bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MetaXcan: Summary Statistics Based Gene-Level Association Method Infers Accurate PrediXcan Results

Alvaro Barbeira[1], Kaanan P. Shah[2], Jason M. Torres[3], Heather E Wheeler[4], Eric S. Torstenson[5], Todd Edwards[5], Tzintzuni Garcia[6], Graeme I Bell[7], Dan Nicolae[1], Nancy J Cox[5], Hae Kyung Im[2,*]

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

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

3 Committee on Molecular Metabolism and Nutrition, The 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

6 Center for Research Informatics, The University of Chicago, IL, USA

7 Section of Endocrinology, The University of Chicago, Chicago, IL, 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 role 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 large-scale GWAS or meta-analyses. Here we show that the concordance between PrediXcan and MetaXcan is excellent when the right reference population is used ($R^2 > 0.95$) and robust to population mismatches ($R^2 > 0.85$). We provide open source local and web-based software for easy implementation

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

through `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 [1,2] 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 [3].

Many transcriptome studies have been conducted where genotype and expression levels are assayed for a large number of individuals [4–7]. 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 [8]. 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 [9], 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 imputed 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 addresses some of the key limitations of GWAS [9].

Other groups have also proposed methods based on similar ideas [10]. Comparison with our method will be discussed.

On the other hand, meta-analysis efforts that aggregate results from multiple GWAS studies have been able to identify an increasing number of 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-

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

analysis studies are needed rather than individual level genotype and phenotype 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 1, we illustrate 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, can be viewed as a shortcut that uses the output from a GWAS study to generate the output from PrediXcan. Since MetaXcan only depends 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 2 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. The last factor in the formula can be computed exactly in principle, but we would need some additional information that is unavailable in typical GWAS output. Fortunately, we have found that this factor is very close to 1 and dropping it from the formula does not affect the accuracy of the

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

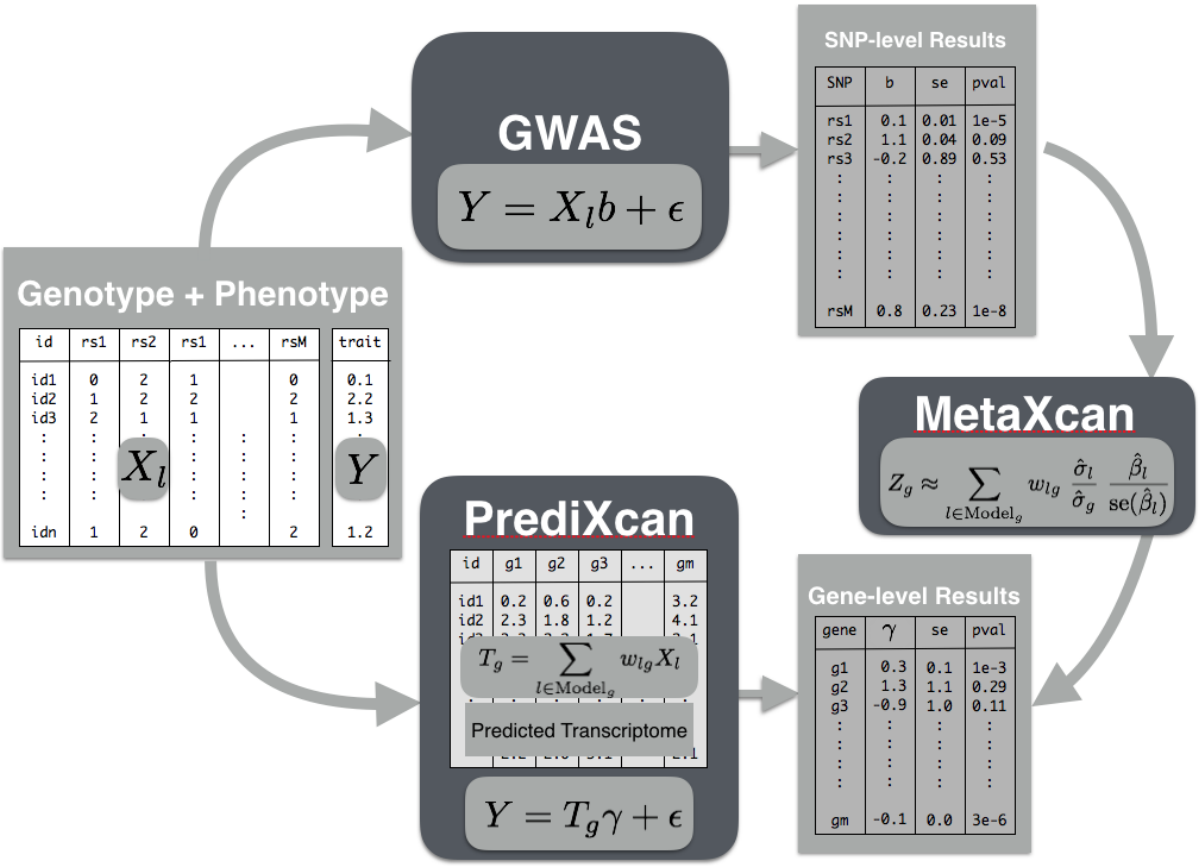

Figure 1. This figure illustrates the MetaXcan method in relationship to GWAS and PrediXcan. Both GWAS and PrediXcan take genotype and phenotype data as input. GWAS computes the regression coefficients of $Y \sim X_l$ using the model $Y = X_l b + \epsilon$, where Y is the phenotype and X_l the individual dosage. The output is the table of SNP-level results. PrediXcan, in contrast, starts first by 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.

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

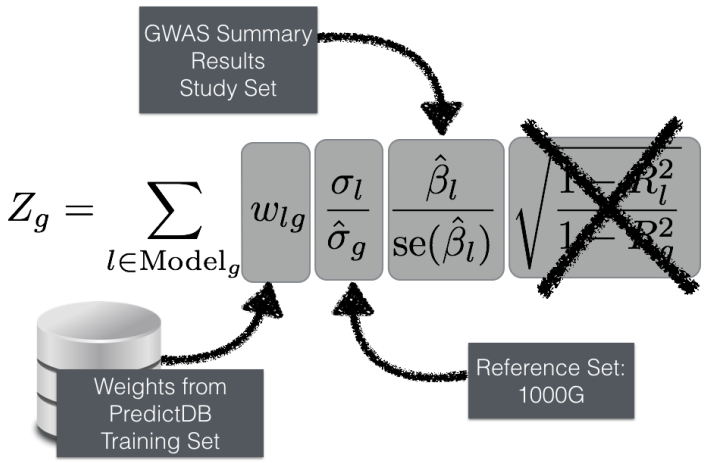

Figure 2. MetaXcan formula. This plot shows the formula to infer PrediXcan gene-level association results using summary statistics. The different sets involved in input data are shown. The study set is where the regression coefficient between the phenotype and the genotype is obtained from. The training 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. The crossed out term was set to 1 as an approximation, since its calculation depends on generally unavailable data. We found this approximation to have negligible impact on the results.

results.

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.

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

- study set,
- training set,
- 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 download information can be found at the `https://github.com/hakyimlab/MetaXcan` resource.

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 generated from a normal distribution, using a single transcriptome model trained on Depression Genes and Network's (DGN) Whole Blood data set [4] downloaded from PredictDB (`http://predictdb.org`). As genotypes we used three ancestral subsets of the 1000 Genomes project: Africans (n=662), East Asians (n=504), and Europeans (n=503). Each set was taken in turn as reference and study set yielding a total of 9 combinations as shown in Figure 3. For each population combination, we computed PrediXcan association results for the simulated phenotype and compared them with results generated from our MetaXcan approach in a scatter plot. This allowed us to assess the effect of ancestral differences between study and reference sets.

As expected, when the study and reference sets are the same, the concordance between MetaXcan and PrediXcan is 100% whereas for sets of different ancestral origin the R^2 drops a few percentage points, with the biggest loss (down to 85%) when the study set is African and the reference set is Asian. This

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

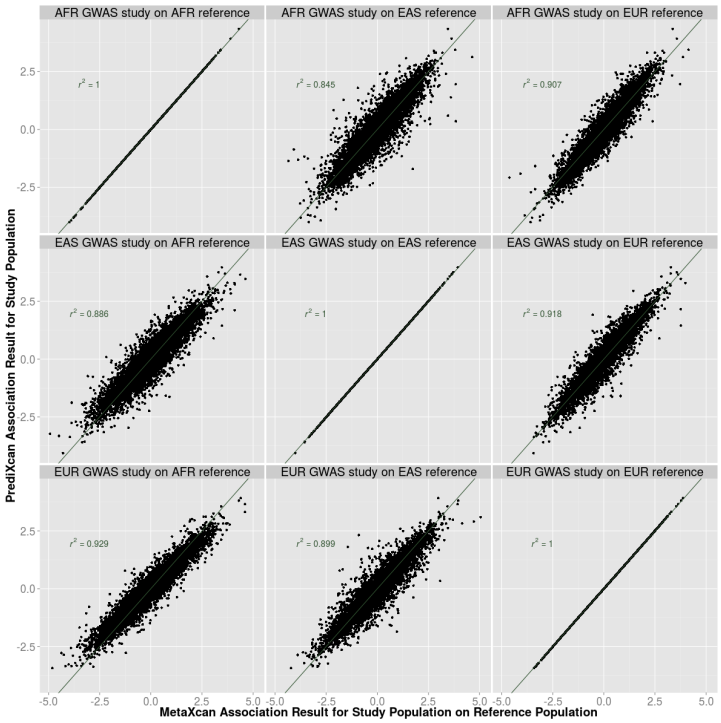

Figure 3. Comparison of PrediXcan and MetaXcan results for a simulated phenotype. Study populations and MetaXcan reference populations were built from European, African, and Asian individuals from the 1000 Genomes Project. Gene Expression model was based on DGN's Whole Blood data.

confirmed that our formula works as expected and that the approach is robust to ethnic differences between study and reference sets.

Performance in cellular growth phenotype from 1000 genomes cell lines

Next we tested with an actual cellular phenotype. Intrinsic growth, a cellular phenotype, was computed based on multiple growth asays for over 500 cell lines from the 1000 Genomes project [11]. We used a subset of values for Europeans (EUR), Africans (AFR), Asians (EAS) individuals.

We compared Z-scores for intrinsic growth generated by PrediXcan and MetaXcan for different combinations of reference and study sets, using whole blood prediction model trained in the DGN cohort. The results are shown in Figure 4. 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

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

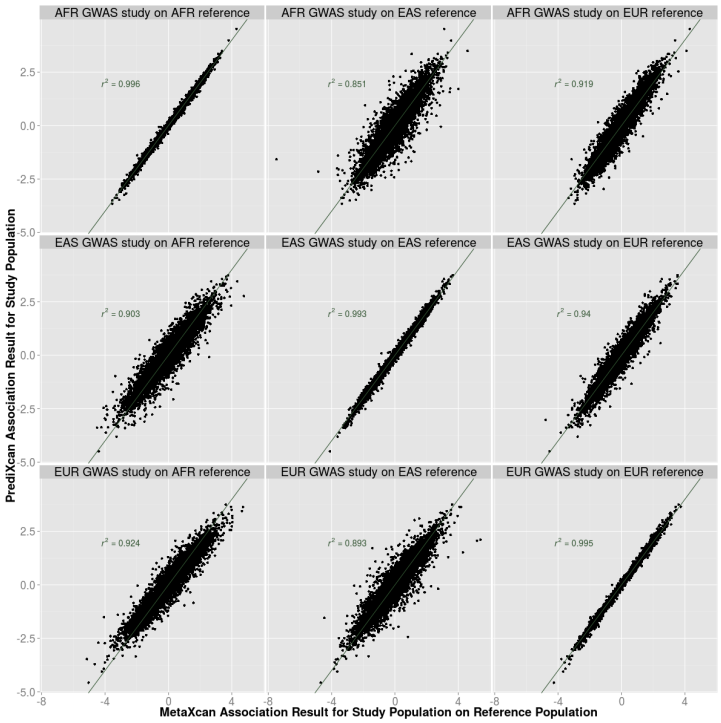

Figure 4. Comparison of PrediXcan and MetaXcan results for a cellular phenotype, intrinsic growth. Study sets and MetaXcan reference sets consisted of European, African, and Asian individuals from the 1000 Genomes Project. Gene Expression model was based on Depression Genes and Networks.

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, thus the study and reference sets were not identical for MetaXcan.

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 5. Other disease phenotypes exhibited similar performance (data not shown). Concordance between MetaXcan and PrediXcan is over 95% in for both diseases (BD $R^2 = 0.956$ and T1D $R^2 = 0.958$). The very small discrepancies 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%).

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

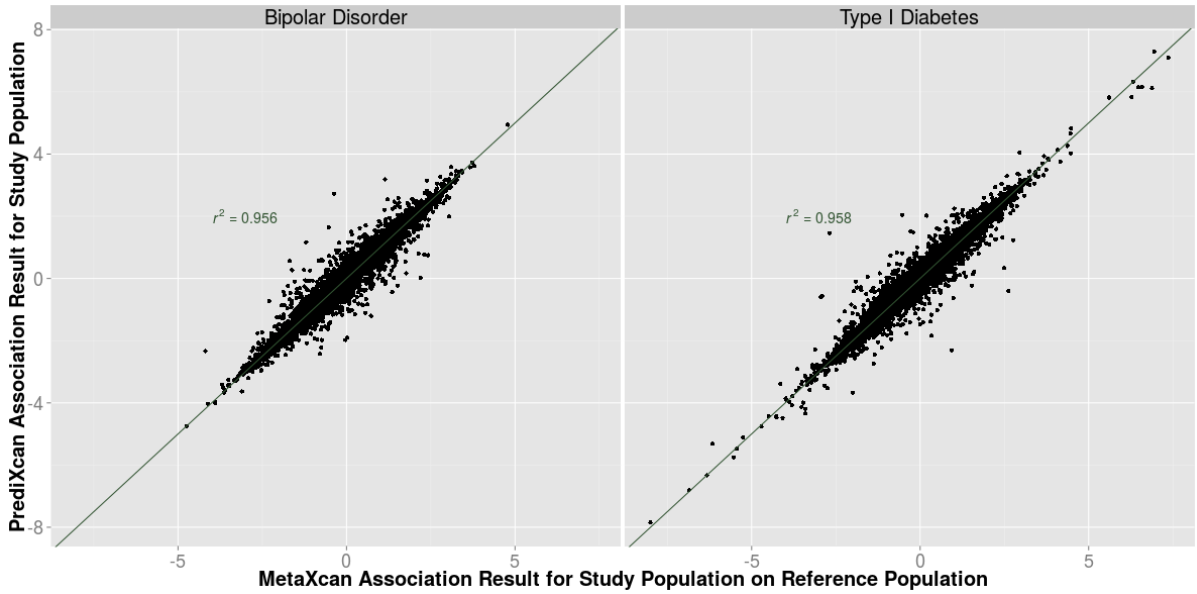

Figure 5. Comparison of PrediXcan results and MetaXcan results for a Type I Diabetes study, and a Bipolar Disorder study. Study data was extracted from Wellcome Trust Case Control Consortium, and MetaXcan reference population were the European individuals from Thousand Genomes Project (same as in previous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

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 [12], 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.

This high concordance also shows that the approximation where we drop the term $\sqrt{\frac{1-R_l^2}{1-R_g^2}}$ does not significantly affect the results.

Software

We make our software publicly available on a GitHub repository: `https://github.com/hakyimlab/MetaXcan`. Instructions for obtaining the weights and covariances for different tissues can be found there. A short working example can be found on the GitHub page; more extensive documentation can be found on the project's wiki page.

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Here we present MetaXcan, a scalable, accurate, and efficient method for integrating 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 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 through PredictDB (`http://predictdb.org`).

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 [10]. 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 [13]. This approach is equivalent to predicting expression levels using BLUP/Ridge Regression, which has been shown to be suboptimal for prediction. This is due to the fact that the local architecture of gene expression traits is sparse so that highly polygenic models underperform

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

more sparse prediction models such as LASSO or Elastic Net with mixing parameters 0.5 or greater [14].

In contrast, MetaXcan is not restricted to one imputation or prediction scheme. It infer the results of PrediXcan using summary statistics through an analytic formula. Thus it can be applied to linear models based on SNP data.

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 and Preliminaries

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

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

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

where $\hat{\gamma}_g$ 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 $\bar{Y} = \sum_{i=1,n} Y_i/n$

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

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

$\Gamma_g = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})/n,$

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 this notation, 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

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

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

Calculation of regression coefficient γ_g

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

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

Calculation of standard error of γ_g

Also from the properties of linear regression we know that

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

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

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

Thus

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

Notice that the right hand side of (9) is dependent on the SNP l while the left hand side is not. This

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

equality will hold 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

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

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

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 loss of power is compensated by the large effect size.

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgments

Grants

We acknowledge the following US National Institutes of Health grants: R01MH107666 (H.K.I.), K12 CA139160 (H.K.I.), T32 MH020065 (K.P.S.), R01 MH101820 (GTEx), P30 DK20595 and P60 DK20595 (Diabetes Research and Training Center), P50 DA037844 (Rat Genomics), P50 MH094267 (Conte). H.E.W. was supported in part by start-up funds from Loyola University Chicago.

References

1. 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).

2. 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).

3. 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.

4. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Research. 2014;24(1):14–24.

5. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PaC, Monlong J, Rivas Ma, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501(7468):506–11. Available from: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3918453{&}tool=pmcentrez{&}rendertype=abstract`.

6. Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, et al. Identification of common genetic variants controlling transcript isoform variation in human whole blood. Nature Genetics. 2015;47(4):345–352. Available from: `http://www.nature.com/doifinder/10.1038/ng.3220`.

bioRxiv preprint doi: https://doi.org/10.1101/045260; this version posted March 23, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

7. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of Cis regulatory variation in diverse human populations. PLoS Genetics. 2012;8(4).

8. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nature genetics. 2013;45(6):580–5. Available from: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4010069{&}tool=pmcentrez{&}rendertype=abstract`.

9. 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`.

10. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature Genetics. 2016;48:245–252.

11. Im HK, Gamazon ER, Stark AL, Huang RS, Cox NJ, Dolan ME. Mixed effects modeling of proliferation rates in cell-based models: Consequence for pharmacogenomics and Cancer. PLoS Genetics. 2012;8(2).

12. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genetics. 2013;9(2).

13. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. Bioinformatics (Oxford, England). 2014;30(20):2906–2914.

14. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, , et al. Survey of the Heritability and Sparsity of Gene Expression Traits Across Human Tissues. bioRxiv. 2016;Available from: `http://biorxiv.org/content/early/2016/03/15/043653.1`.