A Fast and Accurate Gene Level Association Method for Mapping Traits Using Reference Transcriptome Data

Alvaro Barbeira[1], Kaanan P. Shah[2], Jason M. Torres [3], Heather E Wheeler [4], GTEx Consortium ?, Dan Nicolae[1], Nancy J Cox[5], Hae Kyung Im[2,*]

1 Department of Physics, Instituto Tecnologico de Buenos Aires, CABA, Argentina

2 Genetic Medicine, The University of Chicago, Chicago, IL, USA

3 Committee on Molecular Metabolism and Nutrition, University of Chicago, Chicago, IL, USA

4 Departments of Biology and Computer Science, Loyola University Chicago, Chicago, IL, USA

5 Vanderbilt Genetic Institute, Vanderbilt University, Nashville, TN, USA

∗ **E-mail: Corresponding haky@uchicago.edu**

Abstract

To gain biological insight into the discoveries made by GWAS and meta-analysis studies, effective integration of functional data generated by large-scale efforts such as the GTEx Project is needed. PrediXcan is a gene-level approach that addresses this need by estimating the genetically determined component of gene expression. These predicted expression traits can then be tested for association with phenotype in order to test for mediating effects of gene expression levels. Furthermore, due to the polygenic nature of many complex traits, efforts to aggregate multiple GWAS studies and conduct meta-analyses have successfully increased our ability to identify variants of small effect sizes. To take advantage of the results generated by these efforts and to avoid the problems associated with accessing and handling individual-level data (e.g. consent limitations, large computational/storage costs) we have developed an extension of PrediXcan. The new method, MetaXcan, infers the results of PrediXcan using only summary statistics from meta-analyses of GWAS. Here we show that the concordance between PrediXcan and MetaXcan exceeds $R^2 > 0.90$ across reference populations. We demonstrate the power of this approach by computing gene-level association results for xx traits (N>100K) with 40 different tissue models and make the results accessible through `gene2pheno.org`. We also provide open source local and web-based software for easy implementation `https://github.com/hakyimlab/MetaXcan`

Introduction

Over the last decade, GWAS have been successful in identifying genetic loci that robustly associate with multiple complex traits. However, the mechanistic understanding of these discoveries is still limited, hampering the translation of this knowledge into actionable targets. Studies of enrichment of expression quantitative trait loci (eQTLs) among trait-associated variants [?, ?, ?] show the importance of gene expression regulation. Direct quantification of the contribution of different functional classes of genetic variants showed that 80% of phenotype variability (in 12 diseases) can be attributed to DNAase I hypersensitivity sites, further highlighting the importance of transcript regulation in determining phenotypes [?, ?].

Many transcriptome studies have been conducted where genotype and expression levels are assayed for a large number of individuals [?,?,?,?]. The most comprehensive transcriptome dataset, in terms of tissues covered, is the GTEx Project, a large-scale effort where DNA and RNA are collected from multiple tissue samples from nearly 1000 deceased individuals and sequenced to high coverage (cite GTEx paper `http://www.nature.com/ng/journal/v45/n6/full/ng.2653.html?WT.ec_id=NG-201306`). This remarkable resource provides a comprehensive cross-tissue survey of the functional consequences of genetic variation at the transcript level.

To integrate knowledge generated from these large-scale transcriptome studies and shed light on disease biology, we developed PrediXcan [?], a gene-level association approach that tests the mediating effects of gene expression levels on phenotypes. This is implemented on GWAS/sequencing studies (i.e. studies with genome-wide interrogation of DNA variation and phenotypes) where transcriptome levels are estimated with models trained in measured transcriptome datasets (e.g. GTEx). These predicted expression levels are then correlated with the phenotype and provides the basis for a gene-level association test that ameliorates some of the key limitations of GWAS (cite Predixcan paper again).

Other groups have also proposed methods based on similar ideas, cite Gusev et al, the person at ASHG??. Comparison with our method will be examined.

On the other hand, meta-analysis efforts that aggregate results from multiple GWAS studies have been able to identify an increasing number phenotype associations that were not detected with smaller sample sizes. In order to harness the power of these increased sample sizes while keeping the computational burden manageable, we have extended the PrediXcan method so that only summary statistics from meta-analysis studies are needed rather than genotype data.

We will show here that our new method, termed MetaXcan, is a fast, accurate, and efficient way to scale up implementation of PrediXcan and take advantage of the large sample sizes made available through meta-analysis of GWAS.

Results

We have derived an analytic expression that allows us to compute the outcome of PrediXcan using only summary statistics from genetic association studies. Details of the derivation are shown in the Methods section. In Figure **??**, we show the mechanics of MetaXcan in relation to traditional GWAS and our recently published PrediXcan method.

For both GWAS and PrediXcan, the input is the genotype matrix and phenotype vector. GWAS computes the regression coefficient of the phenotype on each marker in the genotype matrix and generates SNP-level results. PrediXcan starts by estimating the genetically-regulated component of the transcriptome (using weights from the publicly available PredictDB database) and then computes regression coefficients of the phenotype on each predicted gene expression level generating gene-level results. MetaXcan, on the other hand, is a shortcut that takes the output from GWAS and generates the output from PrediXcan. Since MetaXcan only uses summary statistics, it can effectively take advantage of large-scale meta analysis results, avoiding the computational and regulatory burden of handling large amounts of protected individual level data.

MetaXcan formula

Figure **??** shows the main analytic expression used by MetaXcan for the Z-score (effect size divided by its standard error) of the association between predicted gene expression and the phenotype. The input variables are the weights used to predict the expression of a given gene w_{lg}, the variance and covariances of the markers included in the prediction of the expression level of the gene, and the GWAS coefficient for each marker. To compute the last factor exactly, we would need some additional information, i.e. statistics on the association between the GWAS population and the phenotype. Since this information is unavailable in most cases, as a further approximation we are discarding this factor; we will later show

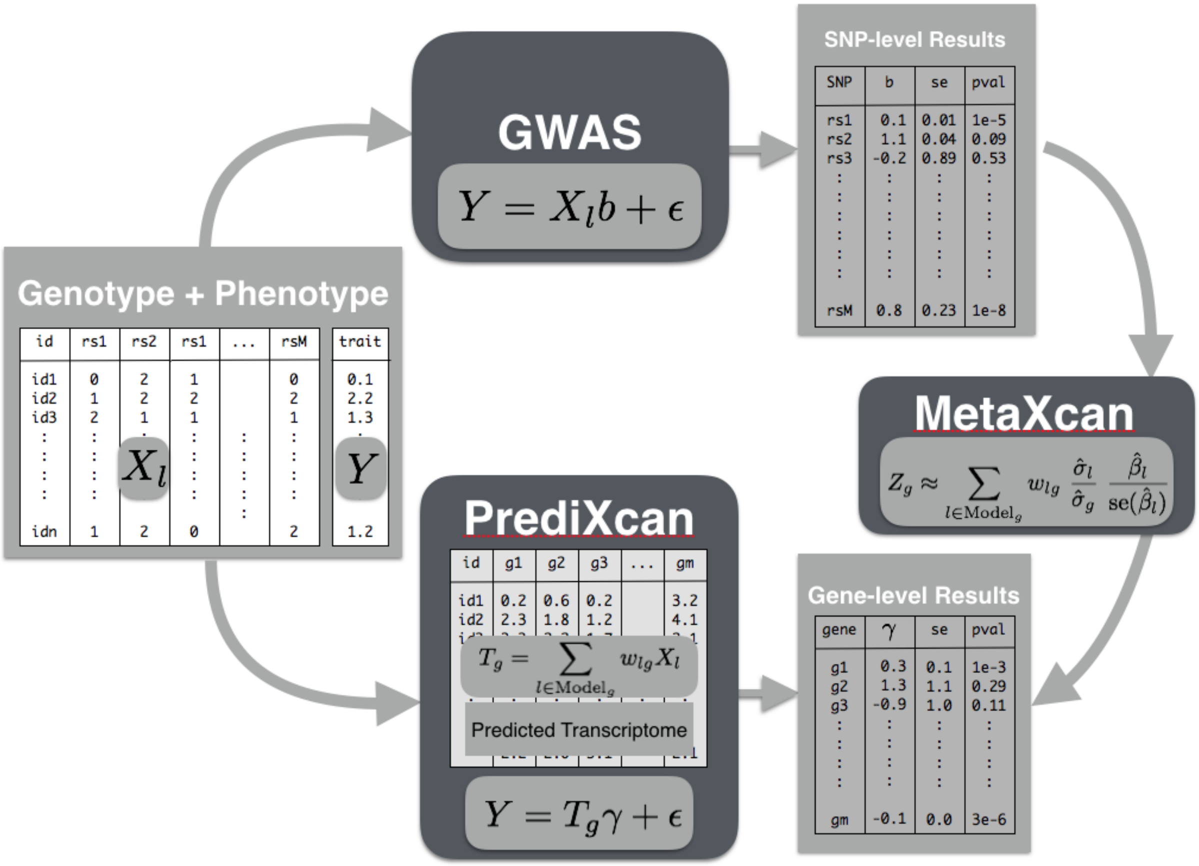


Figure 1. This figure illustrates the MetaXcan method in relationship to GWAS and PrediXcan. Both GWAS and PrediXcan take genotype and phenotype data as input. GWAS computes the regression coefficients of $Y \sim X_l$ using the model $Y = X_l b + \epsilon$, with Y being the phenotype and X_l individual dosage. The output is the table of SNP-levle results. PrediXcan, in contrast, starts first by computing the predicting/imputing the transcriptome. Then it calculates the regression coefficients of the phenotype Y on each gene's predicted expression T_g. The output is a table of gene-level results. MetaXcan computes the gene-level association results using directly the output from GWAS.

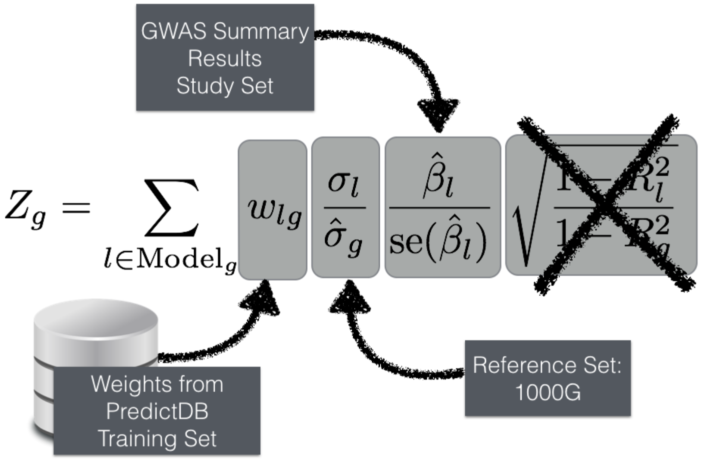

GWAS Summary Results Study Set
$Z_g = \sum_{l \in \text{Model}_g} w_{lg} \frac{\sigma_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)} \cancel{\sqrt{\frac{1-R_l^2}{1-R_g^2}}}$
Weights from PredictDB Training Set
Reference Set: 1000G

Figure 2. MetaXcan formula. This plot shows the formula to infer PrediXcan gene-level association results using summary statistics. The different sets involved in input data are shown. The study set is where the regression coefficient between the phenotype and the genotype is obtained. The training set is the reference transcriptome dataset where the prediction models of gene expression levels are trained. The reference set, in general 1000 Genomes, is used to compute the variances and covariances (LD structure) of the markers used in the predicted expression levels. Both the reference set and training set values are pre-computed and provided to the user so that only the study set results need to be provided to the software.

that dropping it does not affect the results in any appreciable way.

The approximate formula we will use is as follows:

$$Zg \approx \sum_{l \in \text{Model}_g} w_{lg} \, \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \, \frac{\hat{\beta}_l}{\text{se}(\beta_l)} \tag{1}$$

where

- w_{lg} is the weight of SNP l in the prediction of the expression of gene g,
- $\hat{\beta}_l$ is the GWAS regression coefficients for SNP l,
- $\text{se}(\beta_l)$ is standard error of $\hat{\beta}_l$,
- $\hat{\sigma}_l$ is the estimated variance of SNP l, and
- $\hat{\sigma}_g$ is the estimated variance of the predicted expression of gene g.

The inputs are based, in general, on data from three different sources:

- study set,
- training set, and
- population reference set.

The study set is the main dataset of interest from which the genotype and phenotypes of interest are gathered. The regression coefficients and standard errors are computed based on individual -level data from the study set. Training sets are the reference transcriptome datasets used for the training of the prediction models (GTEx, DGN, Framingham, etc.) thus the weights w_{lg} are computed from this set. Finally, the reference sets (e.g. 1000 Genomes) are used to derive variance and covariance (LD) properties of genetic markers, which will usually be different from the study sets.

In the most common use scenario, the user will only need to provide GWAS results using his/her study set. The remaining parameters are pre-computed and hosted in the PredictDB database predictdb url here.

Next we will show the performance of the method, measured as the concordance (R^2) between PrediXcan and MetaXcan results.

Performance in simulated data

We first compared MetaXcan and PrediXcan using simulated phenotypes. For genotypes we use three ancestral subsets of the 1000 Genomes project: Africans (n=662), East Asians (n=504), and Europeans (n=503). Each set was used as reference and study sets yielding a total of 9 combinations as shown in Figure **??**. For each population combination, we computed PrediXcan association results for the simulated phenotype and compared them with results generated from our MetaXcan approach. This allowed us to assess the effect of ancestral differences between study and reference sets.

The z-scores of MetaXcan for each combination of study and reference sets are displayed in Figure **??**, in a scatter plot against their corresponding PrediXcan result. As expected, when the study and reference sets are the same, the concordance between MetaXcan and PrediXcan is 100% whereas when the sets are of different ancestral origin the R^2 drops a few percentage points, with the biggest loss (down to 87%) when the study set is Asian and the reference set is African. This confirmed that our formula works as expected and that the approach is robust to ethnic differences between study and reference sets.

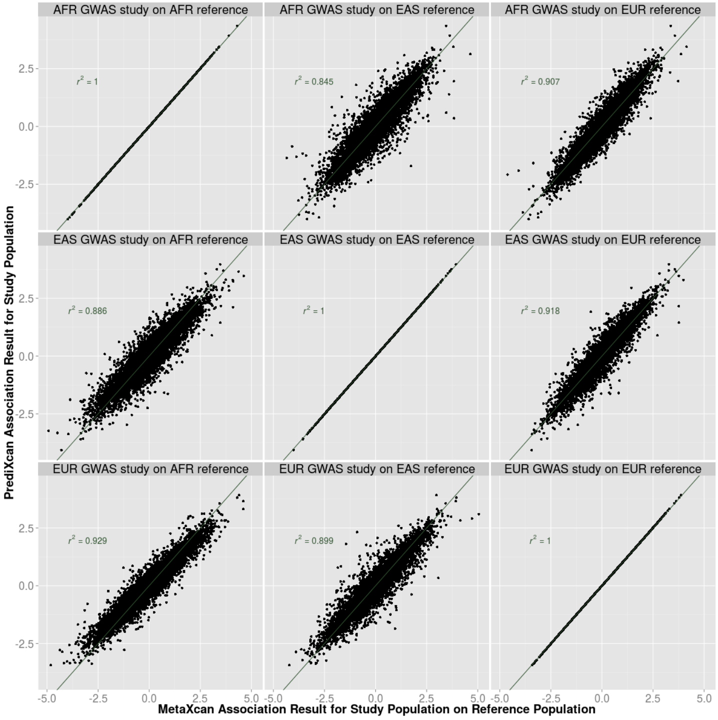


Figure 3. Comparison of PrediXcan and MetaXcan results for a simulated phenotype. Study populations and MetaXcan reference populations were built from European, African, and Asian individuals from the 1000 Genomes Project. Gene Expression model was based on Depression Genes and Networks.

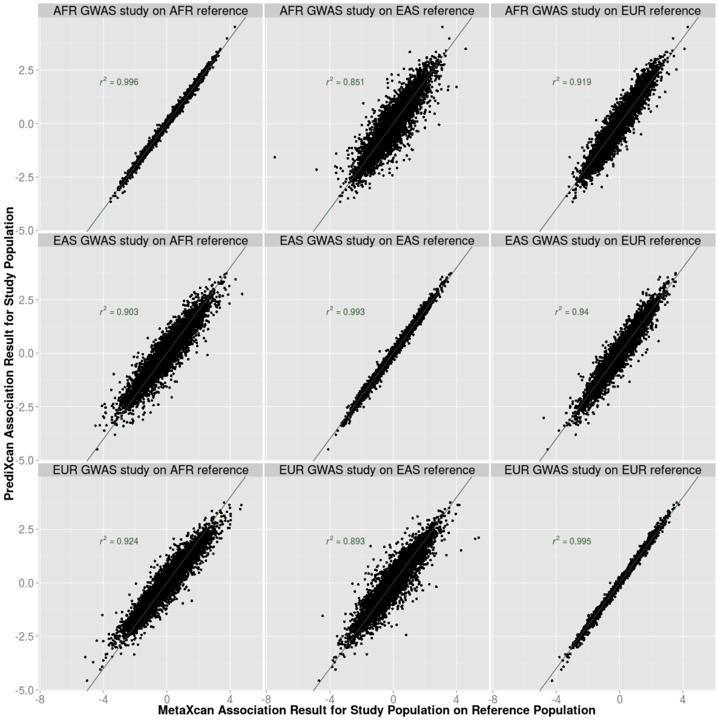


Figure 4. Comparison of PrediXcan and MetaXcan results for a cellular phenotype, intrinsic growth. Study sets and MetaXcan reference sets consisted of European, African, and Asian individuals from the 1000 Genomes Project. Gene Expression model was based on Depression Genes and Networks.

Performance in cellular growth phenotype from 1000 genomes cell lines

Intrinsic growth, a cellular phenotype, is available for XXX cell lines from the 1000 Genomes project. These values are available for the following subsets of Thousand Genomes: European (EUR), African (AFR), Asian (EAS).

We compared z-scores for intrinsic growth generated by PrediXcan and MetaXcan for different combinations of reference and study sets. The results are shown in Figure **??**. Consistent with our simulation study, the MetaXcan results closely match the PrediXcan results. Again, the best concordance occurs when reference and study sets share similar continental ancestry while differences in population slightly reduce concordance. Compared to the plots for the simulated phenotypes, the diagonal concordance is slightly lower than 1; this is due to the fact that more individuals were included in the reference set than in the study set, such that the sets were not identical.

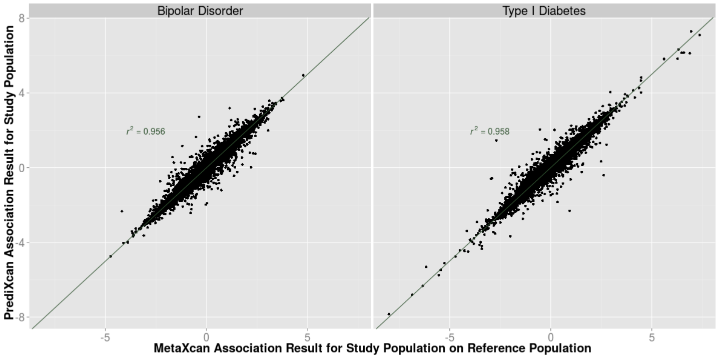


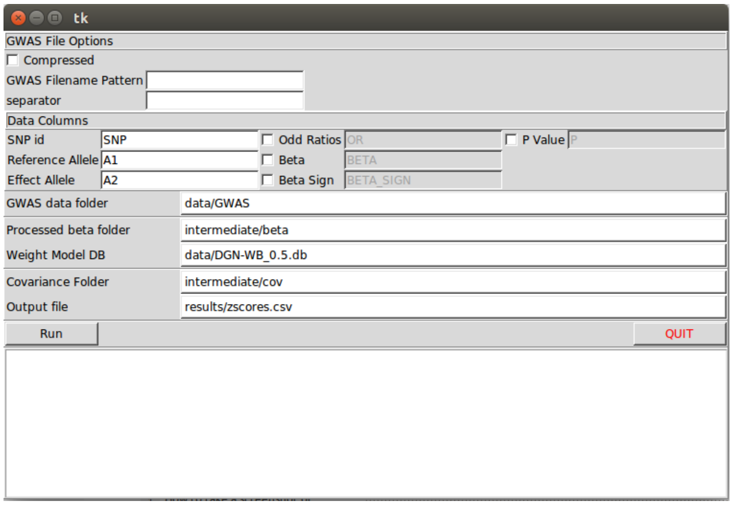
Figure 5. Comparison of PrediXcan results and MetaXcan results for a Type I Diabetes study, and a Biploar Disorder study. Study data was extracted from Wellcome Trust Case Control Consortium, and MetaXcan reference population were the European individuals from Thousand Genomes Project (same as in previous sections)

Performance on disease phenotypes from WTCCC

We show the comparison of MetaXcan and PrediXcan results for two diseases, Bipolar Disorder (BD) and Type 1 Diabetes (T1D) from the WTCCC in Figure **??**. Other disease phenotypes exhibited similar performance (data not shown). Concordance between MetaXcan and PrediXcan is over 99% in for both diseases (BD $R^2 = 0.996$ and T1D $R^2 = 0.995$). The very small differences are explained by differences in allele frequencies and LD between the reference set (1000 Genomes) and the study set (WTCCC). Given this high concordance, we do not expect much improvement when using a reference set that is more similar to the study set. We verified this and, as expected, found that using control individuals from WTCCC as reference set improved the concordance only marginally (0.1%). It is worth noting that the PrediXcan results for diseases were obtained using logistic regression whereas MetaXcan formula is based on linear regression properties. As observed before [**?**, **?**], when the number of cases and controls are relatively well balanced (roughly, at least 25% of cases and controls), linear regression approximation yields very similar results to logistic regression.

Application to large-scale meta analysis results

Finally, we apply MetaXcan to a number of publicly available large consortia meta analysis results with sample sizes exceeding the 100K individuals. The full list of consortia results used are listed on table Txxx. Using gene expression prediction models for 40 different tissues, we performed MetaXcan association. As expected, we find that the top hits are enriched for known disease/trait associated genes.



☐ Compressed
GWAS Filename Pattern
separator
SNP id: SNP
Reference Allele: A1
Effect Allele: A2
☐ Odd Ratios: OR
☐ Beta: BETA
☐ Beta Sign: BETA_SIGN
☐ P Value: P
GWAS data folder: data/GWAS
Processed beta folder: intermediate/beta
Weight Model DB: data/DGN-WB_0.5.db
Covariance Folder: intermediate/cov
Output file: results/zscores.csv

Figure 6. GUI application running in a local environment

We make the results available through on `gene2pheno.org`, which should be useful to the community for replication, validation, or as functional annotations of genes. Disease/trait focused analysis of these results are currently ongoing to fully leverage these results and further our biological understanding of the traits.

Software

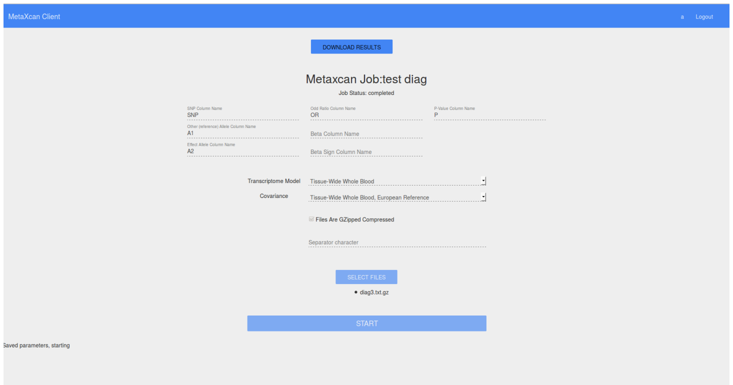
We make our software publicly available on `https://github.com/hakyimlab/MetaXcan`. The weights and covariances for different tissues can be downloaded from ... A short working example can be found on the github page.

For ease of use, we have developed a desktop GUI application, shown in Figure **??**.

We have also developed a web-based version to further simplify the use of the method (Figure **??**) . url:

Discussion

Here we present MetaXcan, a scalable, accurate, and efficient method to integrate reference transcriptome studies to learn about the biology of complex traits and diseases. Our method extends PrediXcan, which maps genes to phenotypes by testing the mediating effects of gene expression levels. This is implemented by predicting gene expression levels and correlating these traits with phenotypes. MetaXcan is a shortcut



Metaxcan Job:test diag

Job Status: completed

SNP Column Name: SNP

Odd Ratio Column Name: OR

P-Value Column Name: P

Other (reference) Allele Column Name: A1

Beta Column Name

Effect Allele Column Name: A2

Beta Sign Column Name

Transcriptome Model: Tissue-Wide Whole Blood

Covariance: Tissue-Wide Whole Blood, European Reference

☑ Files Are GZipped Compressed

Separator character

- diag3.txt.gz

Saved parameters, starting

Figure 7. Web Based version

that uses SNP-level association results and combines them to reproduce the results of PrediXcan, without the need to use individual level data.

MetaXcan shares most of the benefits of PrediXcan: a) it directly tests the regulatory mechanism through which genetic variants affect phenotype; b) it provides gene-level results which are better functionally characterized than genetic variants, easier to validate within model systems, and carry a smaller multiple testing burden; c) the direction of the effects are known, facilitating identification of therapeutic targets; d) reverse causality is largely avoided since predicted expression levels are based on germline variation, which are not affected by onset of disease; e) it can be systematically applied to existing GWAS studies; f) tissue-specific analysis can be performed using all the models we have made available on PredictDB, with over 40 tissues available.

The difference between the reference sets (used to estimate LD and allele frequencies) and study set (used to compute GWAS/meta analysis summary statistics) is the main cause of the small differences between MetaXcan and PrediXcan results. We have shown here that even when the populations are quite different, the concordance is very high. Thus, MetaXcan is robust to ancestral differences between study and reference sets.

Even though the method was derived with linear regression in mind, in case-control designs, the approximation generates results that are in almost full concordance with exact results generated with PrediXcan and logistic regression.

Methods similar in spirit to PrediXcan have been reported cite gusev/bogdan, ASHG guy?. Gusev et al also propose a method comparable to MetaXcan that is based only on summary statistics. Their method called Transcriptome-Wide Association Study (TWAS) imputes the SNP level z-scores into gene level z-scores using the method Pasaniuc and others have published [?]. In contrast, MetaXcan infers the results of PrediXcan using summary statistics through an analytically derived formula (see next section for the derivation details). The formula from both methods are similar but MetaXcan is more general since it allows any weighting scheme for the prediction of expression level, i.e. not limited to one imputation scheme. Also, we have employed transcriptome models built with Elastic Net and a mixing parameter $alpha = 0.5$ in thsi particular application, whereas the other approach employed a model equivalent to Ridge Regression, which has been shown to be inferior in performance [?].

In summary, we present an accurate and computationally efficient gene-level association method that integrates functional information from reference transcriptome dataset into GWAS and large scale meta-analysis results to inform the biology of complex traits.

Methods

Derivation of MetaXcan Formula

The goal of MetaXcan is to infer the results of PrediXcan using only GWAS summary statistics. Individual level data are not needed for this algorithm. We will define some notations for the derivation of the analytic expressions of MetaXcan.

Notation

Y is the n-dimensional vector of phenotype for individuals $i = 1, n$.
X_l is the allelic dosage for SNP l.
T_g is the predicted expression (or estimated GREx, genetically regulated expression) for gene g.
We model the phenotype as linear functions of X_l and T_g

$$Y = X_l \beta_l + \eta$$

$$Y = T_g \gamma_g + \epsilon$$

$\hat{\gamma}_g$ and $\hat{\beta}_l$ are the estimated regression coefficients of Y regressed on T_g and X_l, respectively. $\hat{\gamma}_g$ is the result (effect size for gene g) we get from PrediXcan whereas $\hat{\beta}_l$ is the result from a GWAS for SNP l.

We will denote as Var and Cov the operators that computes the sample variance and covariances, i.e.

$\mathrm{Var}(Y) = \sum_{i=1,n}(Y_i - \bar{Y})^2/n$ with $\bar{Y} = \sum_{i=1,n} Y_i/n$

$\hat{\sigma}_l^2 = \mathrm{Var}(X_l)$

$\hat{\sigma}_g^2 = \mathrm{Var}(T_g)$

$\hat{\sigma}_Y^2 = \mathrm{Var}(Y)$

$\Gamma_g = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})/n$ where $\mathbf{X}'$ is the $n \times p$ matrix of SNP data and $\bar{\mathbf{X}}$ is a $n \times p$ matrix where column l has the column mean of $\mathbf{X}_l$ (p being the number of SNPS in the model for gene g).

With these notations, our goal is to infer PrediXcan results ($\hat{\gamma}_g$ and its standard error) using only GWAS results (β_l and se), estimated variances of SNPs ($\hat{\sigma}_l^2$), covariances between SNPs in each gene model (Γ_g), and prediction model weights w_{lg}.

Input: β_l, $\mathrm{se}(\beta_l)$, $\hat{\sigma}_l^2$, Γ_g, w_{lg}. **Output:** $\hat{\gamma}_g$, $\mathrm{se}(\hat{\gamma}_g)$.

Next we list the properties and definitions used in the derivation:

$$\hat{\gamma}_g = \frac{\mathrm{Cov}(T_g, Y)}{\mathrm{Var}(T_g)} = \frac{\mathrm{Cov}(T_g, Y)}{\hat{\sigma}_g^2} \tag{2}$$

and

$$\hat{\beta}_l = \frac{\mathrm{Cov}(X_l, Y)}{\mathrm{Var}(X_l)} = \frac{\mathrm{Cov}(X_l, Y)}{\hat{\sigma}_l^2} \tag{3}$$

The proportion of variance explained by the covariate (T_g or X_l) can be expressed as

$$R_g^2 = \hat{\gamma}_g^2 \; \frac{\hat{\sigma}_g^2}{\hat{\sigma}_Y^2}$$

$$R_l^2 = \hat{\gamma}_l^2 \; \frac{\hat{\sigma}_l^2}{\hat{\sigma}_Y^2}$$

By definition

$$T_g = \sum_{l \in \mathrm{Model}_g} w_{lg} X_l \tag{4}$$

The $\mathrm{Var}(T_g) = \hat{\sigma}_g^2$ can be computed as

$$\begin{aligned}
\hat{\sigma}_g^2 &= \mathrm{Var}\left(\sum_{l \in \mathrm{Model}_g} w_{lg} X_l\right) \\
&= \mathrm{Var}(\mathbf{W}_g \mathbf{X}_g) && \text{where } \mathbf{W}_g \text{is the vector of } w_{lg} \text{for SNPs in the model of } g \\
&= \mathbf{W}_g' \mathrm{Var}(\mathbf{X}_g) \mathbf{W}_g && \text{where } \Gamma_g \text{ is the} \mathrm{Var}(\mathbf{X}_g) = \text{ covariance matrix of } \mathbf{X}_g \\
&= \mathbf{W}_g' \Gamma_g \mathbf{W}_g && (5)
\end{aligned}$$

Calculation of regression coefficient γ_g

$\hat{\gamma}_g$ can be expressed as

$$\begin{aligned}
\hat{\gamma}_g &= \frac{\mathrm{Cov}(T_g, Y)}{\hat{\sigma}_g^2} \\
&= \frac{\mathrm{Cov}(\sum_{l \in \mathrm{Model}_g} w_{lg} X_l, Y)}{\hat{\sigma}_g^2} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \mathrm{Cov}(X_l, Y)}{\hat{\sigma}_g^2} && \text{by linearity of Cov} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \hat{\beta}_l \sigma_l^2}{\hat{\sigma}_g^2} && \text{using Eq } \textbf{??} \qquad (6)
\end{aligned}$$

Calculation of standard error of γ_g

Also from the properties of linear regression we know that

$$\mathrm{se}(\hat{\gamma}_g) = \sqrt{\mathrm{Var}(\hat{\gamma}_g)} = \frac{\hat{\sigma}_\epsilon}{\sqrt{n \hat{\sigma}_g^2}} = \frac{\hat{\sigma}_Y^2 (1 - R_g^2)}{n \hat{\sigma}_g^2} \qquad (7)$$

In this equation, σ_Y / n is not necessarily known but can be estimated using the analogous equation (**??**) for beta

$$\mathrm{se}(\hat{\beta}_l) = \frac{\hat{\sigma}_Y^2 (1 - R_l^2)}{n \hat{\sigma}_l^2} \qquad (8)$$

Thus

$$\frac{\hat{\sigma}_Y^2}{n} = \frac{\mathrm{se}(\hat{\beta}_l)^2 \hat{\sigma}_l^2}{(1 - R_l^2)} \qquad (9)$$

Notice that the right hand side of (**??**) is dependent on the SNP l while the left hand side is not. This equality will only approximately in our implementation since we will be using approximate values for $\hat{\sigma}_l^2$, i.e. from reference population, not the actual study population.

Calculation of Z score

To assess the significance of the association, we need to compute the ratio of the effect size γ_g and standard error $\mathrm{se}(\gamma_g)$, or Z score,

$$Z_g = \frac{\hat{\gamma}_g}{\mathrm{se}(\hat{\gamma}_g)} \tag{10}$$

with which we can compute the p value as

$$p = 2 \ \mathrm{pnorm}(-|Z_g|) \tag{11}$$

$$\begin{aligned}
Zg &= \frac{\hat{\gamma}_g}{\mathrm{se}(\hat{\gamma}_g)} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \hat{\beta}_l \sigma_l^2}{\hat{\sigma}_g^2} \sqrt{\frac{n}{\hat{\sigma}_Y^2} \frac{\hat{\sigma}_g^2}{(1 - R_g^2)}} && \text{using Eq. } \textbf{??} \text{ and } \textbf{??} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \hat{\beta}_l \sigma_l^2}{\hat{\sigma}_g} \sqrt{\frac{(1 - R_l^2)}{\mathrm{se}(\hat{\beta}_l)^2 \hat{\sigma}_l^2}} \sqrt{\frac{1}{(1 - R_g^2)}} \\
&= \sum_{l \in \mathrm{Model}_g} w_{lg} \ \frac{\sigma_l}{\hat{\sigma}_g} \ \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} \ \sqrt{\frac{1 - R_l^2}{1 - R_g^2}} && (12) \\
&\approx \sum_{l \in \mathrm{Model}_g} w_{lg} \ \frac{\sigma_l}{\hat{\sigma}_g} \ \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} && (13)
\end{aligned}$$

Based on results with actual and simulated data we have found that the last approximation does reduce power since the deviation is only noticeable when the correlation between the SNP or the predicted expression and the phenotype is large, i.e. large effect sizes. When the effects are large enough for the approximation to impact the association, the loss of power is compensated by the large effect size.

Acknowledgments

References

References

1. Alexander Gusev, *Integrative approaches for large-scale transcriptome-wide association studies*, 1994.