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

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

1 Dept?, ITBA?, Buenos Aires, Argentina

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

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

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

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

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

Abstract

Effective integration of functional data such as the ones generated by GTEx and other efforts are needed to gain biological insights from the discoveries made by GWAS and meta analysis studies. PrediXcan is a gene-level approach that addresses this need. It consists of estimating the genetically determined levels of gene expression and correlating these with phenotype to test the mediating effects of a gene. In addition, due the polygenic nature of many complex traits, meta analysis efforts have formed and successfully aggregated multiple GWAS studies increasing our ability to identify variants of smaller 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 (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 traditional GWAS/meta analysis output. Here we show that the concordance between PrediXcan and MetaXcan is excellent and robust to differences between reference –used to compute the LD structure– and study populations. Thus, MetaXcan is a scalable, accurate and efficient gene-level association test well suited for application to ever increasing sample sizes.

We provide open source local and web-based software for easy implementation `https://github.com/hakyimlab/MetaXcan`

Introduction

Over the last decade, GWAS have been succesful in identifying genetic loci that are robustly associated with multiple complex traits. However, the mechanistic understanding of these discoveries is still limited hampering the translation of this knowledge into actionable targets. Studies on enrichment of expression trait loci, eQTL, among trait-associated variants [**?**, **?**] show the importance of gene expression regulation in this missing link. 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 reference transcriptome studies have been conducted where genotype and expression levels are assayed for a large number of individuals []. The most comprehensive in terms of tissues covered is the GTEx project, a large scale effort where multiple tissue samples from nearly 1000 deceased individuals are being collected and their DNA and RNA sequenced to high coverage. This remarkable resource provides a comprehensive cross tissue survey of 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 have proposed PrediXcan [**?**], a gene-level association approach that test the mediating effects of gene expression levels on phenotypes. This is implemented in GWAS/sequencing studies by predicting the transcriptome levels with models trained in reference transcriptome datasets. These predicted expression levels are correlated with the phenotype with the idea that causal genes will tend to be significantly associated.

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 genetic markers 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 PrediXcan so that only summary statistics, i.e. results from meta analysis studies are needed.

We will show here that the new method termed MetaXcan is a fast, accurate, and efficient way to scale up implementation of PrediXcan to the large sample sizes used in meta analysis studies.

Results

We have derived an analytic expression that allows us to compute the outcome of PrediXcan using only summary statistics, i.e. the outcome of GWAS. Details of the derivation are shown in the Methods section. In Figure 1, we show the mechanics of MetaXcan in relation to traditional GWAS and our recently published PrediXcan methods.

For both GWAS and PrediXcan, the input is the genotype matrix and phenotype vector, Y. GWAS computes the regression coefficient of Y on each marker in the genotype matrix and generates SNP-level results. PrediXcan, imputes the transcriptome using models from the publicly available PredictDB database, computes the regression coefficient of Y on each predicted expression level, T, and generates gene-level results. MetaXcan is a shortcut method that takes the output from GWAS and generates the output from PrediXcan.

the genotype matrix and weights our publicly available PredictDB database are used to compute the predicted transcriptome and gene-level association results between the predicted transcript levels and the phenotype are reported. MetaXcan is a shortcut approach that uses

In a nutshell, MetaXcan computes the results of PrediXcan using results from GWAS or from meta analysis of GWAS.

MetaXcan formula

Figure 2 shows the main analytic expression used by MetaXcan, i.e. the Zscore (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, the GWAS result for each marker. To be exact we would need some additional information to compute the last factor but we will show later that dropping it does not affect the results in any substantive way.

The approximate formula we will use is as follows:

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

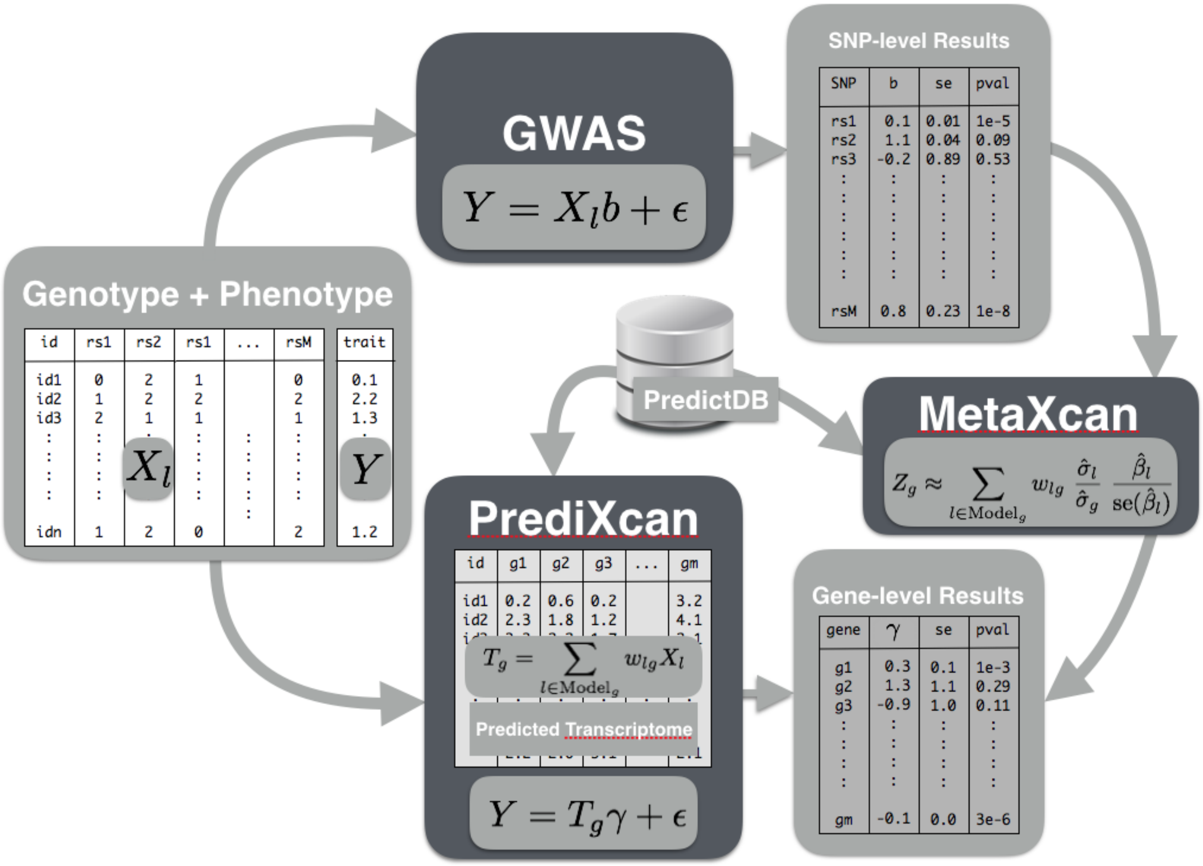

Figure 1. caption here

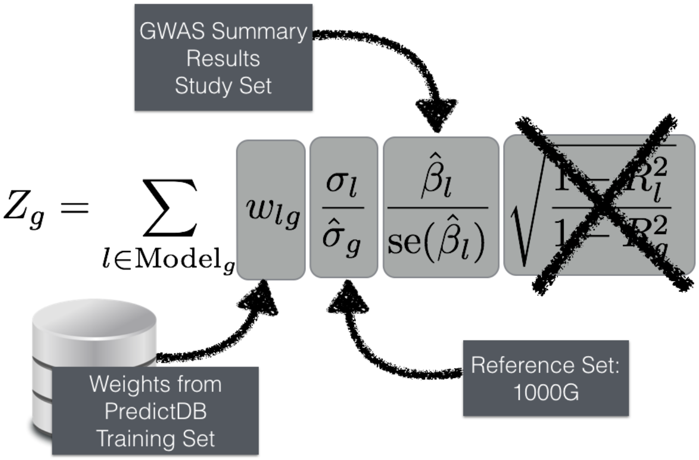
GWAS Summary Results Study Set
$Z_g = \sum_{l \in \text{Model}_g} w_{lg} \frac{\sigma_l}{\hat\sigma_g} \frac{\hat\beta_l}{\text{se}(\hat\beta_l)} \cancel{\sqrt{\frac{1-R_l^2}{1-R_g^2}}}$
Weights from PredictDB Training Set
Reference Set: 1000G

Figure 2. caption here

where

- w_{lg}: weight of SNP l for gene g
- $\hat{\beta}_l$: GWAS regression coefficients for each SNP used to predict the gene's expression
- se(β_l): standard error of $\hat{\beta}_l$
- σ_l: variance of SNP l
- $\hat{\sigma}_g$: variance of the genetic model for gene g

The inputs are based, in general, on data from three different sources: study set, training set, and reference set. Study set is the main dataset of interest and the one where 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 datasets used for the training of the prediction models (GTEx, DGN, Framingham) thus the weights w_{lg} are computed in this set. Finally reference sets (e.g. 1000 Genomes or other publicly available sets) 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 case 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

To test the performance we ran MetaXcan and PrediXcan using simulated phenotypes. For genotypes we use three subsets of the 1000 Genomes project: one with Africans individuals (n=662), one with East Asians (n=504), and one by Europeans (n=503). Each set was used as reference and study sets yielding a total of 9 combinations as shown in Figure 3.

PrediXcan association results were obtained for the simulated phenotype at each of these population subsets, and compared against MetaXcan association results for different population combinations. This allowed us to assess the effect of ethnic differences between study set and reference set.

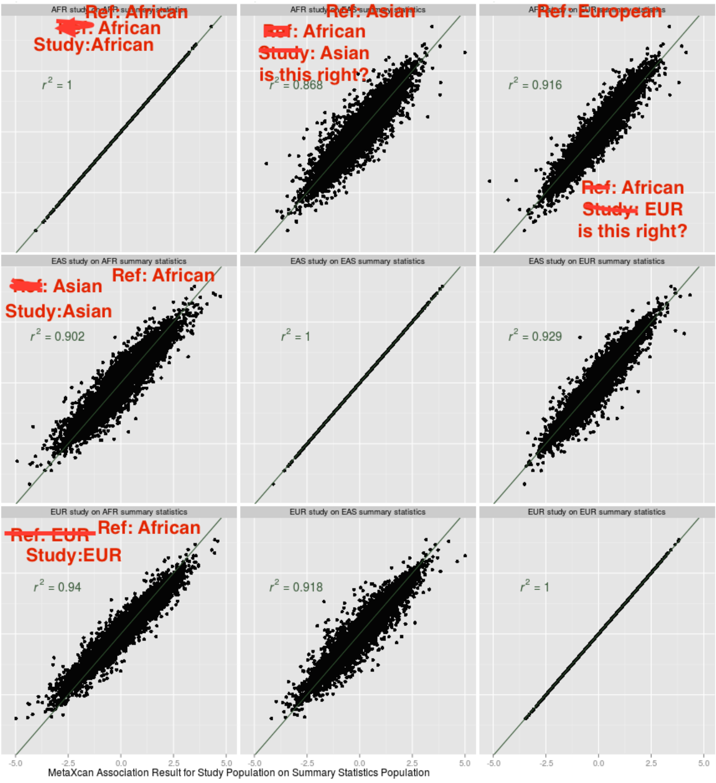

Figure 3. Comparison of PrediXcan and MetaXcan results for a synthetic phenotype. Study populations and MetaXcan reference populations were built from European, African, and Asian individuals from the 1000 Genomes Project. Gene Expression model was based on Depression Genes and Networks.

The z-scores of MetaXcan for each combination of study and reference sets is displayed in Figure 3, 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 ethnic origin the R^2 drop 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. Next, we examined the performance using a real phenotype of cell lines from the 1000 genomes project.

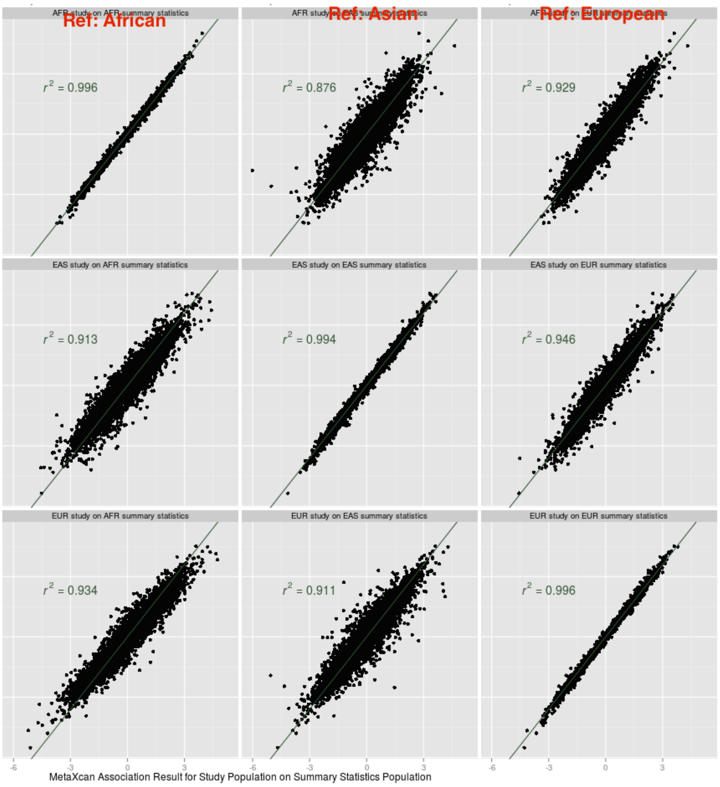

Figure 4. Comparison of PrediXcan and MetaXcan results for a Intrinsic Growth phenotype. Study populations and MetaXcan reference populations were built from European, African, and Asian individuals from the 1000 Genomes Project. Gene Expression model was based on Depression Genes and Networks. Dot color accounts for number of SNPS in each gene.TODO (WTCCC)

Performance in cellular growth phenotype from 1000 genomes cell lines

The intrinsic growth, a cellular phenotype, is available for XXX cell lines from the 1000 Genomes project. These values are available for the following ethnic groups: European (CEU), African (YRI), Asian (CHB, JPT), African American (ASW?).

As for the simulated phenotype, we compared MetaXcan vs PrediXcan for different combinations of reference and study sets. The results are shown in Figure4. MetaXcan results closely match PrediXcan results with best concordance when the reference and study sets are from the same ethnicity while differences in ethnicities slitghly 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 a few more individuals were included in the reference set than in the study set.

Next, we assessed the performance using disease phenotypes from the WTCCC.

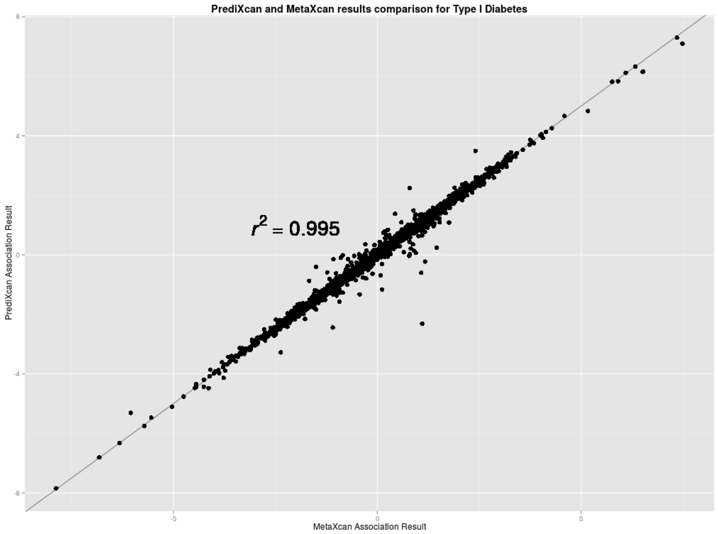

Figure 5. Comparison of PrediXcan results and MetaXcan results for a Type I Diabetes study. Diabetes study data was extracted from Wellcome Trust Case Control Consortium, and MetaXcan reference population were the European individuals from Thousand Genomes Project (same as in previous sections)

Performance on disease phenotypes from WTCCC

Here we show the comparison for two diseases, Bipolar Disorder (BD) and Type 1 Diabetes (T1D) from the WTCCC. These comparisons are displayed in Figures 5 and 6. Other disease phenotypes showed similar performance (not shown).

As expected concordance between MetaXcan and PrediXcan is well over 99% (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 change 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%).

TODO: make one figure with both T1D and Bipolar Disease

Application to large-scale meta analysis results

We downloaded the publicly available meta analysis results from multiple consortia listed on table Txxx. Using gene expression prediction models for 39 different tissues, we performed MetaXcan association and

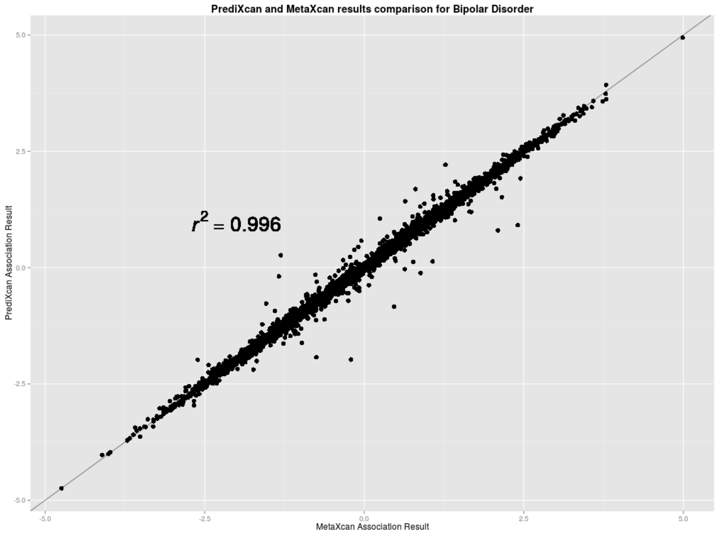

Figure 6. Comparison of PrediXcan results and MetaXcan results for a Bipolar Disorder study. Bipolar Disorder study data was extracted from Wellcome Trust Case Control Consortium, and MetaXcan reference population were the European individuals from Thousand Genomes Project (same as in previous sections)

make the results available through on `gene2pheno.org`.

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.

TODO: include GUI screen shot

We have also developed a web-based version to further simplify the use of the method. 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/imputing expression levels of genes and correlating the predicted expression levels with phenotypes. MetaXcan is a shortcut to PrediXcan that uses SNP level association results and combines them to reproduce the results of PrediXcan, without the need to use individual level data. Roughly speaking, MetaXcan flips the association and prediction steps taking advantage of the (almost) linearity of them.

MetaXcan shares most of the benefits of PrediXcan: a) it directly tests the molecular regulatory mechanism through which genetic variant affect phenotype, b) it provides gene-level results, which are much better functionally characterized the genetic variants, they are easier to validate with other model systems using orthology, multiple testing burden is substantially reduced; 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) drives the concordance between MetaXcan and PrediXcan. We have shown here that even when the populations are quite different, the reduction in concordance is small. Thus MetaXcan is robust to ethnic differences between study and reference sets.

Even thought the method was derived with linear regression in mind, in the case of 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 Zscores into gene level Zscores using the method Pasaniuc and others have published [**?**]. In contrast, MetaXcan infers the results of PrediXcan using summary statistics with an analytic formula. 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. From what we are finding about the genetic architecture of gene expression traits, we have evidence that indicates that the imputation scheme used in TWAS (the summary statistics one) is sub-optimal []. Gusev et al's summary stat method is equivalent to kriging the Zscores, thus the performance is tied to the predictive performance of kriging gene expression traits. Notice that kriging and ridge regression are equivalent in terms of prediction. Since we are finding that ridge regression (mixing parameter = 1 in Elastic net) has lower performance consistently across all genes, we can expect that the imputation of Zscore approach will perform worse than MetaXcan, if we use elastic net weights, for example.

Note: Here we use GWAS in a more general sense covering common and rare variation association study where a dense genome-wide interrogation of genetic variation is association with phenotypes of interest.

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).
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 a gene's model).

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

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

Calculation of regression coefficient γ_g

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

$$\begin{aligned}
\hat{\gamma}_g &= \frac{\mathrm{Cov}(T_g, Y)}{\hat{\sigma}_g^2} \\
&= \frac{\mathrm{Cov}(\sum_{l \in \mathrm{Model}_g} w_{lg} X_l, Y)}{\hat{\sigma}_g^2} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \mathrm{Cov}(X_l, Y)}{\hat{\sigma}_g^2} && \text{by linearity of Cov} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \hat{\beta}_l \sigma_l^2}{\hat{\sigma}_g^2} && \text{using Eq 3} \qquad (6)
\end{aligned}$$

Calculation of standard error of γ_g

Also from the properties of linear regression we know that

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

In this equation, σ_Y / n is not necessarily known but can be estimated using the analogous equation (7) for beta

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

Thus

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

Notice that the right hand side of (9) 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

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

$$\begin{aligned}
Zg &= \frac{\hat{\gamma}_g}{\mathrm{se}(\hat{\gamma}_g)} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \hat{\beta}_l \sigma_l^2}{\hat{\sigma}_g^2} \sqrt{\frac{n}{\hat{\sigma}_Y^2} \frac{\hat{\sigma}_g^2}{(1 - R_g^2)}} && \text{using Eq. 6 and 7} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \hat{\beta}_l \sigma_l^2}{\hat{\sigma}_g} \sqrt{\frac{(1 - R_l^2)}{\mathrm{se}(\hat{\beta}_l)^2 \hat{\sigma}_l^2}} \sqrt{\frac{1}{(1 - R_g^2)}} \\
&= \sum_{l \in \mathrm{Model}_g} w_{lg} \, \frac{\sigma_l}{\hat{\sigma}_g} \, \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} \, \sqrt{\frac{1 - R_l^2}{1 - R_g^2}} && (12) \\
&\approx \sum_{l \in \mathrm{Model}_g} w_{lg} \, \frac{\sigma_l}{\hat{\sigma}_g} \, \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} && (13)
\end{aligned}$$

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 the affect the approximation, 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.