Summary Statistics Based Method to Infer Phenotypic Consequences of Gene Regulation

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

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

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

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

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

5 Section of Endocrinology, The University of Chicago, Chicago, IL, USA

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

Abstract

To understand the biological mechanisms underlying thousands of genetic variants robustly associated with complex traits, scalable methods that integrate GWAS and functional data generated by large-scale efforts are needed. Here we propose a method termed MetaXcan that addresses this need by estimating the effect of DNA variation on expression levels and subsequently testing the downstream effect on complex traits using only summary level data.

We show that the concordance between results obtained using individual level data using PrediXcan is excellent when the right reference population is used ($R^2 > 0.95$) and robust to population mismatches ($R^2 > 0.85$).

We also show the connection with other methods such as SMR and TWAS, which can be thought of special cases when selecting the appropriate prediction model (Alvaro: add proof of this to the methods section). We apply MetaXcan to 117 phenotypes across 44 human tissue expression models and show its ability to provide novel mechanistic insights.

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 [**?**]. 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 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 [**?**]. A method based on similar ideas has been proposed [**?**].

On the other hand, meta-analysis efforts that aggregate results from multiple GWAS studies have been able to identify an increasing number of 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 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. Taking advantage of publicly available large scale meta analysis

results, we train prediction models for expression in 44 human tissues with the latest GTEx release data (V6p), apply it to 100+ phenotypes, and start to build a comprehensive catalog of phenotypic consequences of gene regulation.

Results

Inferring PrediXcan results with summary statistics

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 results.

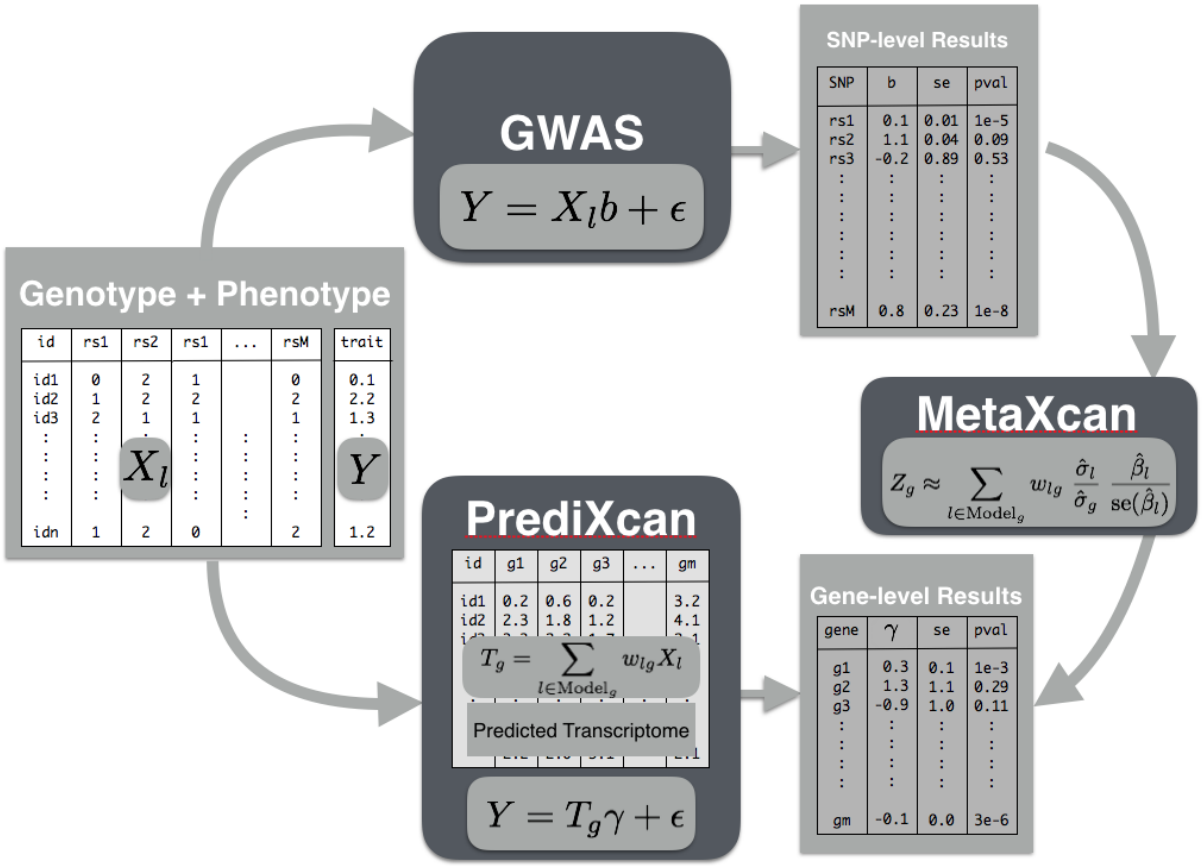


Figure 1. This figure illustrates the MetaXcan method in relationship to GWAS and PrediXcan. Both GWAS and PrediXcan take genotype and phenotype data as input. GWAS computes the regression coefficients of $Y \sim X_l$ using the model $Y = X_l b + \epsilon$, where Y is the phenotype and X_l the individual 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.

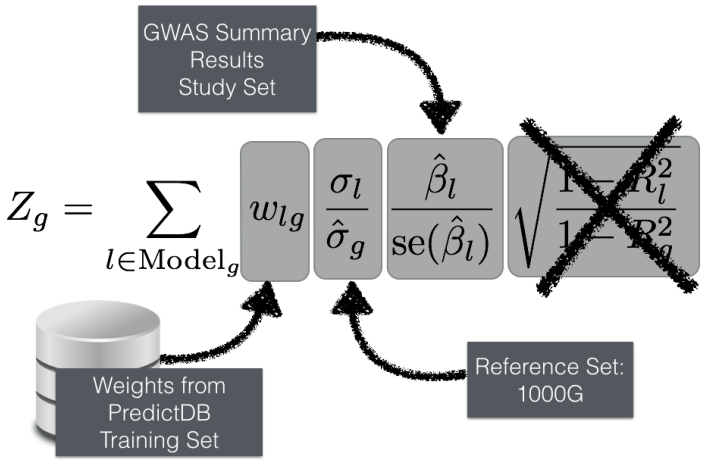


Figure 2. MetaXcan formula. This plot shows the formula to infer PrediXcan gene-level association results using summary statistics. The different sets involved in input data are shown. The study set is where the regression coefficient between the phenotype and the genotype is obtained from. The training 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.

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,
- 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.

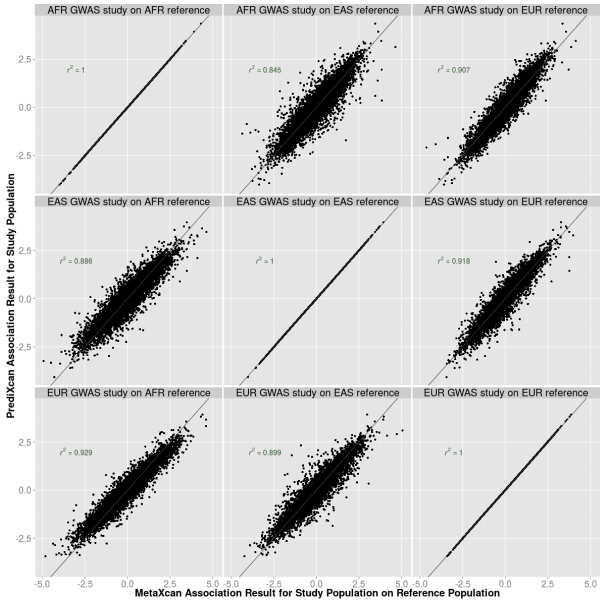


Figure 3. Comparison of PrediXcan and MetaXcan results for a simulated phenotype. Study populations and MetaXcan reference populations were built from European, African, and Asian individuals from the 1000 Genomes Project. Gene Expression model was based on DGN's Whole Blood data.

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

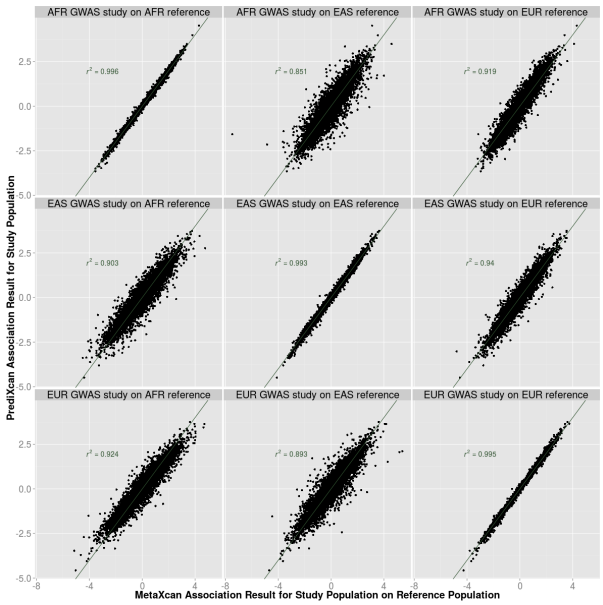


Figure 4. Comparison of PrediXcan and MetaXcan results for a cellular phenotype, intrinsic growth. Study sets and MetaXcan reference sets consisted of European, African, and Asian individuals from the 1000 Genomes Project. Gene Expression model was based on Depression Genes and Networks.

Performance in cellular growth phenotype from 1000 genomes cell lines

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 [**?**]. 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 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%).

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.

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.

Comparison with SMR (Yang/Vischer 2016 NG)

Figure 6 compares

Yang et al have proposed SMR [**?**], a summary data based Mendelian randomization that integrates eQTL results to determine target genes of complex trait-associated GWAS loci.

They derive an approximate chi-square statistic (Eq 5 in SMR) for the mediating effect of the target gene expression on the phenotype. We write the expression in terms of the inverse of the statistic that is easier to interpret.

$$\frac{1}{T_{\text{SMR}}} = \frac{1}{Z^2_{\text{eqtl}}} + \frac{1}{Z^2_{\text{GWAS}}} \tag{2}$$

where T_{SMR} = (effect size/standard error)^2 , Z_{eqtl} is the Z score (=effect size / standard error) of the association between SNP and gene expression, and Z_{GWAS} is the Z score of the association between SNP and trait.

Thus the inverse of the square of the Wald statistic derived by Yang et al is the sum of two inverse

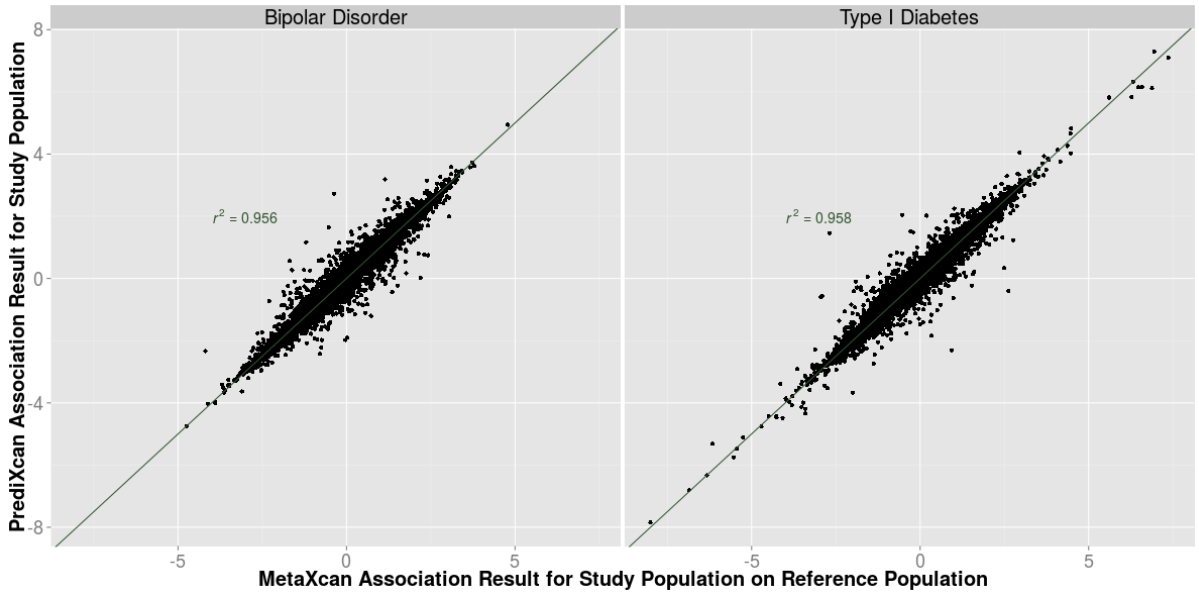


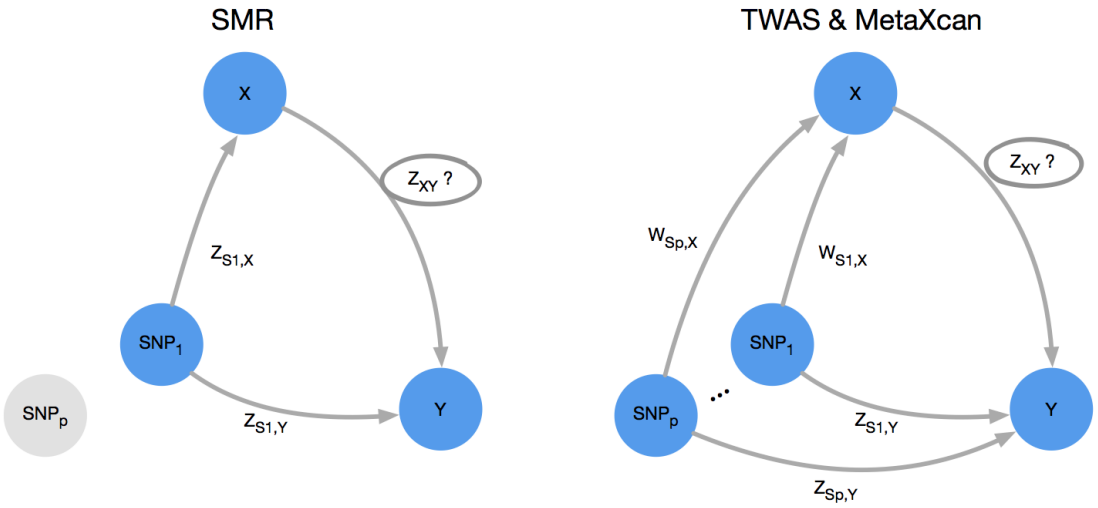
Figure 5. Comparison of PrediXcan results and MetaXcan results for a Type I Diabetes study, and a Bipolar Disorder study. Study data was extracted from Wellcome Trust Case Control Consortium, and MetaXcan reference population were the European individuals from Thousand Genomes Project (same as in previous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

χ^2 (Z is asymptotically normally Z2 is chisquare 1/Z2 is inverse chisquare).

We know that the sum of two inverse χ^2 random variables does not yield a χ^2, so that there are additional assumptions that need to be made. Since each term is inverse χ^2, for the sum to be inverse χ^2, one of the two terms must be negligible compared to the other. SMR assumes that the eQTL effect is strong, i.e. when the first term on the RHS is small. Under this assumption the T_{SMR} is approximately χ^2.

SMR is applied transcriptome wide using the top eQTL for each gene. If we use the top eQTL as the sole predictor of gene expression in PrediXcan/MetaXcan, the association statistic is equal to 1/Z2(GWAS). Thus under the assumption of strong eQTL effect, SMR is equivalent to MetaXcan using top eQTL to predict expression level of the (potential) target gene.

SMR will work well when there is a single variant that affects expression levels and through them alters the phenotype. In this case, an advantage of SMR over MetaXcan is that it incorporates uncertainty in the eQTL association $\frac{1}{Z^2_{\text{eqtl}}}$. However, in the range where the statistic T_{SMR} is approximately correct, the difference between MetaXcan and SMR results is also negligible. Extension of MetaXcan that takes into



$$\frac{1}{Z^2_{X,Y}} = \frac{1}{Z^2_{S1,X}} + \frac{1}{Z^2_{S1,Y}}$$

single causal SNP implicitly assumed

approximate Chi2 only valid when 1/ Z^2_{S1,X} is very small compared to 1/Z^2_{S1,Y}

if top eQTL is used as predictor of gene expression we get the same result in the range where approx is valid

$$Z_{X,Y} = W \cdot Z_{S,Y} / (W \Sigma_{s,s} W')^{\frac{1}{2}}$$

multiple causal SNP allowed

TWAS considers weights = \Sigma_{x,s}\Sigma^{-1}_{s,s} = BLUP/ridge regression prediction

Our derivation shows that the more general weights can be used

Approximation is made more explicit

Figure 6. CAPTION. TODO generate better figures

account the uncertainty in the prediction models will be ideal.

Comparison with TWAS (Gusev/Pasaniuc NG)

Gusev et al have proposed a method comparable to MetaXcan that is based only on summary statistics. This 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 [**?**]. This approach is equivalent to predicting expression levels using BLUP/Ridge Regression, which has been shown to be suboptimal [**?**]. This is due to the fact that the local architecture of gene expression traits is sparse so that highly polygenic models underperform more sparse prediction models such as LASSO or Elastic Net with mixing parameters 0.5 or greater.

Prediction models across human tissues

Using the release version 6p (dbGaP Accession phs000424.v6.p1) from GTEx, we have trained prediction models for expression levels of 44 human tissues with a total of 1,091,787 gene tissue pairs. Among these 203,494 yielded prediction models with cross validated q value < 0.05 (FDR computed among each tissue models), which were saved into the Predictdb database and used for subsequent analysis.

We use SNPs within 1Mb upstream of the TSS and 1Mb downstream of the TES. We use elastic net with a mixing parameter of 0.5, a multivariate linear model estimated via penalized maximum likelihood. As reported in [**?**, **?**] overall performance does not vary for a range of values of the mixing parameter except when the model becomes very close to ridge regression (fully polygenic). Based on this we chose to use elastic net with 0.5 as mixing parameter, which will be more robust to low qualify genotype or imputation.

Supplementary Table 1: list of tissues, sample sizes, summary R2, # with qval ¡ 0.05, # total attempted training, # total successful models.

Catalog of the phenotypic consequences of gene regulation

Next we downloaded summary statistics of meta analysis of 113 phenotypes (encompassing xxx main phenotypes) from 16 consortia. The full list of consortia and phenotypes is shown in Supplementary Table 2. We computed association between these phenotypes and the predicted expression levels in 44 human tissues using elastic net models described in the previous section and a whole blood model. TODO

Alvaro: create top eQTL model using the Blood eqtl summary data and run metaxcan. this should serve as an additional training set replication for whole blood.

To facilitate query of the results, we created a web application that allows filtering the results by gene, phenotype, tissue, p value, and prediction performance (gene2pheno.org). For each trait we assigned ontology terms from EFO and HPO.

Supplementary Table 2: List of consortia and phenotypes for which gene level association are available. Gene/tissue pair results were considered significant when p-value¡0.05/total number of gene/tissue pairs tested (which ranged from xxe-6 - xxe-6).

Manhattan and qqplots for each phenotype (all tissue results combined) can be found in Supplementary Figure xx.

Mostly genome-wide significant genes tend to cluster around SNP level genome-wide significant loci or sub-genome wide significant loci. Because of the reduction in multiple testing or an increase in power because it takes into account of the combined effects of multiple variants, sub-genome-wide significant genes can become gw significant in MetaXcan. Alvaro TODO: find examples. Do we find examples where MetaXcan p values are more significant than GWAS pvalues? I am guessing not but please check.

Results of MetaXcan tend to be more significant as R2 goes up, i.e. the genetic component of expression is larger. The trend is seen both when results are averaged across all tissues for a given phenotype or across all phenotypes for a given tissue. Representative phenotypes were chosen. Using squared zscores or abs zscores did not change results much.

Similarly, results of MetaXcan tend to be more significant as prediction p values are more significant. The trend is seen both when results are averaged across all tissues for a given phenotype or across all phenotypes for a given tissue. Representative phenotypes were chosen. Using squared zscores or abs zscores did not change results much.

Enrichment of genes in ClinVar

Genes implicated in ClinVar to obesity, rheumatoid arthritis, diabetes, Alzheimer's, Crohn's disease, ulcerative colitis, age-related macular degeneration, and autism show inflated significance among metaxcan association results for corresponding diseases. Schizophrenia genes (20 available) do not show any inflation (probably a limitation of the ClinVar database).

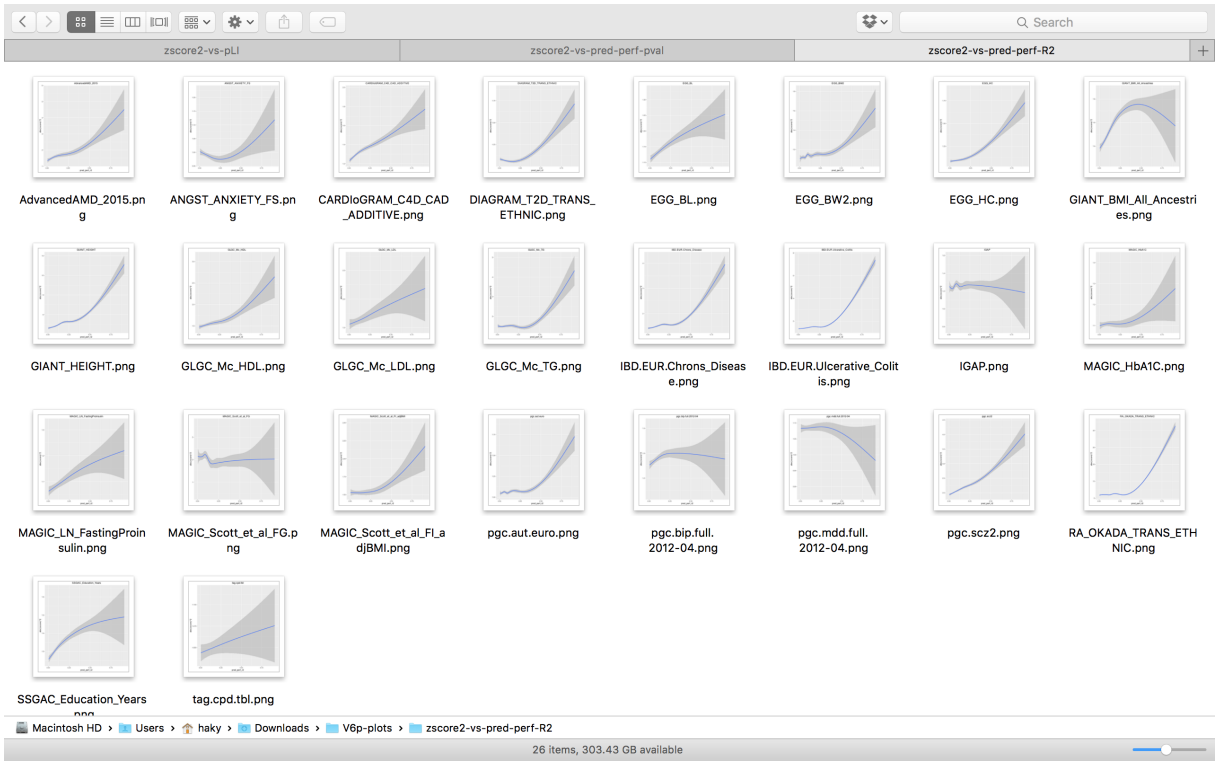


Figure 7. zscore2 vs predicted performance R2 by phenotype TODO generate better figures, move these large figures to supplementary. Combine all four Z2 vs R2 and pppval into one figure.

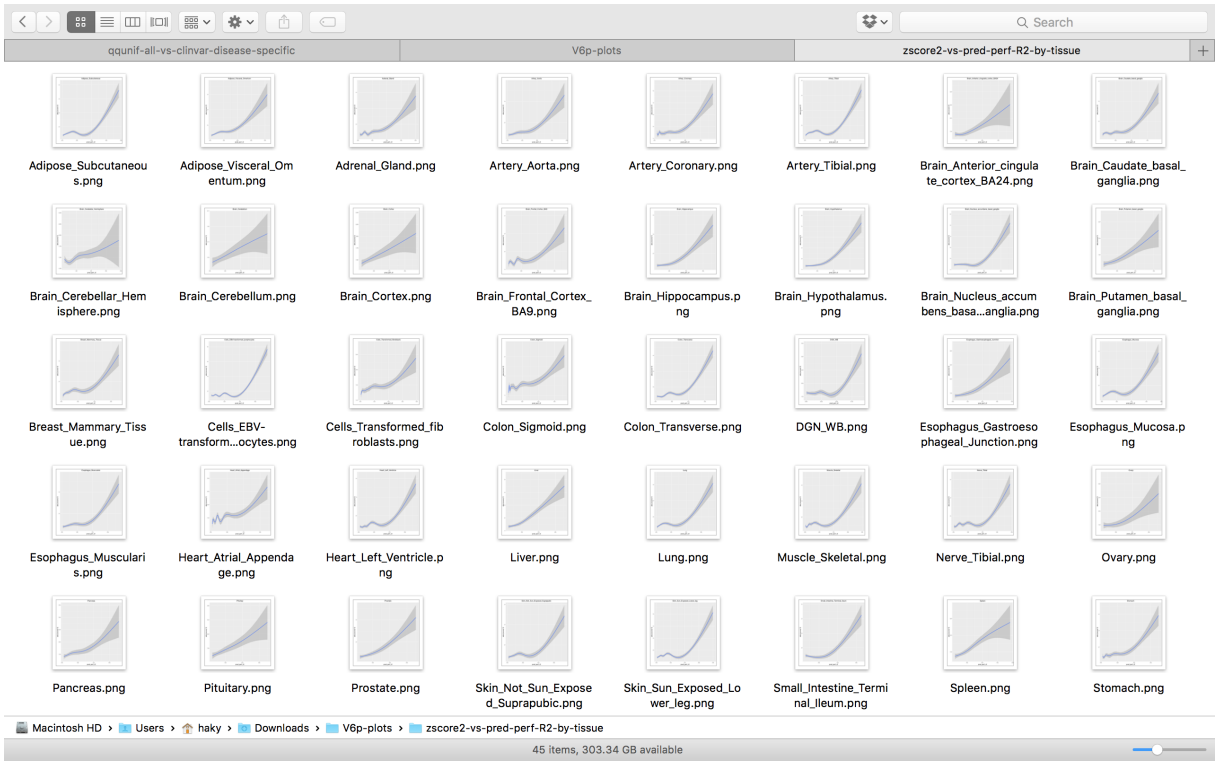


Figure 8. zscore2 vs predicted performance R2 by tissue TODO generate better figures, move these large figures to supplementary. Combine all four Z2 vs R2 and pppval into one figure.

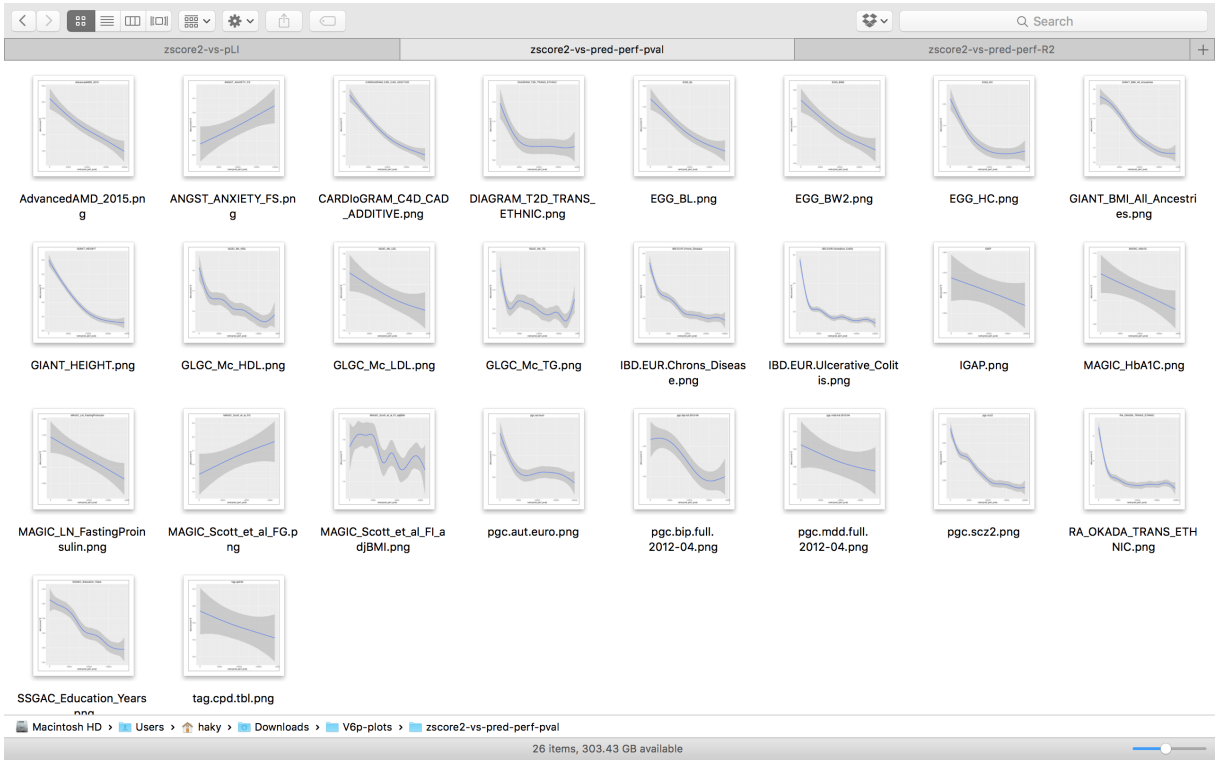


Figure 9. zscore2 vs predicted performance pval by phenotype TODO generate better figures

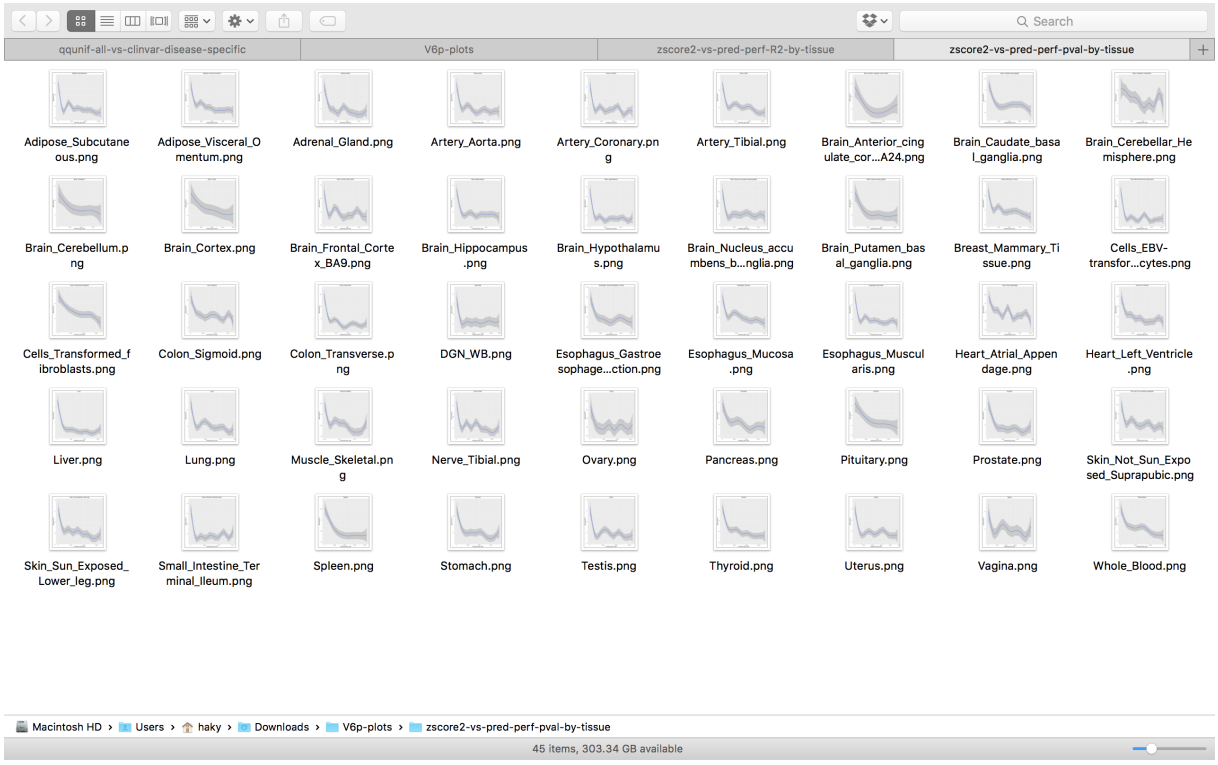


Figure 10. zscore2 vs predicted performance pval by tissue TODO generate better figures

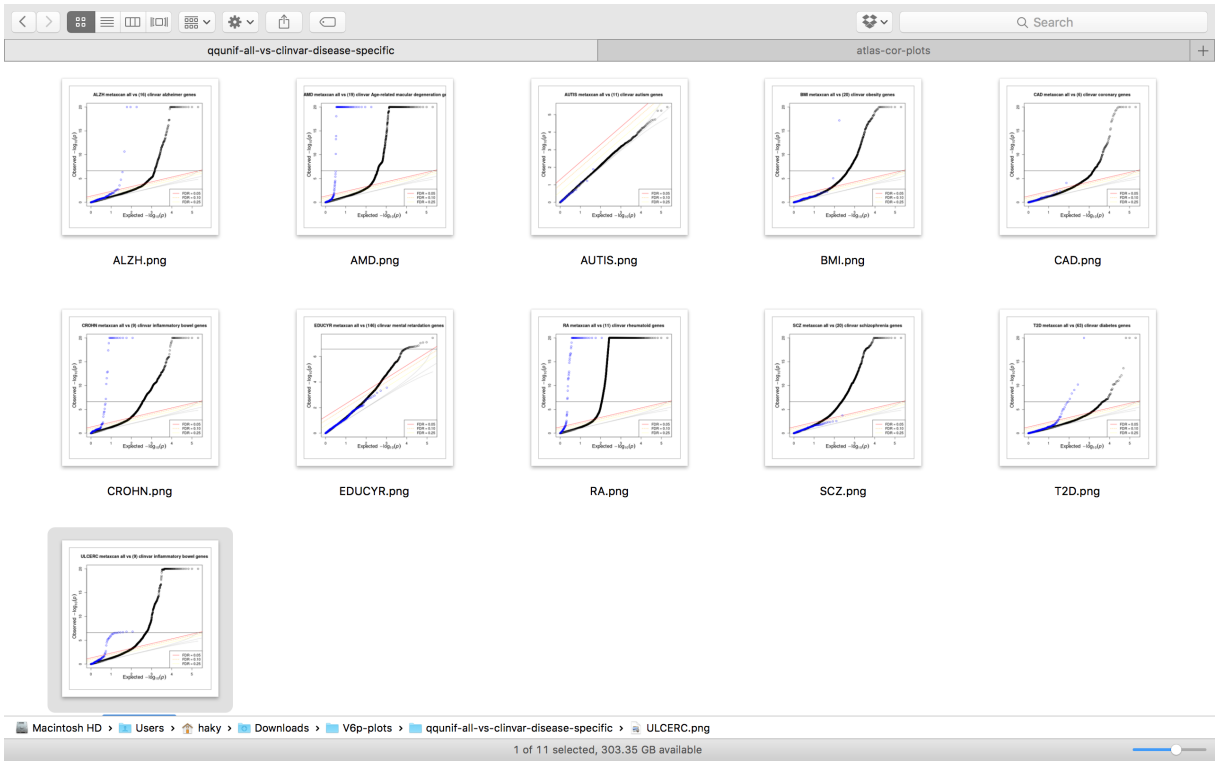


Figure 11. ClinVar genes enriched among significant metaxcan genes for most diseases we tested except for schizophrenia. TODO generate better figures

qqplot comparing all pvals vs tissue specific pvals

When comparing p-value distributions for different tissues, we were expecting to find that tissues relevant to the disease would show up as more significant. We find this to be the case for coronary artery disease: the tissues with most significant genes were Artery tissues (coronary, tibial, and aorta) and liver. Also LDL cholesterol associated genes showed enrichment of genes predicted in liver. However that was not the case for HDL and TG, for which whole blood models were most enriched among top results. In T2D, muscle models depart the most from the reference distribution. But this is due to only a handful of genes. Also TCF7L2, a positive control gene for T2D, only shows significant association in aortic artery.

What is the difference between regions with multiple target genes and single target genes?

What is the difference between regions with genes significant across multiple tissues vs single tissues?

how to distinguish causal vs passenger association?

Regulation may not always be identifiable in the tissue of action

There is value in looking at the results across all tissues even though the causal mechanism that alters disease risk may not happen in all tissues, the regulation may be detectable in other tissues.

For example, rs11206510 and rs2479394 were shown to be eQTL in VAF (Visceral Abdominal Fat) in the STARLET study and a study with morbidly obese patients (Greenawalt, Genome Res 2011).

However in GTEX rs11206510 was associated with PCSK9 only at p=0.035, which in the context of genome wide study with multiple testings would not reach significance.

However in GTEx, we find that the regulation of PCSK9 is much more active in tibial nerve. It is unlikely that tibial nerve is the tissue of action of cardio-metabolic traits. (p=1e-14 and 1e-6)

Conditions of sample treatment (extraction, treatment) and demographic variables results in making tibial nerve the most conducive environment to detect the link between these variants and PCSK9.

Results across all tissues as replicated experiments

Since it has been reported that a large portion of cis regulation is shared across tissues we hypothesize that gene level associations can be detected using tissues that are not necessarily the tissue of action of the disease. Also the complexity of gene regulation and perhaps environmental state dependence of

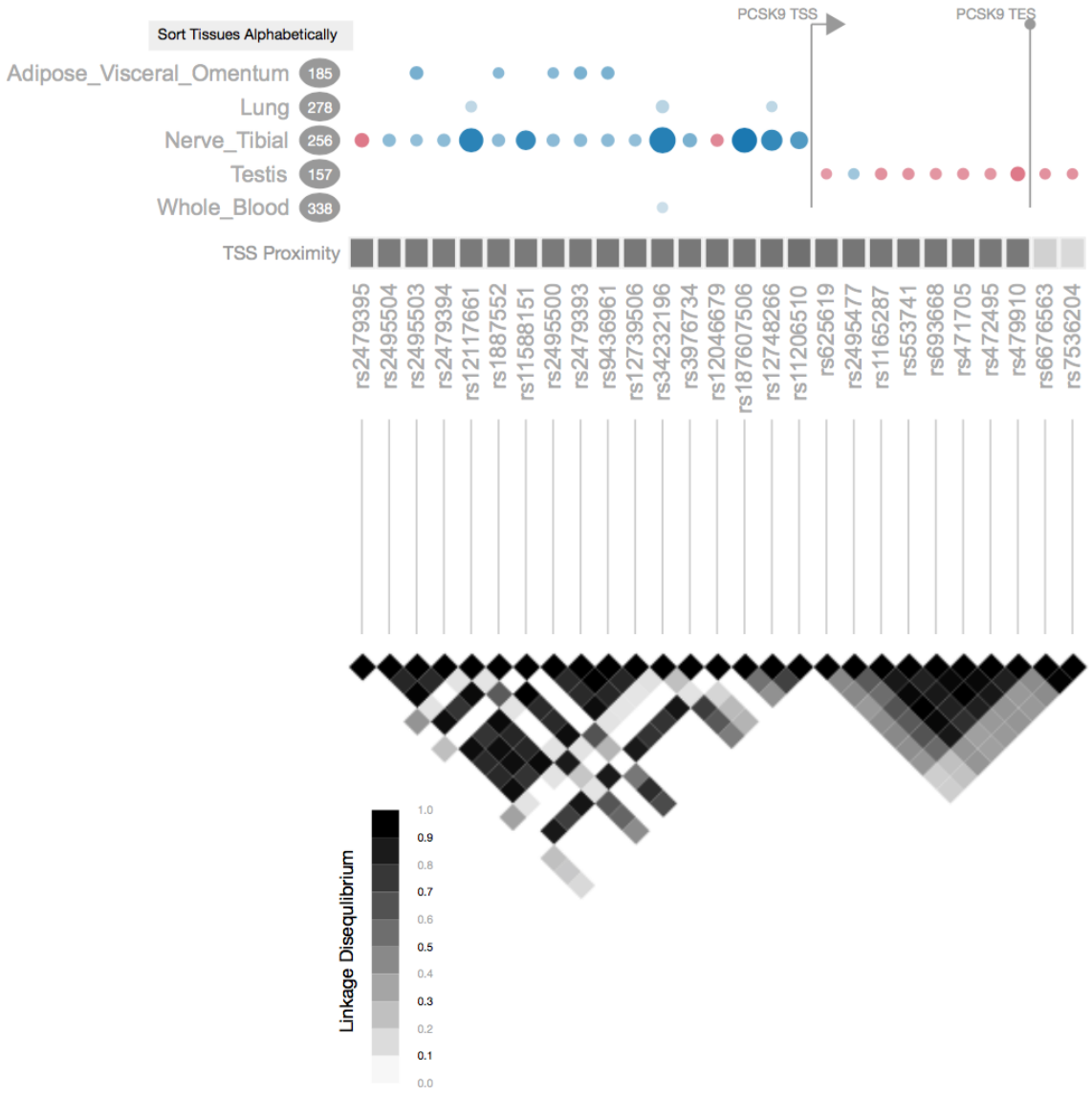


Figure 12. PCSK9 most actively regulated in Tibial Nerve in GTEx. TODO generate better figures

regulatory processes our models may be missing some of the associations because they are based on a one-time snapshot of the transcriptome averaged over many cells. Our results suggest that expanding our analysis to tissues other than the ones we expect to be involved in the phenotype of interest increases our ability to identify disease relevant genes.

Correlation between phenotypes

For a given tissue, we computed the correlation between the zscores from different phenotypes. The correlation is consistent with the results reported by Bulik-Sullivan et al.

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.

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

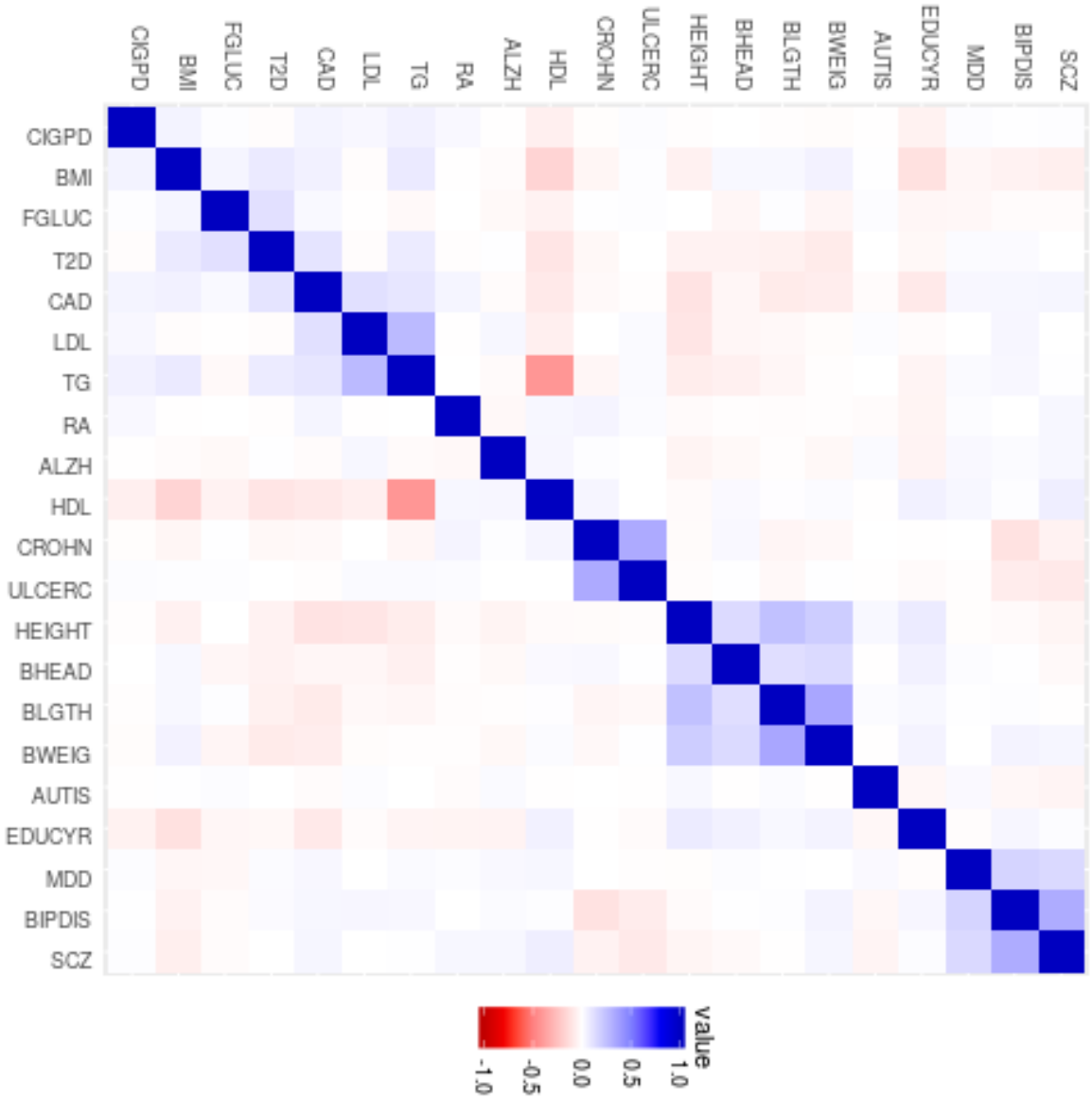


Figure 13. Correlation between traits. TODO generate better figures

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 relatively well balanced (less than 3 to 1 ratio between groups) 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 [**?**]. 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 ImpG, a method proposed by Pasaniuc and others [**?**]. 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 more sparse prediction models such as LASSO or Elastic Net with mixing parameters 0.5 or greater [**?**].

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 any linear models based on SNP data. For example, given the LD we are able to "predict" structural variation or more complex genetic variation as a linear function of SNP data. MetaXcan allows us to infer the assocation between these variants and complex traits using existing publicly available summary statistics.

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.

We show that MetaXcan includes other state of the art methods such as TWAS and SMR as special cases when the appropriate prediction model is selected.

Furthermore, we train prediction models for 44 human tissue expression levels from GTEx and apply to 117 phenotypes with meta analysis results publicly available from 16 large consortia.

As expected, better predicted genes (larger heritable component or more significant predictions) are

more likely to be identified as associated with complex traits. We also find that for the disease genes we have tested ClinVar genes are highly enriched.

Due to the large shared regulation across tissues, we found that examining results in tissues that are not traditionally considered relevant for the disease of interest increase our ability to identify disease genes.

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.

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

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

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

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

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

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, 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:

$$\hat{\gamma}_g = \frac{\text{Cov}(T_g, Y)}{\text{Var}(T_g)} = \frac{\text{Cov}(T_g, Y)}{\hat{\sigma}_g^2} \tag{3}$$

and

$$\hat{\beta}_l = \frac{\text{Cov}(X_l, Y)}{\text{Var}(X_l)} = \frac{\text{Cov}(X_l, Y)}{\hat{\sigma}_l^2} \tag{4}$$

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

$$T_g = \sum_{l \in \text{Model}_g} w_{lg} X_l \tag{5}$$

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

$$\begin{aligned}
\hat{\sigma}_g^2 &= \text{Var}\left(\sum_{l \in \text{Model}_g} w_{lg} X_l\right) \\
&= \text{Var}(\mathbf{W}_g \mathbf{X}_g) && \text{where } \mathbf{W}_g \text{is the vector of } w_{lg} \text{for SNPs in the model of } g \\
&= \mathbf{W}_g' \text{Var}(\mathbf{X}_g) \mathbf{W}_g && \text{where } \Gamma_g \text{ is the} \text{Var}(\mathbf{X}_g) = \text{ covariance matrix of } \mathbf{X}_g \\
&= \mathbf{W}_g' \Gamma_g \mathbf{W}_g && (6)
\end{aligned}$$

Calculation of regression coefficient γ_g

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

$$\begin{aligned}
\hat{\gamma}_g &= \frac{\text{Cov}(T_g, Y)}{\hat{\sigma}_g^2} \\
&= \frac{\text{Cov}(\sum_{l \in \text{Model}_g} w_{lg} X_l, Y)}{\hat{\sigma}_g^2} \\
&= \sum_{l \in \text{Model}_g} \frac{w_{lg} \text{Cov}(X_l, Y)}{\hat{\sigma}_g^2} && \text{by linearity of Cov} \\
&= \sum_{l \in \text{Model}_g} \frac{w_{lg} \hat{\beta}_l \sigma_l^2}{\hat{\sigma}_g^2} && \text{using Eq 4} \qquad (7)
\end{aligned}$$

Calculation of standard error of γ_g

Also from the properties of linear regression we know that

$$\text{se}(\hat{\gamma}_g) = \sqrt{\text{Var}(\hat{\gamma}_g)} = \frac{\hat{\sigma}_\epsilon}{\sqrt{n \hat{\sigma}_g^2}} = \frac{\hat{\sigma}_Y^2 (1 - R_g^2)}{n \hat{\sigma}_g^2} \qquad (8)$$

In this equation, σ_Y / n is not necessarily known but can be estimated using the analogous equation (8) for beta

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

Thus

$$\frac{\hat{\sigma}_Y^2}{n} = \frac{\text{se}(\hat{\beta}_l)^2 \hat{\sigma}_l^2}{(1 - R_l^2)} \qquad (10)$$

Notice that the right hand side of (10) is dependent on the SNP l while the left hand side is not. This

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 $\text{se}(\gamma_g)$, or Z score,

$$Z_g = \frac{\hat{\gamma}_g}{\text{se}(\hat{\gamma}_g)} \tag{11}$$

with which we can compute the p value as

$$p = 2 \text{ pnorm}(-|Z_g|) \tag{12}$$

$$\begin{aligned}
Zg &= \frac{\hat{\gamma}_g}{\text{se}(\hat{\gamma}_g)} \\
&= \sum_{l \in \text{Model}_g} \frac{w_{lg}\hat{\beta}_l\sigma_l^2}{\hat{\sigma}_g^2} \sqrt{\frac{n}{\hat{\sigma}_Y^2} \frac{\hat{\sigma}_g^2}{(1-R_g^2)}} && \text{using Eq. 7 and 8} \\
&= \sum_{l \in \text{Model}_g} \frac{w_{lg}\hat{\beta}_l\sigma_l^2}{\hat{\sigma}_g} \sqrt{\frac{(1-R_l^2)}{\text{se}(\hat{\beta}_l)^2\hat{\sigma}_l^2}} \sqrt{\frac{1}{(1-R_g^2)}} \\
&= \sum_{l \in \text{Model}_g} w_{lg} \frac{\sigma_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)} \sqrt{\frac{1-R_l^2}{1-R_g^2}} && (13) \\
&\approx \sum_{l \in \text{Model}_g} w_{lg} \frac{\sigma_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)} && (14)
\end{aligned}$$

Based on results with actual and simulated data we have found that the last approximation does not affect our ability to identify the association. This is due to the fact that even with a small decrease in power due to the approximation is compensated by the overall high power induced by the large effect size.

Effect size by one standard deviation in predicted expression

TODO: effect size output interpretation. confirm that it is the effect (on the scale of the phenotype used

in the GWAS) of one standard deviation change in prediction gene expression To derive the analytic expression that computes the effect of one standard deviation in the predicted expression on the trait. TODO Alvaro: write down the derivation of the formula here

Expression model training

TODO: Scott - write up method to perform prediction.

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.