**** Compiled on 2018/04/06 at 14:31:06 Chicago Time ****

Integrating Predicted Transcriptome From Multiple Tissues Improves Association Detection

Alvaro N. Barbeira[1], Milton D. Pividori[1], Jiamao Zheng[1], Heather E. Wheeler[2,3], Dan L. Nicolae[1,4,5], Hae Kyung Im [1,5,*]

1 Section of Genetic Medicine, The University of Chicago, Chicago, IL, USA

2 Department of Biology, Loyola University Chicago, Chicago, IL, USA

3 Department of Computer Science, Loyola University Chicago, Chicago, IL, USA

4 Department of Statistics, The University of Chicago, Chicago, IL, USA

5 Department of Human Genetics, The University of Chicago, Chicago, IL, USA

∗ Corresponding author haky@uchicago.edu

Abstract

Integration of genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) studies is needed to improve our understanding of the biological mechanisms underlying GWAS hits, and our ability to identify therapeutic targets. Gene-level association test methods such as PrediXcan can prioritize candidate targets. However, limited eQTL sample sizes and absence of relevant developmental and disease context restricts our ability to detect associations. Here we propose an efficient statistical method that leverages the substantial sharing of eQTLs across tissues and contexts to improve our ability to identify potential target genes: MulTiXcan. MulTiXcan integrates evidence across multiple panels while taking into account their correlation. We apply our method to a broad set of complex traits available from the UK Biobank and show that we can detect a larger set of significantly associated genes than using each panel separately. To improve applicability, we developed an extension to work on summary statistics: S-MulTiXcan, which we show yields highly concordant results with the individual level version. Results from our analysis as well as software and necessary resources to apply our method are publicly available.

Author summary

We develop a new method, MulTiXcan, to test the effect of gene expression regulation on complex traits, integrating information available across multiple tissue studies. We show this approach has higher power than traditional single-tissue methods. We extend this method to use only summary-statistics from public GWAS. We apply these methods to over 200 complex traits available in the UK Biobank cohort, and 100 complex traits from public GWAS and discuss the findings.

Introduction

Recent technological advances allow interrogation of the genome to a high level of coverage and precision, enabling experimental studies that query the effect of genotype on both complex and molecular traits. Among these, GWAS have successfully associated genetic loci to human complex traits. GWAS meta-analyses with ever increasing sample sizes allow the detection of associated variants with smaller effect sizes [1–3]. However, understanding the mechanism underlying these associations remains a challenging problem, requiring follow-up studies and a wide array of techniques such as prioritization [4] and pathway analysis [5].

Another approach is the study of quantitative trait loci (eQTLs), measuring association between genotype and gene expression. These studies provide a wealth of biological information but tend to have smaller sample sizes. A similar observation applies to QTL studies of other traits such methylation, metabolites, or protein levels.

The importance of gene expression regulation in complex traits [6–9] has motivated the integration of eQTL studies and GWAS. To examine these mechanisms we developed PrediXcan [10], a method that tests the mediating role of gene expression variation in complex traits. We also developed an extension that accurately infers its gene-level association results using summary statistics data: S-PrediXcan [11]. This allows the rapid exploration of information available in publicly available GWAS summary statistics results, at a dramatically reduced computational burden.

Due to sharing of eQTLs across multiple tissues, we have shown the benefits of an agnostic scanning across all available tissues [11]. Despite the increased multiple testing burden (for Bonferroni correction, the total number of gene-tissue pairs must be used when determining the threshold) we gain considerably in number of significant genes. However, given the substantial correlation between different tissues [12], Bonferroni correction can be too stringent increasing the false negative rate.

In order to aggregate evidence more efficiently, here we present a method termed MultiXcan that tests the joint effects of gene expression variation from different tissues. Furthermore, we develop and implement a method that only needs summary statistics from a GWAS: Summary-MulTiXcan (S-MulTiXcan for short). We make our implementation publicly available to the research community in `https://github.com/hakyimlab/MetaXcan`. We apply this method to traits from the UK Biobank study and over a hundred public GWAS, and publish the results in `http://gene2pheno.org`.

Results

Combining Information Across Tissues Through Multivariate Regression

To combine information across tissues, we regress the phenotype of interest on the predicted expression of the gene in multiple tissues as follows:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{t}_1 g_1 + \mathbf{t}_2 g_2 + \cdots + \mathbf{t}_p g_p + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the n-dimensional phenotype vector, $\boldsymbol{\mu}$ is an intercept term, $\mathbf{t}_i$ is predicted expression of the gene in tissue i, g_i is its effect size, and $\mathbf{e}$ an error term with variance σ_e^2; p is the number of available tissue models.

Expression predictions from many of these tissues are highly correlated. To avoid numerical issues caused by collinearity, we compute the principal components of the predicted expression matrix and discard the axes of smallest variation. Additional covariates can be added to the regression seamlessly. Fig 1-a displays an overview of the method; see further details in the Methods section. We illustrate prediction correlation across models in Supp. Fig 1.

MulTiXcan Detects More Associations Than Single-Tissue PrediXcan

We applied our method to 222 traits from UK Biobank [13]. We used prediction models for 44 tissues trained with Genotype-Tissue Expression (GTEx) samples [12]. We found that it can detect many more associations than PrediXcan using a specific tissue or even when aggregating results from all tissues (Bonferroni-corrected for all gene-tissue pairs tested).

More specifically, we compared three approaches for assessing the significance of a gene jointly across all tissues: 1) running PrediXcan using the most relevant tissue; 2) running PrediXcan using all tissues, one tissue at a time; 3) running MulTiXcan. Fig 1-b illustrates a schematic representation of the results from each approach. Overall, MulTiXcan detects more associations than PrediXcan, as shown in Fig 2-c. Supplementary Data 1 contains a summary of associations per trait. See Supplementary Data 2 and 3 for the list of significant MulTiXcan and PrediXcan results respectively.

We examined the High-Cholesterol trait results in closer detail. We used 50,497 cases and 100,994 controls. After Bonferroni correction, MulTiXcan was able to detect a larger number of significantly associated genes (251) than PrediXcan using all tissues (196) or only a single tissue (whole blood, 33). 172 genes were detected by both PrediXcan and MulTiXcan. Fig 2-a compares the number of significantly associated genes in MulTiXcan, and PrediXcan both for a single tissue (whole blood) and all tissues. Fig 2-b shows the QQ-plot for associations in these three approaches. There are 79 genes associated to high cholesterol via MulTiXcan and not PrediXcan. Among them, we find many genes related to lipid metabolism (*APOM* [14], *PAFAH1B2* [15]), glucose transport(*SLC5A6* [16]), and vascular processes (*NOTCH4* [17]).

MulTiXcan Results Can Be Inferred From GWAS Summary Results

To expand the applicability of our method to massive sample sizes and to studies where individual level data are not available, we have extended our method to use GWAS summary results rather than individual-level data.

We call this extension Summary-MulTiXcan (S-MulTiXcan). Here we derive an analytic expression that relates the joint regression estimates (g_j) to the marginal regression estimates (γ_j as obtained from S-PrediXcan), assuming a known LD structure from a reference panel. We display a conceptual overview of the method in Fig 4-a. See details in the Methods Section.

Fig 3 displays a few examples of the general agreement between the individual-level MulTiXcan and S-MulTiXcan. The summary-based version's results tend to be slightly less significant than MulTiXcan, as illustrated in Supp. Fig 2.

To reduce false positives due to LD misspecification, we discard any significant association result for a gene when the best single tissue result has p-value greater than 10^{-4}. This is rather conservative since it is possible that evidence with modest significance from weakly correlated tissues can lead to very significant combined association. We have, in fact, found several such instances using individual level data.

Application to a broad set of complex traits

We applied S-MulTiXcan to over a 100 traits on publicly available GWAS. As with the individual level method, we observed that S-MulTiXcan detects more associations than S-PrediXcan in most cases, as shown in Fig 4-b, after discarding suspicious associations. We also show the QQ-plots and total number of detected association for a sample trait (Schizophrenia) on Fig 4-c and 4-d.

These results have been incorporated into the publicly available catalog at `http://gene2pheno.org`. The list of analyzed traits can be found in Supplementary Data 4. Supplementary Data 5 contains a summary of significant associations for each trait. Supplementary Data 6 lists the significant S-MulTiXcan results for each trait.

Highlight Of Associations Identified By Summary-MulTiXcan

We examined the biological relevance of some of the genes detected by our new method that was missed by looking at one tissue at a time (S-PrediXcan).

For example, in the Early Growth Genetics (EGG) Consortium's Body-Mass Index (BMI) study, S-MulTiXcan detects three genes not significant in S-PrediXcan: *POMC* (p-value=1.4×10^{-6}, tied to childhood obesity [18]); *RACGAP1* (p-value= 1.2×10^{-10}; embryogenesis [19], cell growth and differentiation, [20]); and *TUBA1B* (p-value=1.23×10^{-09}, circadian cycle processes and psychological disorders [21], suggesting a behavioral pathway).

In the CARDIoGRAM+C4D Coronary Artery Disease (CAD) study, S-MulTiXcan detected 12 associations not significant in S-PrediXcan. The top result was *AS3MT* (p-value=4.3×10^{-9}), related to arsenic metabolism; interestingly, environmental and toxicological studies link arsenic exposure and *AS3MT* polymorphisms with cardiovascular disease [22, 23]. Associations previously linked to CAD included *CDKN2B* (p-value$< 1.0 \times 10^{-6}$, [24]) *HECTD4* (p-value$< 2.3 \times 10^{-6}$, [25]). Other interesting S-MulTiXcan findings were *CLCC1* (pvalue=1.2×10^{-7}, a gene for chloride channel activity); *IREB2* (p-value=2.1×10^{-7}, recently linked to pulmonary conditions, [26]), and *ADAM15* (p-value=2.5×10^{-07}, from the disintegrin and metalloproteinase family, linked to atherosclerosis [27], atrial fibrillation [28], and other vascular processes [29, 30]).

The list of significant S-MulTiXcan and S-PrediXcan results for all traits can be found in Supplementary Data 6 and 7.

Discussion

Motivated by the widespread sharing of regulatory processes across tissues [12], we propose here a method (MulTiXcan) that aggregates information across multiple tissues and improves the identification of genes significantly associated with complex traits. To expand the applicability of our approach we have extended the method to accommodate GWAS studies where only summary results are available (S-MulTiXcan). We show through applications to hundreds of traits the performance of both individual and summary based methods. We also show that the summary based method provides a reasonably good approximation to the individual level results.

As any method relying on a reference panel, S-MulTiXcan might be inaccurate when the study population has a different Linkage Disequilibrium (LD) structure than the reference panel. For example, should two models for a gene yield predicted expressions that are lowly correlated in the reference panel but highly correlated in the study population, then this method underestimates their correlation. To avoid this misspecification, a reference panel matching the study population should be used when available (i.e. using East Asian population from 1000 Genomes if the study set is composed of East Asian individuals).

A limitation of PrediXcan and S-PrediXcan is LD contamination, i.e. when causal loci for the trait and expression are different but in LD. We have addressed this in S-PrediXcan through an additional colocalization filtering step. For MulTiXcan, this could be avoided by restricting the analysis to gene-tissue pairs with high colocalization probability.

Here we showed the advantages of our joint estimation method through application to multiple traits with publicly available GWAS results as well as the ones available in the UK Biobank. These results include many novel associations of interest, which we make publicly available to the research community in `http://gene2pheno.org`.

Software And Resources

We make our software publicly available on a GitHub repository: `https://github.com/hakyimlab/MetaXcan`. Prediction model weights and covariances for different tissues can be downloaded from PredictDB.org. A short working example can be found on the GitHub page; more extensive documentation can be found on the project's wiki. The results of S-MulTiXcan applied to the 44 human tissues and a broad set of phenotypes can be queried on `http://gene2pheno.org`.

Methods

Definitions, Notation And Preliminaries

Let us consider a GWAS study of n samples, and assume availability of prediction models in p different tissues. Each model j is a collection of prediction weights w_i^j.

Let:

- $\mathbf{y}$ be an n-vector of phenotypes, assumed to be centered for convenience.
- $\mathbf{X}$ the genotype matrix, where each column X_l is the n-vector genotype for SNP l. We assume it coded in the range [0,2] but it can be defined in another range, or standardized.
- $\tilde{\mathbf{t}}_j = \sum_{i \in \text{model}_j} w_i^j X_i$ be the predicted expression in tissue j. This is the independent variable used by single-tissue PrediXcan.
- $\mathbf{t}_j$ be the standardization of $\tilde{\mathbf{t}}_j$ to $mean = 0$ and $standard\ deviation = 1$.

In our application, we will consider $p = 44$ models for a given gene's expression, trained on GTEx data. This method is easily extensible to support incorporation of other covariates, or correction by them.

MulTiXcan

MulTiXcan consists of fitting a linear regression of the phenotype on predicted expression from multiple tissue models jointly:

$$\begin{aligned} \mathbf{y} &= \sum_{j=1}^{p} \mathbf{t}_j g_j + \mathbf{e} \\ &= \mathbf{T}\mathbf{g} + \mathbf{e}, \end{aligned} \tag{2}$$

where $\mathbf{y}$ is a vector of phenotypes for n individuals, $\mathbf{t}_j$ is an n-vector of standardized predicted gene expression for model j, g_j is the effect size for the predicted gene expression j, $\mathbf{e}$ is an error term with variance σ_e^2, and p is the number of tissues; thus $\mathbf{T}$ is a data matrix where each column j contains the values from $\mathbf{t}_j$, and $\mathbf{g}$ is the p-vector of effect sizes g_j. One of this columns is a constant intercept term.

The high degree of eQTL sharing between different tissues induces a high correlation between predicted expression levels. In order to avoid collinearity issues and numerical instability, we decompose the predicted expression matrix into principal components and keep only the eigenvectors of non negligible variance. To select the number of components, we used a condition number threshold of $\frac{\lambda_{\text{max}}}{\lambda_i} < 30$, where λ_i is an eigenvalue of the matrix $\mathbf{T}^t\mathbf{T}$. A range of values between 10 and 100 yielded similar results for significance in real data. See the next section for additional details in the number of components used.

Lastly, we use an F-test to quantify the significance of the joint fit.

We use Bonferroni correction to determine the significance threshold. For MulTiXcan, we use the total number of genes with a prediction model in at least one tissue, which yields a threshold approximately at $0.05/17500 \sim 2.9 \times 10^{-6}$. For PrediXcan across all tissues, we use the total number of gene-tissue pairs, which yields a threshold approximately at $0.05/200,000 \sim 2.5 \times 10^{-7}$.

Application To UK Biobank Data

We used the same covariates reported in [31], which include the first ten genotype principal components, sex, age, genotyping array, and depending on the trait others such as body mass index (BMI), weight or height. We used 44 models trained on GTEx tissues from release version v6p. For diseases, we used twice as many healthy individuals as controls, selected at random. For the MulTiXcan-significant associations in the 222 traits, the median number of available models is 11 ($1Q = 7$, $3Q = 16$), with $\sim 77\%$ components surviving PCA thresholding.

Summary-MulTiXcan

We have demonstrated that S-PrediXcan can accurately infer PrediXcan results from GWAS Summary Statistics and LD information from a reference panel [11], with the added benefits of reduced computational and regulatory burden. Here we extend MulTiXcan in a similar fashion.

Summary-MulTiXcan (S-MulTiXcan) infers the individual-level MulTiXcan results, using univariate S-PrediXcan results and LD information from a reference panel. It consists of the following steps:

- Computation of single tissue association results with S-PrediXcan.
- Estimation of the correlation matrix of predicted gene expression for the models using the Linkage Disequilibrium (LD) information from a reference panel (typically GTEx or 1000 Genomes [32])
- Discarding components of smallest variation from this correlation matrix to avert collinearity and numerical problems (Singular Value Decomposition, analogue to PC analysis in individual-level data).
- Estimation of joint effects from the univariate (single-tissue) results and expression correlation.
- Discarding suspicious results, suspect to be false positives arising from LD-structure mismatch.

Joint Analysis Estimation From Marginal Effects

To derive the multivariate regression (2) effect sizes and variances using the marginal regression (3) estimates, we employ a technique presented in [33].

More specifically, we want to obtain the multivariate regression coefficient estimates for g_j (2) using the estimates from the marginal regression:

$$\mathbf{y} = \mathbf{t}_j \gamma_j + \epsilon_j. \tag{3}$$

where we assume $\mathbf{y}$ centered for convenience, and ϵ_j is the marginal regression error term with variance σ_ϵ^2 (i.e. we assume a common variance σ_ϵ^2 for all j).

First, notice that the solution to the multivariate regression in Eq (2) is

$$\hat{\mathbf{g}} = \left(\mathbf{T}^t \mathbf{T}\right)^{-1} \mathbf{T}^t \mathbf{y} \tag{4}$$

$$\mathrm{var}(\hat{\mathbf{g}}) = \sigma_e^2 (\mathbf{T}^t \mathbf{T})^{-1} \tag{5}$$

whereas the solution to the marginal regression in Eq (3) is:

$$\hat{\gamma} = \mathbf{D}^{-1} \mathbf{T}^t \mathbf{y} \tag{6}$$

$$\mathrm{var}(\hat{\gamma}) = \sigma_\epsilon^2 \mathbf{D}^{-1} \quad \text{with } \mathbf{D} = \mathrm{diag}(\mathbf{T}^t \mathbf{T}) \tag{7}$$

where $\boldsymbol{\gamma}$ is the vector of effect sizes γ_j, and $\mathbb{1}$ is the $p \times p$ identity matrix. Please note that, since the $\mathbf{t}_j$ are standardized, then $\mathbf{D} = (n-1)\mathbb{1}$ and $se(\gamma_j) = \sqrt{var(\gamma_j)} = \frac{\sigma_\epsilon}{\sqrt{n-1}}$.

From (6) we get $\mathbf{T}^t\mathbf{y} = \mathbf{D}\hat{\boldsymbol{\gamma}}$, which we replace in (4) and obtain the relationship between marginal and joint estimates:

$$\hat{\mathbf{g}} = \left(\mathbf{T}^t\mathbf{T}\right)^{-1}\mathbf{D}\hat{\boldsymbol{\gamma}} \tag{8}$$

To compute the variance of the estimated effect sizes (5) we use the variance of the phenotype as a conservative estimate of σ_e^2 and LD information from reference samples as described next.

Estimating Expression Correlation From A Reference Panel

As the genotypes from most GWAS are typically unavailable, we must use a reference panel to compute $\mathbf{T}^t\mathbf{T}$, using only those SNPS available in the GWAS results. To do so, notice that:

$$\begin{aligned}
\frac{(\mathbf{T}^t\mathbf{T})_{ij}}{n-1} &= Cor(\mathbf{t}_i, \mathbf{t}_j) \\
&= Cov(\mathbf{t}_i, \mathbf{t}_j) \\
&= \frac{Cov\left(\tilde{\mathbf{t}}_i, \tilde{\mathbf{t}}_j\right)}{\sqrt{\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_i)\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_j)}} \\
&= \frac{Cov\left(\sum_{a \in \mathrm{model}_i} w_a^i X_a, \sum_{b \in \mathrm{model}_j} w_b^j X_b\right)}{\sqrt{\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_i)\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_j)}} \\
&= \frac{\sum_{\substack{a \in \mathrm{model}_i \\ b \in \mathrm{model}_j}} w_a^i w_b^j Cov\left(X_a, X_b\right)}{\sqrt{\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_i)\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_j)}} \\
&= \frac{\sum_{\substack{a \in \mathrm{model}_i \\ b \in \mathrm{model}_j}} w_a^i w_b^j \Gamma_{ab}}{\sqrt{\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_i)\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_j)}},
\end{aligned} \tag{9}$$

where Γ_{ij} are the elements of the covariance matrix $\mathbf{\Gamma} = \widehat{\mathrm{var}}\,(\mathbf{X}) = (\mathbf{X} - \bar{\mathbf{X}})^{\mathbf{t}}(\mathbf{X} - \bar{\mathbf{X}})/(n-1)$. We compute the variances as in the S-PrediXcan analysis:

$$\begin{aligned}\widehat{\mathrm{var}}(\tilde{\mathbf{t}}_j) &= \hat{\sigma_j^2} \\ &= (\mathbf{W}^j)^t\,\mathbf{\Gamma}^j\,\mathbf{W}^j \\ &= \sum_{\substack{a\in\mathrm{model}_j\\ b\in\mathrm{model}_j}} w_a^j w_b^j \Gamma_{ab}^j\end{aligned} \tag{10}$$

Addressing Singularity Of The Correlation Matrix

Given the high degree of correlation among many of the prediction models, $\mathbf{T}^t\mathbf{T}$ is close to singular and its inverse cannot be reliably calculated for many genes. To address this problem, we compute the pseudo-inverse via Singular Value Decomposition, decomposing the correlation matrix into its principal components and removing those with small eigenvalues. In other terms, we will restrict the analysis to axes of largest variation of the expression data. This is analogous to the principal components-based approach used with individual level data. We denote with $\mathbf{\Sigma}^+$ the pseudo-inverse for any matrix $\mathbf{\Sigma}$. We use the same condition number from individual-level MultiXcan ($\frac{\lambda_{\max}}{\lambda_i} < 30$) as threshold. For S-MulTiXcan-significant associations across 100 public traits, we found a median number of available models of 9 ($1Q = 5$, $3Q = 15$), with $\sim 80\%$ of components surviving the SVD threshold.

Estimating Significance

To quantify significance, we use the fact that the regression coefficient estimates follow a (approximate) multivariate normal distribution: $\hat{\mathbf{g}} \sim \mathcal{N}(\mathbf{g}, \sigma_e^2\,(\mathbf{T}^t\mathbf{T})^{-1})$. Under the null hypothesis of no association, it follows that $\hat{\mathbf{g}}^t \frac{\mathbf{T}^t\mathbf{T}}{\sigma_e^2} \hat{\mathbf{g}} \sim \chi_p^2$. We can then replace $\hat{\mathbf{g}}$ with its estimate from the marginal regression:

$$\begin{aligned}
\frac{\hat{\mathbf{g}}^t(\mathbf{T}^t\mathbf{T})\hat{\mathbf{g}}}{\sigma_e^2} &= \frac{\hat{\boldsymbol{\gamma}}^t\mathbf{D}\left(\mathbf{T}^t\mathbf{T}\right)^{-1}\mathbf{T}^t\mathbf{T}\left(\mathbf{T}^t\mathbf{T}\right)^{-1}\mathbf{D}\hat{\boldsymbol{\gamma}}}{\sigma_e^2} \\
&= \frac{\hat{\boldsymbol{\gamma}}^t\mathbf{D}}{\sigma_e}\left(\mathbf{T}^t\mathbf{T}\right)^{-1}\frac{\mathbf{D}\hat{\boldsymbol{\gamma}}}{\sigma_e} \\
&\approx \frac{\hat{\boldsymbol{\gamma}}^t\mathbb{1}(n-1)}{\sigma_\epsilon}\left(\mathbf{T}^t\mathbf{T}\right)^{-1}\frac{(n-1)\mathbb{1}\hat{\boldsymbol{\gamma}}}{\sigma_\epsilon} \\
&\approx \hat{\boldsymbol{\gamma}}^t\frac{\sqrt{n-1}}{\sigma_\epsilon}\left(\frac{\mathbf{T}^t\mathbf{T}}{n-1}\right)^{-1}\frac{\sqrt{n-1}}{\sigma_\epsilon}\hat{\boldsymbol{\gamma}} \\
&\approx \hat{\mathbf{z}}^t Cor(\mathbf{T})^{-1}\hat{\mathbf{z}},
\end{aligned}$$

where $Cor(\mathbf{T})$ is the autocorrelation of $\mathbf{T}$, and $\hat{\mathbf{z}}$ is the p-vector of marginal analysis z-scores, $\gamma_j/se(\gamma_j)$. We have used $\sigma_e^2 \approx \sigma_\epsilon^2$ as an approximation (i.e. the residual variance of the *marginal* regression as approximation of the residual variance of the *joint* regression). This simplification is conservative, and based on our comparison to the individual multivariate results we consider the loss of efficiency acceptable.

In practice, we will use the SVD pseudo-inverse $Cor(\mathbf{T})^+$ as explained in the previous section, and a χ^2-test: $\hat{\mathbf{z}}^t Cor(\mathbf{T})^+\hat{\mathbf{z}} \sim \chi_k^2$, with k the number of components surviving the SVD pseudoinverse.

Implementation And Computation

Prediction Models were obtained from PredictDB.org resource. These models were trained using Elastic Net as implemented in R's package *glmnet* [34], with a mixing parameter $\alpha = 0.5$, over 44 tissue studies from GTEx' release version 6p. The underlying GTEx study data was obtained from dbGaP with accesion number phs000424.v6.p1. Please see [11] for details. We implemented MulTiXcan and S-MulTiXcan working up from existing software in the MetaXcan package.

UK Biobank genotype data for $487,409$ individuals was downloaded and processed in the Bionimbus Protected Data Cloud (PDC), a secure biomedical cloud operated at FISMA moderate as IaaS with an NIH Trusted Partner status for analyzing and sharing protected datasets. We computed GWAS results using BGENIE, a program for efficient GWAS for multiple continuous traits [35]. We selected 222 traits available for these individuals, covering continuous phenotypes such as height and self reported diseases such as asthma. We used different covariate groups for these phenotypes as in [31]. Age, sex and the top ten principal components were used in all cases. For diseases, we randomly sampled twice as many healthy controls as there were cases. Gene expression prediction was computed on the genotype data using the 44 GTEx models.

When running MulTiXcan, we used the same covariates and data as in the GWAS. On most continuous phenotypes, there were between 300, 000 and 400, 000 individuals with available data determined by the intersection of covariates and traits. For the case of self reported diseases, we found a number of cases ranging from a few hundreds (i.e. Acne) to 50, 000 (i.e. High Cholesterol). We also ran S-PrediXcan on 105 public GWAS traits (the same analyzed in [11], see Supplementary Data 4 for details).

Acknowledgments

Grants

We acknowledge the following US National Institutes of Health grants: R01MH107666 (H.K.I.), R01 MH101820 (GTEx)

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc. Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035). Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to contract N01MH000028. Statistical Methods development grants were made to the University of Geneva (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania (MH101822). The data used for the analyses described in this manuscript were obtained from dbGaP accession number phs000424.v6.p1 on 06/17/2016.

This work was completed in part with resources provided by Bionimbus [36], and the Center for

Research Informatics. The Center for Research Informatics is funded by the Biological Sciences Division at the University of Chicago with additional funding provided by the Institute for Translational Medicine, CTSA grant number UL1 TR000430 from the National Institutes of Health.

Figures

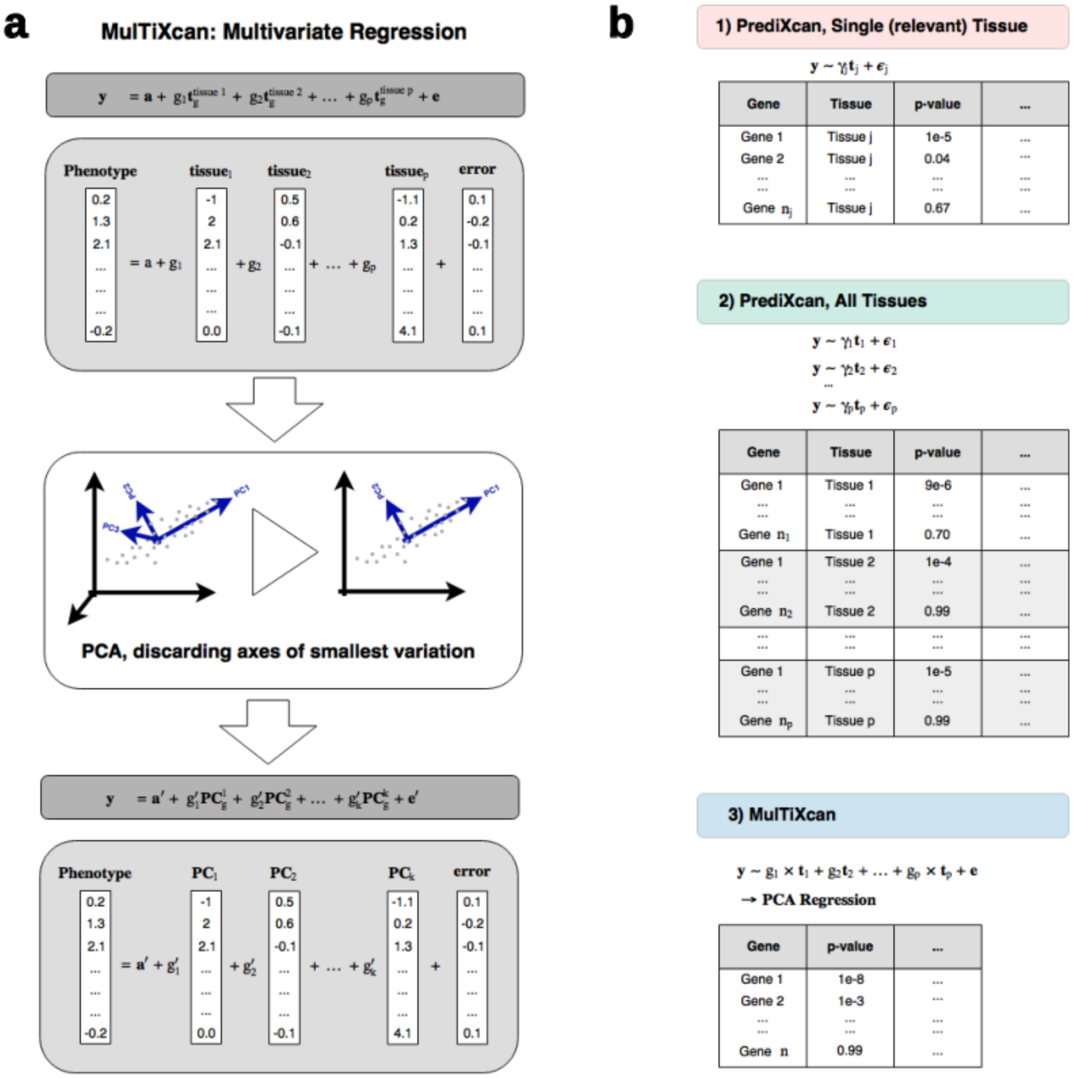


Figure 1. MulTiXcan method.
Panel a illustrates MulTiXcan method. Predicted expression from all available tissue models are used as explanatory variables. To avoid multicolinearity, we use the first k Principal Components of the predicted expression. $\mathbf{y}$ is a vector of phenotypes for n individuals, $\mathbf{t}_g^{\text{tissue}\,j}$ is the standardized predicted gene expression for tissue j, $\mathbf{g}_j$ is its effect size, $\mathbf{a}$ is an intercept and $\mathbf{e}$ is an error term.
Panel b shows a schematic representation of MulTiXcan results compared to classical PrediXcan, both for a single relevant tissue and all available tissues in agnostic scanning. $\mathbf{y}$ is a vector of phenotypes for n individuals, $\mathbf{t}_j$ is the standardized predicted gene expression for model j, g_j is its effect size in the joint regression, γ_j is its effect size in the marginal regression using only prediction j, $\mathbf{e}$ and ϵ_j are error terms.

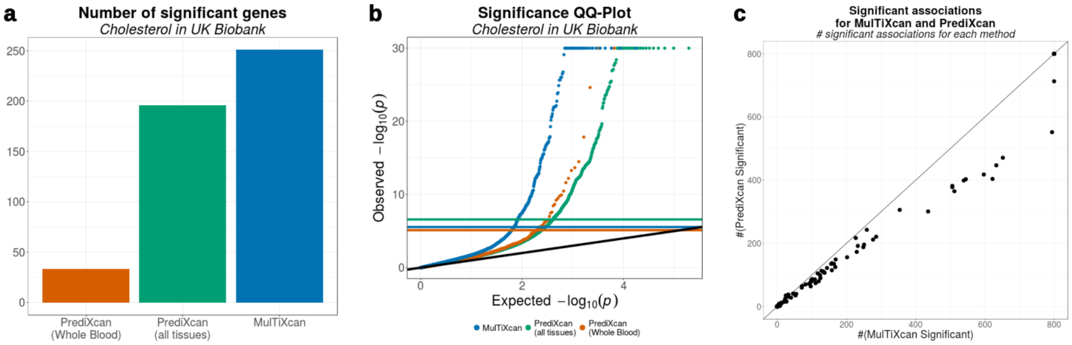


Figure 2. Joint testing across all tissues increases number of significant genes.
Panel a shows the number of discoveries in each method for Cholesterol trait. MulTiXcan is able to detect more findings (251 significant associations) than either of PrediXcan approaches (33 using only Whole Blood and 196 using all 44 GTEx tissues).
Panel b compares the distribution of MulTiXcan's p-values to PrediXcan's p-values for the Cholesterol trait in the UK Biobank cohort. Both PrediXcan with a single tissue model (GTEx Whole Blood) and 44 models (GTEx v6p models) are shown. Notice that Bonferroni-significance levels are different for each case, since 6588 genes were tested in PrediXcan for Whole Blood, 195532 gene-tissue pairs for all GTEx tissues, and 17434 genes in MulTiXcan. P-values were truncated at 10^{-30} for visualization convenience.
Panel c compares the number of significant associations discovered by MulTiXcan and PrediXcan for 222 traits from UK Biobank. These numbers were thresholded at 800 for visualization purposes.

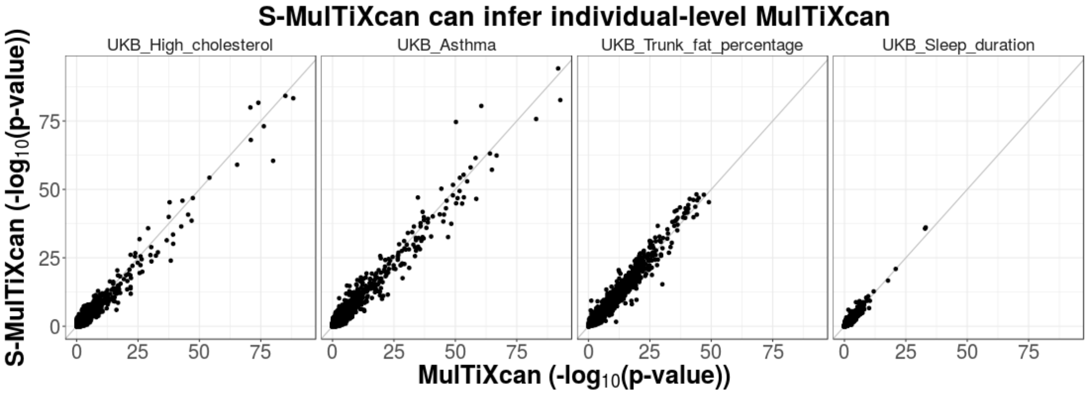


Figure 3. MulTiXcan results can be inferred from GWAS summary statistics and a reference panel. This figure compares S-MulTiXcan to MulTiXcan in three UK Biobank phenotypes. GTEx individuals were used as a reference panel for estimating expression correlation in the study population. The summary data-based method shows a good level of agreement with the individual-based method. In cases where the LD-structure between reference and study cohorts is mismatched, the summary-based method becomes less accurate. For example in Asthma, two genes are significantly overestimated; however it tends to be conservative for most genes.

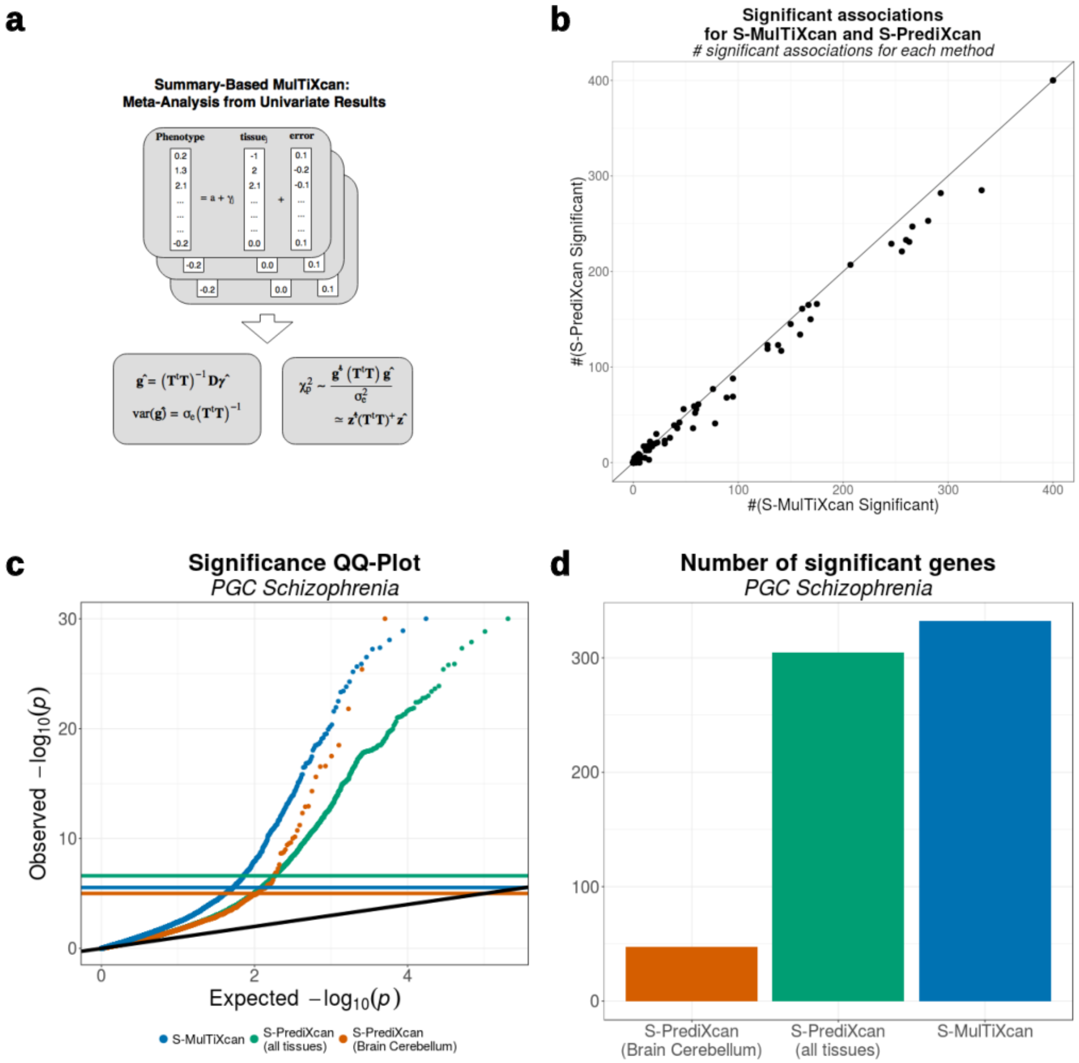


Figure 4. Comparison between S-PrediXcan and S-MulTiXcan.

Panel a illustrates the S-MulTiXcan method: the marginal univariate S-PrediXcan effect sizes are computed, then the joint effect sizes are estimated from them. The significance is quantified through an omnibus test.

Panel b compares the number of associations significant via S-MulTiXcan versus those significant via S-PrediXcan, for the same GWAS Studies. In most cases, S-MulTiXcan detects a larger number of exclusive significant associations. The number of discoveries was thresholded at 200 for visualization purposes.

Panel c displays QQ-Plots for the association p-values from S-MulTiXcan and S-PrediXcan in Schizophrenia, using a model trained on brain's cerebellum, and S-PrediXcan associations for all 44 GTEx tissues.

Panel d shows the number of significant associations in Schizophrenia for each method as a bar plot.

Tables

References

1. Smoller JW, Craddock N, Kendler K, Lee PH, Neale BM, Nurnberger JI, et al. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet. 2013;381(9875):1371–9. Available from: `http://discovery.ucl.ac.uk/1395494/$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/23453885`.

2. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nature genetics. 2013;45(1):25–33. Available from: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3679547{&}tool=pmcentrez{&}rendertype=abstract`.

3. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature Genetics. 2012;44(9):981–990. Available from: `http://www.ncbi.nlm.nih.gov/pubmed/22885922$\delimiter"026E30F$nhttp://www.nature.com/doifinder/10.1038/ng.2383`.

4. Moreau Y, Tranchevent LC. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nature Reviews; Genetics. 2012;13(8):523–536.

5. Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. American Journal of Human Genetics. 2010;86(1):6–22.

6. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genetics. 2010;6(4).

7. Nicolae DL, Gamazon E, Zhang W, Duan S, Eileen Dolan M, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. PLoS Genetics. 2010;6(4).

8. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016;352(6285):600–604. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27126046.

9. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. American Journal of Human Genetics. 2014;95(5):535–552.

10. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. Nature genetics. 2015;47(9):1091–1098. Available from: http://dx.doi.org/10.1038/ng.3367.

11. Barbeira A, Dickinson SP, Torres JM, Bonazzola R, Zheng J, Torstenson ES, et al. Integrating tissue specific mechanisms into GWAS summary results. bioRxiv. 2017;Available from: http://www.biorxiv.org/content/early/2017/05/21/045260.

12. Aguet F, Brown AA, Castel S, Davis JR, Mohammadi P, Segre AV, et al. Local genetic effects on gene expression across 44 human tissues. bioRxiv. 2016;Available from: http://biorxiv.org/content/early/2016/09/09/074450.

13. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Medicine. 2015;12(3).

14. Xu N, Dahlbäck B. A novel human apolipoprotein (apoM). The Journal of biological chemistry. 1999;274(44):31286–90. Available from: http://www.jbc.org.ezproxy.lib.ucalgary.ca/content/274/44/31286.full{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/10531326.

15. Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. American Journal of Human Genetics. 2014;94(2):223–232.

16. Wright EM, Turk E. The sodium/glucose cotransport family SLC5; 2004.

17. Gridley T. Notch signaling in vascular development and physiology. Development (Cambridge, England). 2007;134(15):2709–2718.

18. Kuehnen P, Mischke M, Wiegand S, Sers C, Horsthemke B, Lau S, et al. An alu element-associated hypermethylation variant of the POMC gene is associated with childhood obesity. PLoS Genetics. 2012;8(3).

19. Grewal S, Carver JG, Ridley AJ, Mardon HJ. Implantation of the human embryo requires Rac1-dependent endometrial stromal cell migration. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(42):16189–16194. Available from: `http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed{&}id=18838676{&}retmode=ref{&}cmd=prlinks{%}5Cnpapers2://publication/doi/10.1073/pnas.0806219105`.

20. Hallstrom TC, Mori S, Nevins JR. An E2F1-Dependent Gene Expression Program that Determines the Balance between Proliferation and Cell Death. Cancer Cell. 2008;13(1):11–22.

21. Byrne EM, Heath AC, Madden PAF, Pergadia ML, Hickie IB, Montgomery GW, et al. Testing the role of circadian genes in conferring risk for psychiatric disorders. American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics. 2014;165(3):254–260.

22. Gong G, O'Bryant SE. Low-level arsenic exposure, AS3MT gene polymorphism and cardiovascular diseases in rural Texas counties. Environmental Research. 2012;113:52–57.

23. Moon K, Guallar E, Navas-Acien A. Arsenic exposure and cardiovascular disease: An updated systematic review; 2012.

24. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide Association Analysis of Coronary Artery Disease. New England Journal of Medicine. 2007;357(5):443–453. Available from: `http://www.nejm.org/doi/abs/10.1056/NEJMoa072366`.

25. Lu X, Wang L, Chen S, He L, Yang X, Shi Y, et al. Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. Nature Genetics. 2012;44(8):890–894.

26. DeMeo DL, Mariani T, Bhattacharya S, Srisuma S, Lange C, Litonjua A, et al. Integration of Genomic and Genetic Approaches Implicates IREB2 as a COPD Susceptibility Gene. American Journal of Human Genetics. 2009;85(4):493–502.

27. Oksala N, Levula M, Airla N, Pelto-Huikko M, Ortiz RM, JÃďrvinen O, et al. ADAM-9, ADAM-15, and ADAM-17 are upregulated in macrophages in advanced human atherosclerotic plaques in aorta and carotid and femoral arteriesâĂŤTampere vascular study. Annals of Medicine. 2009;41(4):279–290. Available from: http://dx.doi.org/10.1080/07853890802649738.

28. Arndt M, Lendeckel U, Röcken C, Nepple K, Wolke C, Spiess A, et al. Altered expression of ADAMs (A Disintegrin And Metalloproteinase) in fibrillating human atria. Circulation. 2002;105(6):720–725.

29. Xie B, Shen J, Dong A, Swaim M, Hackett SF, Wyder L, et al. An Adam15 amplification loop promotes vascular endothelial growth factor-induced ocular neovascularization. FASEB journal : official publication of the Federation of American Societies for Experimental Biology. 2008;22(8):2775–83. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2493454{&}tool=pmcentrez{&}rendertype=abstract.

30. Komiya K, Enomoto H, Inoki I, Okazaki S, Fujita Y, Ikeda E, et al. Expression of ADAM15 in rheumatoid synovium: up-regulation by vascular endothelial growth factor and possible implications for angiogenesis. Arthritis research & therapy. 2005;7(6):R1158–R1173. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1297561{&}tool=pmcentrez{&}rendertype=abstract.

31. Ge T, Chen CY, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. PLoS Genetics. 2017;13(4).

32. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. Available from: http://www.nature.com/doifinder/10.1038/nature15393$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/26432245.

33. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics. 2012;44(4):369–375. Available from: http://www.nature.com/doifinder/10.1038/ng.2213.

34. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33(1):1–22. Available from: http://www.jstatsoft.org/v33/i01/.

35. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. Genome-wide genetic data on ˜500,000 UK Biobank participants. bioRxiv. 2017;p. 166298. Available from: https://www.biorxiv.org/content/early/2017/07/20/166298.

36. Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, et al. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. Journal of the American Medical Informatics Association : JAMIA. 2014 Nov;21(6):969–975. Available from: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-002155.

Supplementary Material

Supplementary Data

Supplementary Data 1. Summary statistics for UK Biobank traits used in the MulTiXcan analysis. MulTiXcan was run for 222 traits on UK Biobank. Summary statistics for significant results included in **supp-data-ukb-multixcan-stats.txt**. Columns are: **tag**: trait, gene2pheno.org display name; **n_predixcan_significant**: Number of Bonferroni-significant PrediXcan results; **n_MulTiXcan_significant** number of Bonferroni-significant results for MulTiXcan; **n_predixcan_only** number of results only significant in PrediXcan; **n_MulTiXcan_only** number of results only significant in MulTiXcan.

Supplementary Data 2. Significant associations for MulTiXcan on UK Biobank. Significant results included in **supp-data-ukb-multixcan-significant.txt**. Columns are: **phenotype**: trait, gene2pheno.org display name; **gene**: Ensembl id; **gene_name**: HUGO name; **pvalue**: p-value of the S-MulTiXcan association; **n_models** number of prediction models available for the gene; **n_used** number of independent components surviving PCA selection; **n_samples**: number of individuals available.

Supplementary Data 3. Significant associations for PrediXcan on UK Biobank. Significant results included in **supp-data-ukb-p-significant.txt**. Columns are: **Phenotype**: trait, gene2pheno.org display name; **model**: GTEx tissue where the model was trained; **gene**: Ensembl Id; **gene_name**: HUGO name; **model** GTEx tissue where model was trained; **zscore** PrediXcan association Z-score, **pvalue** PrediXcan association p-value; **n_samples**: number of individuals available.

Supplementary Data 4. List of Genome-wide Association Meta Analysis (GWAMA) Consortia and phenotypes. Data included in **supp-data-gwas-traits.txt**. Columns are consortium name, study name, gene2pheno.org display name, study sample size, study population, URL of portal where data was downloaded from, link to pubmed entry if available.

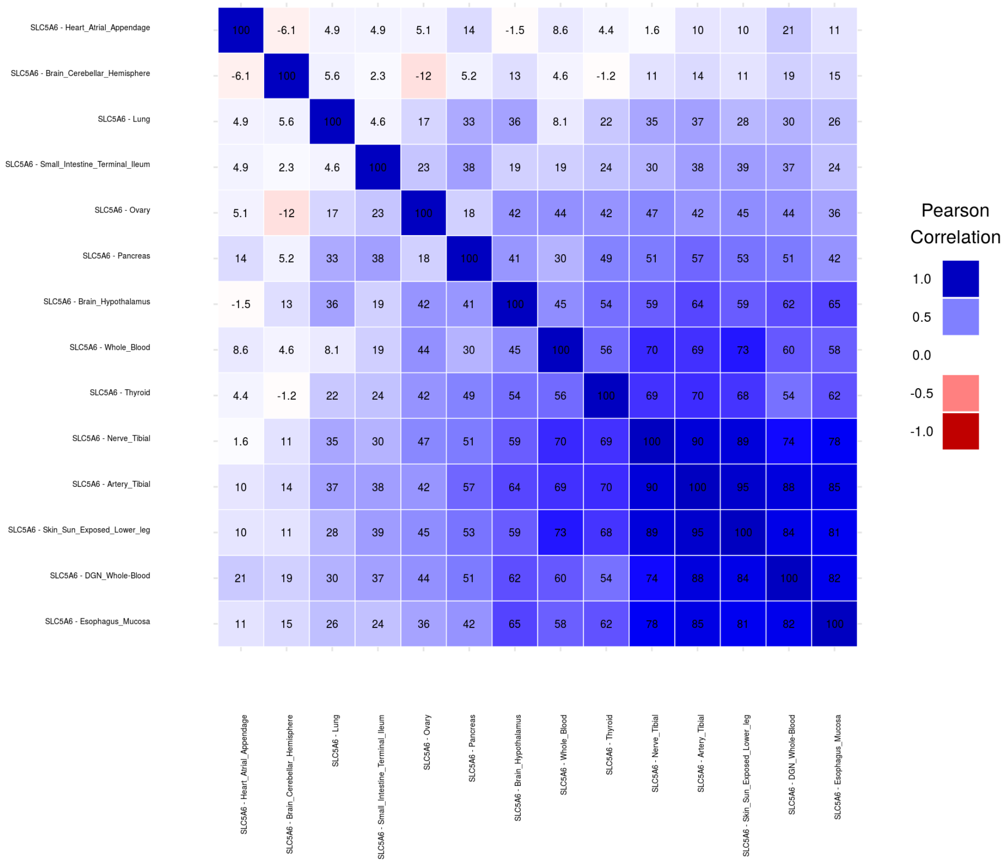
Supplementary Data 5. Summary statistics for traits used in the MulTiXcan analysis. MulTiXcan was run for 105 public GWAS. Summary statistics for significant results included in **supp-data-gwas-smultixcan-stats.txt**. Columns are: **tag**: gene2pheno.org display name; **consortium**: Consortium Name; **name**: study name; **n_spredixcan_significant**: Number of Bonferroni-significant S-PrediXcan results; **n_sMulTiXcan_significant** number of Bonferroni-significant results for MulTi-

Xcan; **n_spredixcan_only** number of results only significant in S-PrediXcan; **n_sMulTiXcan_only** number of results only significant in S-MulTiXcan.

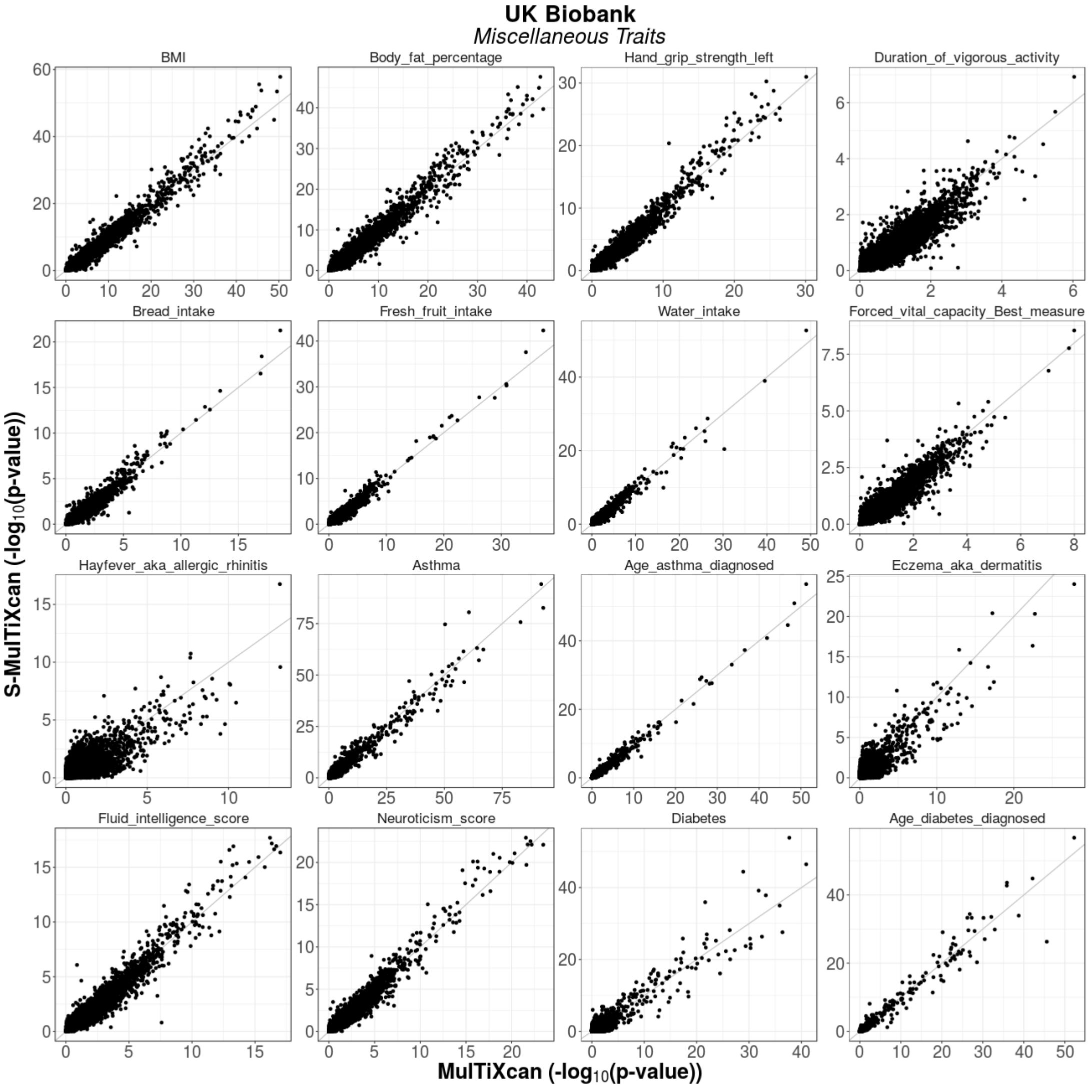
Supplementary Data 6. Significant associations for Summary-MulTiXcan on public GWAS. Significant results included in **supp-data-gwas-smultixcan-significant.txt**. Columns are: **tag**: gene2pheno.org display name; **consortium**: Consortium Name; **name**: study name; **gene**: Ensembl id; **gene_name**: HUGO name; **pvalue**: p-value of the S-MulTiXcan association; **n** number of S-PrediXcan results available for the gene; **n_indep** number of independent components surviving SVD; **p_i_best** best p-value of S-PrediXcan;**t_i_best** tissue that presented best S-PrediXcan result; **p_i_worst** worst p-value of S-PrediXcan; **t_i_worst** tissue that presented worst S-PrediXcan result; **suspicious**: whether the result was discarded as a potential false positive.

Supplementary Data 7. Significant associations for Summary-PrediXcan on public GWAS. Significant results included in **supp-data-gwas-sp-significant.txt**. Columns are: **consortium**: Consortium Name; **name**: study name; **tag**: gene2pheno.org display name; **gene**: Ensembl Id; **gene_name**: HUGO name; **model** GTEx tissue where model was trained; **zscore** S-PrediXcan association Z-score, **pvalue** S-PrediXcan association p-value.

455 Supplementary Figures



Supplementary Figure 1. Predicted expression correlation for gene *SLC5A6*. We observe a high degree of predicted expression correlation, in agreement with recent publications on the high degree of mechanism sharing across tissues [12]. This behavior is exhibited in most genes.



Supplementary Figure 2. Summary-MulTiXcan vs MulTiXcan for Miscellaneous Traits. There is a satisfactory agreement between the individual-level and the summary-level versions of MulTiXcan in UK Biobank traits.