ is the selected eQTLs for the gene. If the individual-level genotype data are not available, we could derive MTWAS test statistics with GWAS summary statistics. Specifically, the MTWAS $Z$-score can be approximated with [2, 28]

$$Z = \frac{\hat{\gamma}}{se(\hat{\gamma})} \approx \sum_{j \in A} \hat{\beta}_j \hat{\sigma} \hat{\sigma} z_j,$$ \hspace{1cm} (4)

where $z_j$ is the $z$-score for the $j$-th SNP in GWAS summary statistics; $\hat{\sigma}^2$ is the sample variance of SNP $j$; $\hat{\sigma}^2$ is the sample variance of the expression level of the gene. We removed the major histocompatibility complex (MHC) region (6p21.3; GRCh38 coordinates 6:28,510,120-33,480,577) in the TWAS analysis of UKBB phenotypes.

Compared methods
PrediXcan
PrediXcan is a TWAS method testing the molecular mechanisms through which genetic variation affects phenotype [2]. For each gene on each target tissue, PrediXcan trains a prediction model
with elastic net, using the accessible genome variation and gene expression levels. The prediction
models are used to impute gene expression in a typical GWAS, where the gene expression is usually
unobserved, for the association test with complex disease. Barbeira et al. [3] extends the PrediXcan to
S-PrediXcan, which can be applied when only GWAS summary statistics are available. The PrediXcan
software is available at https://github.com/hakyimlab/PrediXcan/tree/master/
Software.

**UTMOST**

UTMOST is a TWAS method that trains a cross-tissue expression imputation model by using the
genotype information and matched expression data from multiple tissues. Specifically, the cross-tissue
expression imputation is formulated as a penalized multivariate regression problem, and the effect
sizes are estimated 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). The UTMOST
software is available at https://github.com/Joker-Jerome/UTMOST.

**Data availability**

The genotype and gene expression data of GTEx version 8 project were downloaded from the database
of Genotypes and Phenotypes (accession number phs000424.v8.p2); The genotype and gene ex-
pression data of GEUVADIS LCLs were downloaded from the EBI ArrayExpress database with
accession code E-GEUV-1, https://www.ebi.ac.uk/arrayexpress/experiments/
E-GEUV-1/. GWAS summary statistics from the UKBB were downloaded from the repository
http://www.nealelab.is/uk-biobank. The effective sample sizes for the binary UKBB
phenotypes were calculated by $\frac{n_{\text{case}}n_{\text{control}}}{n_{\text{case}}+n_{\text{control}}}$. The LD matrix was estimated with UKBB European an-
cestry samples, which can be downloaded from https://pan.ukbb.broadinstitute.org
[29]. We partitioned the genome into 1,703 independent blocks using LDetect [30], based on the
1000G reference panel with European ancestry [https://bitbucket.org/nygcresearch/
ldetect-data/src/master/].

**Data preprocessing**

We followed the quality control and sample exclusion process provided by GTEx portal for the
genotype and gene expression datasets [6]. SNPs with minor allele frequency (MAF) < 0.05 or
with strand-ambiguities were removed. The gene expression data was first normalized through rank-
based inverse normal transformation, and was further adjusted for sex, top 3 principal components
from genotypes, and top 15 peer factors (to quantify batch effects and experimental confounders)
[2, 31] were used to remove potential confounding factors. For GTEx, we used 47 tissues with
European ancestry sample sizes larger than 100 for model training. For GEUVADIS datasets, we used
lymphoblastoid cell lines of 344 individuals with European ancestry to perform an external validation
of transcriptome imputation models.
Predictable genes
We considered two thresholds for identifying genes that are significantly predictable, including prediction $R^2 > 0.01$, which is widely used in TWAS analysis such as PrediXcan [2], and a more stringent threshold. We performed a Fisher transformation to the Pearson correlation $r$ between the predicted and true gene expression, which is defined as $z = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$. Under the null hypothesis, $\sqrt{N - 3} \cdot z$ follows a standard normal distribution. The stringent threshold for predictable genes is defined as $|z| > 1.96 \cdot \sqrt{N - 3}$ (so that the p-value < 0.05).

Pathway enrichment analysis
We tested the enrichment of ct-genes and ts-genes in KEGG pathway [19] using the R package "clusterProfiler"[32]. For multiple testing error control, we used the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR) at level 0.05 [33].

Funding
This work was supported in part by National Science Foundation DMS-2015411 to J.S.L.

Acknowledgements
The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. This research has been conducted using the publicly available UK Biobank GWAS summary statistics released by Neale’s lab.

References


Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx
- supplementary.pdf