Integrating tissue specific mechanisms into GWAS summary results

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], GTEx Consortium, 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 Medical Center, Nashville, TN, USA

4 Departments of Biology and Computer Science, Loyola University 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 inferring the downstream consequences of genetically regulated components of molecular traits on complex phenotypes using summary data only. MetaXcan allows multiple causal variants and flexible multivariate models extending the capabilities of existing methods and enabling the testing of more complex processes. As an example application, we trained prediction models of gene expression levels in 44 human tissues and inferred the consequences of their regulation in 40 complex phenotypes. Our examination of this broad set of human tissues revealed many novel genes and re-identified known ones with patterns of regulation in expected as well as unexpected tissues.

Introduction

Over the last decade, GWAS have been successful in identifying genetic loci that robustly associate with human 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 [1–3] show the importance of gene expression regulation. Direct quantification of the contribution of different functional classes of genetic variants showed that 80% of the common variant contribution to phenotype variability (in 12 diseases) can be attributed to DNAase I hypersensitivity sites, further highlighting the importance of transcript regulation in determining phenotypes [4].

Many transcriptome studies have been conducted where genotypes and expression levels are assayed for a large number of individuals [5–8]. The most comprehensive transcriptome dataset, in terms of tissues covered, is the one generated by the Genotype-Tissue Expression Project (GTEx); a large-scale effort where DNA and RNA are collected from multiple tissue samples from nearly 1000 individuals and sequenced to high coverage [9]. 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 [10], 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 in a gene-level association test that addresses some of the key limitations of GWAS [10].

A method based on similar ideas was proposed by Gusev et al. [11] called Transcriptome-wide Association Study (TWAS). For the individual level data based version, the main difference between PrediXcan and TWAS resides in the models used for the prediction of gene expression levels in the implementation. An important extension of this approach was implemented by Gusev et al. [11] that allows us to compute gene level association results using only summary statistics. We will refer to this method as TWAS-summary.

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 [12–14]. We will refer to these results as GWAMA (Genome-wide association meta analysis) results. In order to harness the power of these increased sample sizes while keeping the computational burden manageable, methods that uses summary level rather than individual level data such as TWAS-summary are needed.

Zhu et al [15] proposed another method that integrates eQTL data with GWAS results based on summary data. The method, Summary Mendelian Randomization (SMR), uses Wald statistics (effect size/standard error) from GWAS and eQTL studies and derives the effect of the genetic component of gene expression on a phenotype using the delta approximation [16]. By design, this approach uses one eQTL per gene so that in practice only the top eQTL is used per gene.

Here we present a method we call MetaXcan that greatly expands the applicability of the ideas behind PrediXcan by using only summary results. We will show that our method can reproduce PrediXcan results accurately and it is robust to ancestry mismatches between study, reference, and training populations.

In contrast to TWAS-summary, our method can be applied to any linear model of expression traits. The Gaussian imputation underlying TWAS-summary implies a ridge regression based prediction, which has been shown to be suboptimal for gene expression trait modeling [17]. In contrast to SMR, MetaXcan naturally allows for multiple causal variants by using a flexible multivariate linear model to describe expression traits as well as more complex mechanistic processes.

To illustrate the power of MetaXcan, we first train over 1 million elastic net prediction models of gene expression traits, covering protein coding genes across 44 human tissues from GTEx, and then apply it to 40 phenotypes from 17 large meta analysis consortia. We use the results of these to lay the groundwork for building a comprehensive catalog of phenotypic consequences of gene regulation across multiple tissues and contexts.

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 [10].

For both GWAS and PrediXcan, the input is a 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 regres-

sion 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 infer the output from PrediXcan. Since MetaXcan only uses summary statistics, it can effectively take advantage of large-scale meta analysis results, avoiding the computational and regulatory burden of handling large amounts of protected individual-level data.

MetaXcan formula

Figure 1B shows the main analytic expression used by MetaXcan for the Z-score (Wald statistic) of the association between predicted gene expression and a phenotype. The input variables are the weights used to predict the expression of a given gene, 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 additional information that is unavailable in typical GWAS summary statistics output such as sample size and variance of the phenotype. Dropping this factor from the formula does not affect the accuracy of the results as demonstrated in the close to perfect concordance between MetaXcan and PrediXcan results on the diagonal of Figure 2A.

The approximate formula we use is:

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

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,
- $\text{se}(\hat{\beta}_l)$ 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 (e.g. GWAS study set),

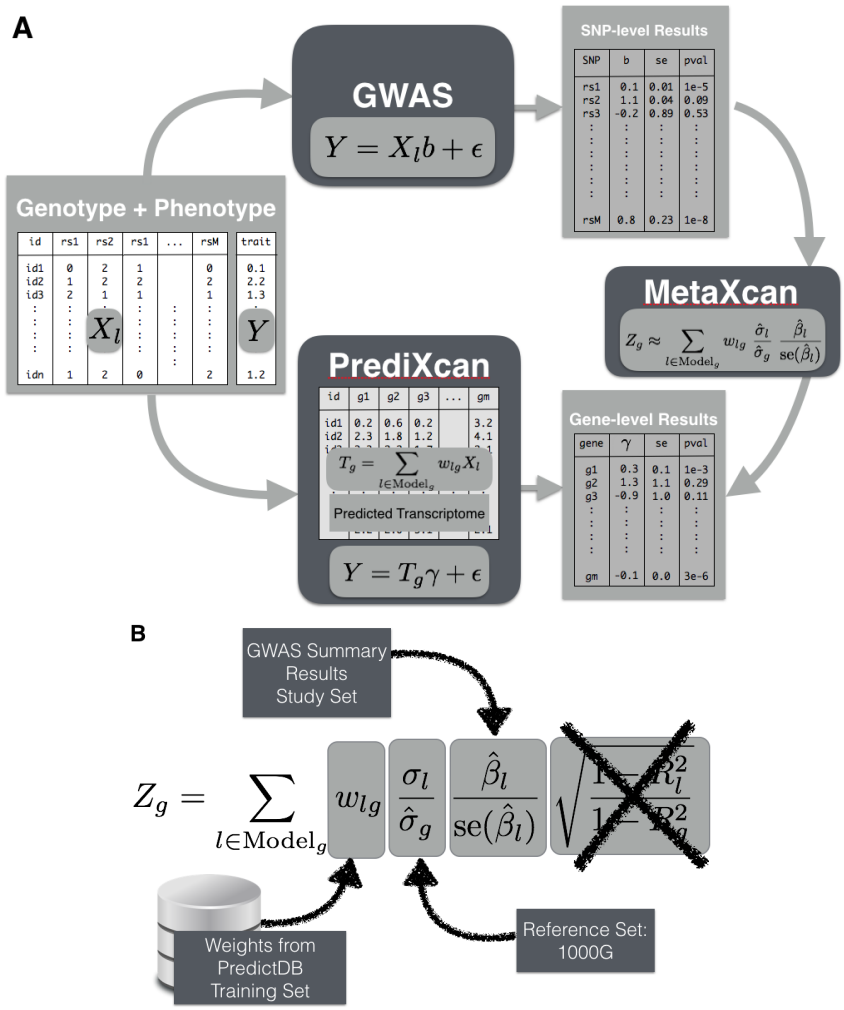


Figure 1. A) Comparison of GWAS, PrediXcan, and MetaXcan This figure illustrates the MetaXcan method in relationship to GWAS and PrediXcan. Both GWAS and PrediXcan take genotype and phenotype data as input. GWAS computes the regression coefficients of Y on X_l using the model $Y = a + X_l b + \epsilon$, where Y is the phenotype and X_l the individual SNP 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 directly computes the gene-level association results using the output from GWAS. **B) 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 regression coefficient between the phenotype and the genotype is obtained from the study set. 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.

- training set (e.g. GTEx, DGN),
- population reference set (e.g. 1000 Genomes).

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 or a SNP-level meta-analysis of multiple GWAS. 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. Training sets are also used to generate variance and covariances of genetic markers, which will usually be different from the study sets. When individual level data are not available from the training set we use population reference sets such as 1000 Genomes data.

In the most common use scenario, users will need to provide only GWAS results using their study set. The remaining parameters are pre-computed, and download information can be found at the `https://github.com/hakyimlab/MetaXcan` resource.

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 [5] downloaded from PredictDB (`http://predictdb.org`). For genotypes we used three ancestral subsets of the 1000 Genomes project: Africans (n=661), 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 2. For each population combination, we computed PrediXcan association results for the simulated phenotype and compared them with results generated using MetaXcan in a scatter plot. In this manner we 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 confirms that our formula works as expected and that the approach is robust to ethnic differences between study and reference sets.

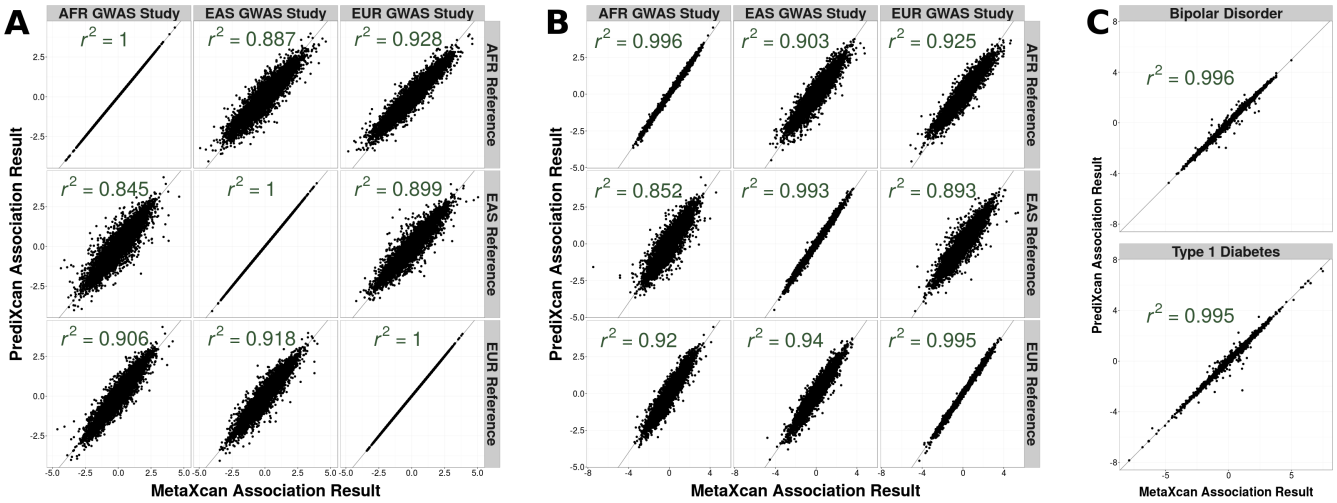


Figure 2. Comparison of PrediXcan and MetaXcan results for **(A)** a simulated phenotype, **(B)** a cellular phenotype, intrinsic growth; Study sets and MetaXcan population reference sets consisted of European, African, and Asian individuals from the 1000 Genomes Project. Gene expression prediction models were based on DGN Cohort. **(C)** Comparison of PrediXcan results and MetaXcan results for a Type 1 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 subset as in previous sections).

Performance in cellular growth phenotype from 1000 genomes cell lines

Next we tested with an actual cellular phenotype - intrinsic growth. This phenotype was computed based on multiple growth assays for over 500 cell lines from the 1000 Genomes project [18]. We used a subset of values for European (EUR), African (AFR), and Asian (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 models trained in the DGN cohort. The results are shown in Figure 2B. 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 2C. Other disease phenotypes exhibited similar performance (data not shown). Concordance between MetaXcan and PrediXcan is over 99% for both diseases (BD $R^2 = 0.996$ and T1D $R^2 = 0.995$). The very small discrepancies are explained by differences in allele frequencies and LD between the reference set (1000 Genomes) and the study set (WTCCC).

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 [19], when the number of cases and controls are relatively well balanced (roughly, at least 25% of a cohort are cases or controls), linear regression approximation yields very similar results to logistic regression. This high concordance also shows that the approximation of dropping the factor $\sqrt{\frac{1-R_l^2}{1-R_g^2}}$ does not significantly affect the results.

Comparison of MetaXcan with other integrative methods based on summary results

Zhu et al. have proposed Summary Mendelian Randomization (SMR) [15], a summary data based Mendelian randomization that integrates eQTL results to determine target genes of complex trait-associated GWAS loci. They derive an approximate χ^2 statistic (Eq 5 in SMR) for the mediating effect of the target gene expression on the phenotype. This approximation is only valid when the eQTL association is much stronger than the GWAS association $\left(Z^2_{S1,X}(\text{eQTL}) >> Z^2_{S1,Y}(\text{GWAS})\right)$. Within the range of validity of the SMR statistic, MetaXcan produces approximately the same result if the top eQTL is used as the prediction model. We demonstrate in the Methods section that the difference is negligible within the range where SMR assumptions hold.

Gusev et al. have proposed Transcriptome-Wide Association Study (TWAS-summary), which imputes the SNP level Z-scores into gene level Z-scores. This is a natural extension of ImpG [20] or DIST [21], which are SNP-based methods that impute summary statistics of unmeasured SNPs using Gaussian imputation [22]. We show that this approach is equivalent to predicting expression levels using BLUP/Ridge Regression, which has been shown to be suboptimal for gene expression traits [17]. 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.

Figure 3 illustrates the SMR, TWAS, and MetaXcan approaches. ImpG and DIST methods are also included in the figure for comparison. All three methods seek to identify target genes by computing the strength of association between the unobserved predicted expression levels (X) of a gene with the complex trait (Y) quantified with $Z_{X,Y}$ or $Z^2_{X,Y}$.

Next we demonstrate the utility of MetaXcan by training prediction models for a broad set of primary tissue expression levels from GTEx and applying them to multiple GWAMA summary results.

Prediction models across 44 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 within each tissue model) and corresponding to protein coding genes. These were saved into the publicly available PredictDB database and used for subsequent analysis.

To build the models, we use SNPs within 1Mb upstream of the TSS and 1Mb downstream of the TES. We use elastic net [23], a multivariate linear model estimated via penalized maximum likelihood, with a mixing parameter of 0.5. As reported in [10, 17] overall performance remains similar for a range of values of the mixing parameter but drops abruptly 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 retains several correlated predictors and is consequently more robust to missing genotype data or low quality imputation. A summary of tissues, sample sizes, and number of attempted and successful models (FDR $< 5\%$) can be found in Supplementary Table 7.

Catalog of the phenotypic consequences of gene regulation

Next we downloaded summary statistics of meta analysis of 40 phenotypes from 17 consortia. The full list of consortia and phenotypes is shown in Supplementary Table 3. We tested association between these phenotypes and the predicted expression levels using elastic net models in 44 human tissues from GTEx as described in the previous section and a whole blood model from the DGN cohort presented in [10].

We used a Bonferroni threshold accounting for all the gene-tissue pairs that were tested (0.05/total number of gene-tissue pairs $\approx$ 2.5e-7). This approach is conservative because the correlation between tissues would make the total number of independent tests smaller than the total number of gene-tissue

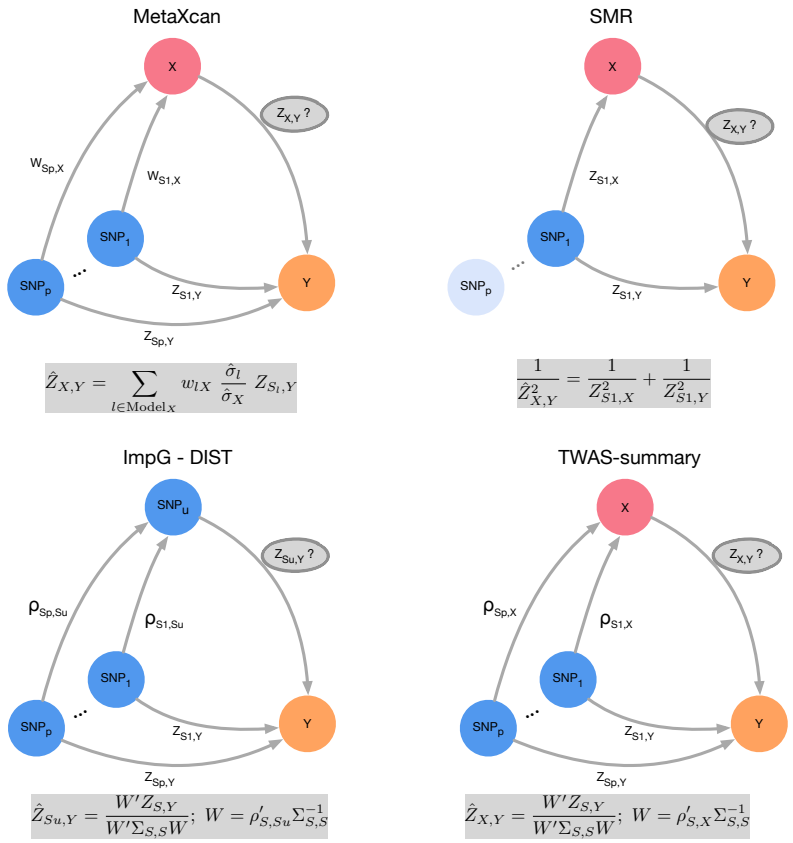


Figure 3. Comparison of MetaXcan with TWAS and SMR. All three methods seek to identify target genes by computing the strength of association between the unobserved predicted expression levels (X) of a gene with the complex trait (Y) quantified with $Z_{X,Y}$ or $Z^2_{X,Y}$. SMR uses the top eQTL (thus implicitly assumes single causal SNP). SMR uses the fact that $Z^2(\text{SMR})$ has an approximate χ^2 distribution when the eQTL association is much stronger than the GWAS association, i.e. $Z_{S_1,X}(\text{eQTL}) >> Z_{S_1,Y}(\text{GWAS})$. ImpG [20] and DIST [21] are precursors of TWAS-summary. These impute the summary results of unmeasured SNPs using a Gaussian imputation scheme. The weights are given by the Best Linear Unbiased Prediction (BLUP) formula. TWAS-summary simply extends this idea to imputed gene expression levels so that it inherits the Gaussian imputation scheme. In contrast, MetaXcan allows for general linear function of SNPs and can be considered a multivariate generalization of SMR and TWAS-Summary.

pairs. Height had the largest number of genes significantly associated with 1691 unique genes (based on a GWAMA of 250K individuals), followed by schizophrenia with 285 unique significant genes (n = 150K individuals), lipid levels, glycemic traits, and immune/inflammatory disorders such as rheumatoid arthritis and inflammatory bowel disease. Other Psychiatric phenotypes had much smaller number of significant genes with 8 significant genes for bipolar disorder (n = 16,731) and none for major depressive disorder (n = 18,759) probably due to smaller sample size but also smaller effect sizes.

Mostly, genome-wide significant genes tend to cluster around known 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 the combined effects of multiple variants, regions with sub-genome-wide significant SNPs can yield genome-wide significant results in MetaXcan. Supplementary Table 2 lists a few examples where this occurs.

As expected, results of MetaXcan tend to be more significant as the genetic component of gene expression increases (larger cross validated prediction performance R^2). Similarly, MetaXcan associations tend to be more significant when 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. All tissues and representative phenotypes are shown in Supplementary Figures 1-4. This trend was also robust to using different monotone functions of the Z-scores.

The full set of results can be queried in our online catalog (gene2pheno.org). This web application allows filtering the results by gene, phenotype, tissue, p value, and prediction performance . For each trait we assigned ontology terms from the Experimental Factor Ontology (EFO) [24] and Human Phenotype Ontology (HPO) [25], if applicable. As the catalog grows, the ontology annotation will facilitate the analysis by hierarchy of phenotypes. Supplementary Table 3 shows the list of consortia and phenotypes for which gene level association are available.

Disease associated genes in ClinVar also associated in MetaXcan

We verified that disease genes listed in ClinVar for obesity, rheumatoid arthritis, diabetes, Alzheimer's, Crohn's disease, ulcerative colitis, and age-related macular degeneration [26] show inflated significance among MetaXcan association results for the corresponding diseases. Schizophrenia and autism genes did not show enriched significance, which is not surprising given the highly polygenic nature of these phenotypes and consequently smaller effect sizes of these disease genes. Genes with small effect sizes are

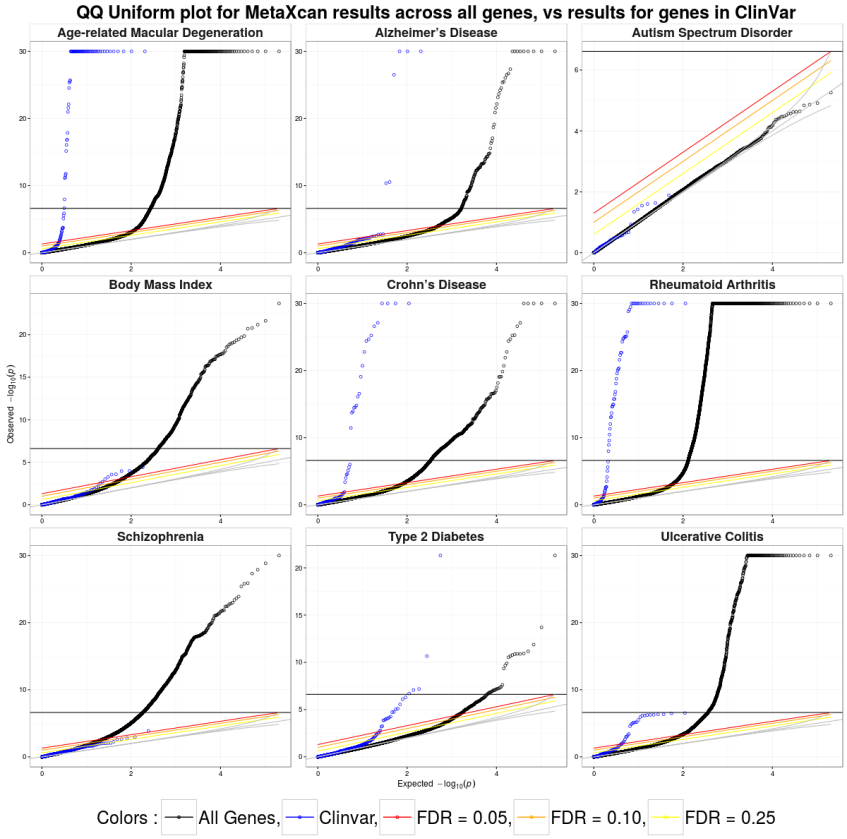


Figure 4. ClinVar genes show significant MetaXcan associations. Genes implicated in ClinVar tended to be more significant in MetaXcan for most diseases we tested, except for schizophrenia and autism. Blue circles correspond to qqplot of genes in ClinVar that were annotated with the phenotype and black correspond to all genes

.

likely to be underrepresented in ClinVar. The list of diseases in ClinVar used to generate the enrichment figures can be found in Supplementary Table 1, along with the corresponding association results.

Agnostic scanning of a broad set of tissues enabled by GTEx

The broad coverage of tissues in our prediction models enabled us to examine the tissue specificity of phenotypic consequences of GWAS signals. We started by computing average enrichment of significance by tissue. We used several measures of enrichment such as the mean Z-scores squared across all genes, or across significant genes for different thresholds, as well as the proportion of significant genes for different thresholds. We also compared the full distribution of the p-values of a given tissue relative to

the remaining tissues. Supplementary Figure 5 shows the average Z-score^2 as a measure of enrichment of each tissue by phenotype.

For low-density lipoprotein cholesterol (LDL-C) levels, liver was the most enriched tissue in significant associations as expected given known biology of this trait. This prominent role of liver was apparent despite the smaller sample size available for building liver models (n=97), which was less than a third of the numbers available for muscle (n=361) or lung (n=278). In general, however, expected tissues for diseases given currently known biology did not stand out as more enriched when we looked at the average across all (significant) genes using various measures of enrichment in our results. For example, the enrichment in liver was less apparent for high-density lipoprotein cholesterol (HDL-C) or triglyceride levels.

Next we focus on a few genes whose functional role has been well established: *C4A* for schizophrenia [27] and *SORT1* [28] and *PCSK9* both for LDL-C and cardiovascular disease. The MetaXcan results for these genes and traits and regulatory activity by tissue (as measured by the proportion of expression explained by the genetic component) are shown in Figure 5 and details can be found in Supplementary Tables 4, 5, and 6

There is strong evidence that *SORT1* has a causal role in LDL-C levels which is likely to affect risk for cardiovascular disease. This gene is most actively regulated in liver (close to 50% of the expression level of this gene is determined by the genetic component) with the most significant MetaXcan association in liver (p-value ≈ 0) consistent with our prior knowledge of lipid metabolism.

Other genes are found to be associated across multiple tissues. *C4A* is one example that was significantly associated with schizophrenia risk across all tissues ($p < 10^{-7}$ in 36 tissue models and $p<0.05$ for the remaining 4 tissue models) even though the effect of this gene is thought to be determined by excessive synaptic pruning in the brain during development.

PCSK9 is a target of several LDL cholesterol lowering drugs currently under trial to lower cardiovacular events [29]. Results from the STARNET study [30] profiled gene expression levels in cardiometabolic disease patients and showed that index SNP rs12740374 to be a strong eQTL for *PCSK9* in visceral fat but not in liver. Consistent with this our MetaXcan results also show highly significant association between *PCSK9* and LDL cholesterol ($p \approx 10^{-13}$) in visceral fat and not in liver (our training algorithm did not yield a prediction model for *PCSK9* i.e. there was no evidence of regulatory activity). In our results, however, the statistical evidence is much stronger in tibial nerve ($p \approx 10^{-27}$). The association between

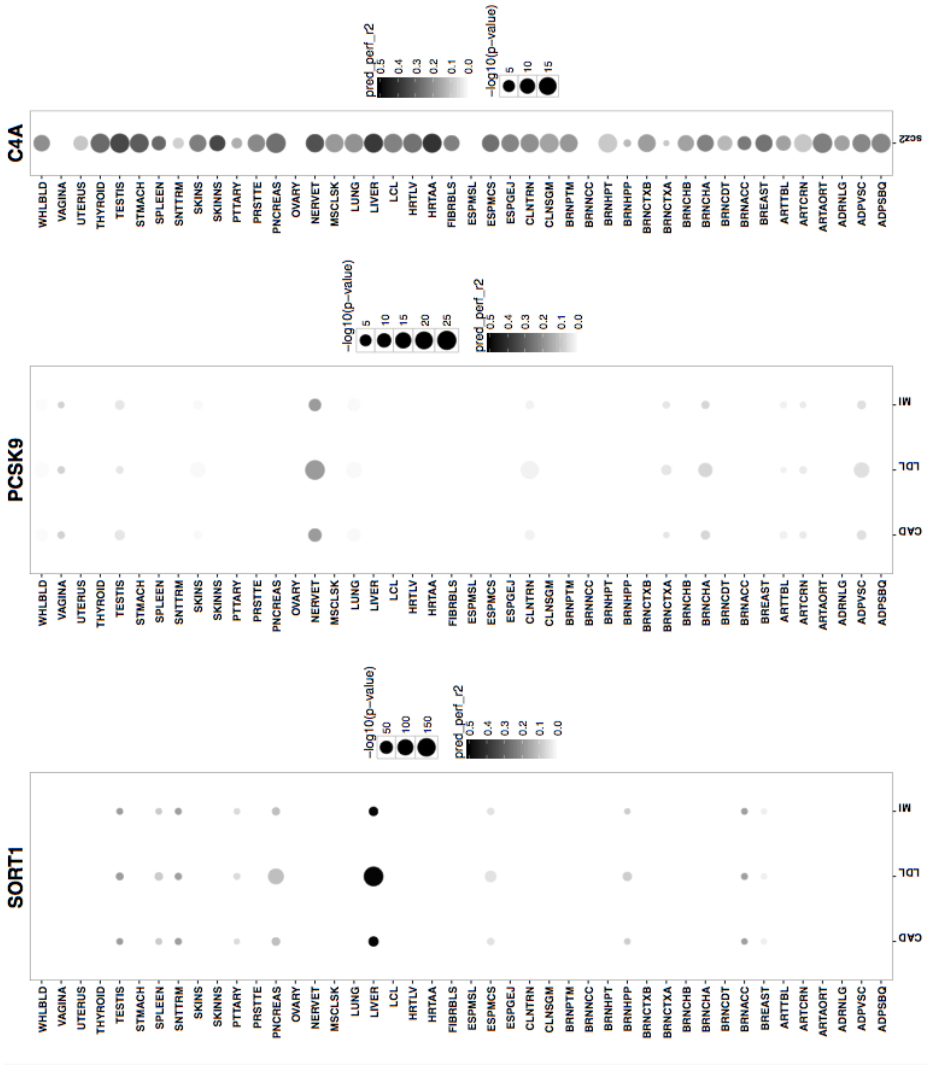


Figure 5. MetaXcan association for PCSK9, SORT1, and C4A. Pred.perf.R2 corresponds to the correlation squared between observed and predicted expression computed via cross validation in the training set. Darker points indicate larger genetic component and consequently more active regulation in the tissue. The size of the points represent the significance of the association between predicted expression and the traits indicated on the right side labels. C4A associations with schizophrenia (SCZ) is found across all tissues. SORT1 association with LDL-C, coronary artery disease (CAD), and myocardial infarction (MI) are most significant in liver. PCSK9 association LDL-C, coronary artery disease (CAD), and myocardial infarction (MI) are most significant in tibial nerve. Tissue abbreviation: Adipose - Subcutaneous (ADPSBQ), Adipose - Visceral (Omentum) (ADPVSC), Adrenal Gland (ADRNLG), Artery - Aorta (ARTAORT), Artery - Coronary (ARTCRN), Artery - Tibial (ARTTBL), Bladder (BLDDER), Brain - Amygdala (BRNAMY), Brain - Anterior cingulate cortex (BA24) (BRNACC), Brain - Caudate (basal ganglia) (BRNCDT), Brain - Cerebellar Hemisphere (BRNCHB), Brain - Cerebellum (BRNCHA), Brain - Cortex (BRNCTXA), Brain - Frontal Cortex (BA9) (BRNCTXB), Brain - Hippocampus (BRNHPP), Brain - Hypothalamus (BRNHPT), Brain - Nucleus accumbens (basal ganglia) (BRNNCC), Brain - Putamen (basal ganglia) (BRNPTM), Brain - Spinal cord (cervical c-1) (BRNSPC), Brain - Substantia nigra (BRNSNG), Breast - Mammary Tissue (BREAST), Cells - EBV-transformed lymphocytes (LCL), Cells - Transformed fibroblasts (FIBRBLS), Cervix - Ectocervix (CVXECT), Cervix - Endocervix (CVSEND), Colon - Sigmoid (CLNSGM), Colon - Transverse (CLNTRN), Esophagus - Gastroesophageal Junction (ESPGEJ), Esophagus - Mucosa (ESPMCS), Esophagus - Muscularis (ESPMSL), Fallopian Tube (FLLPNT), Heart - Atrial Appendage (HRTAA), Heart - Left Ventricle (HRTLV), Kidney - Cortex (KDNCTX), Liver (LIVER), Lung (LUNG), Minor Salivary Gland (SLVRYG), Muscle - Skeletal (MSCLSK), Nerve - Tibial (NERVET), Ovary (OVARY), Pancreas (PNCREAS), Pituitary (PTTARY), Prostate (PRSTTE), Skin - Not Sun Exposed (Suprapubic) (SKINNS), Skin - Sun Exposed (Lower leg) (SKINS), Small Intestine - Terminal Ileum (SNTTRM), Spleen (SPLEEN), Stomach (STMACH), Testis (TESTIS), Thyroid (THYROID), Uterus (UTERUS), Vagina (VAGINA), Whole Blood (WHLBLD).

PCSK9 and coronary artery disease is also significant in tibial nerve ($p \approx 10^{-8}$) but only nominally significant in visceral fat ($p \approx 0.02$). Accordingly, in our training set (GTEx), there is much stronger evidence of regulation of this gene in tibial nerve compared to visceral fat.

These examples show the importance of studying the regulation in a broad set of tissues and contexts and emphasize the challenges of determining causal tissues of complex traits based on in-silico analysis alone.

Discussion

Here we propose MetaXcan, a method that integrates genetically regulated components of molecular traits into large-scale GWAS results, to gain insight into the mechanisms that link genetic variation to phenotypic variation.

MetaXcan scales up the applicability of ideas behind PrediXcan and allows us to build a mechanism testing framework using prediction models of complex molecular processes and publicly available GWAMA summary results. Any molecular process that can be represented as linear functions of SNP variation can be encoded into prediction models which are in turn used to infer the phenotypic consequences via MetaXcan. These processes include, for example, expression levels of genes, intron usage, methylation status, telomere length, within different spatial, temporal, and developmental contexts.

As an example application of this framework, we trained transcriptome models in 44 human tissues from GTEx and estimated their effect on phenotypes from multiple publicly available GWAMA studies. We find known disease and trait associated genes active in relevant tissues but we also discover patterns of regulatory activity in tissues that are not traditionally associated with the trait. Further investigation of context and tissue specificity of these processes is needed but our results emphasize the importance of methods that integrate functional data across a broad set of tissues and contexts to augment our ability to identify novel target genes and provide mechanistic insight.

To facilitate broad adoption of the MetaXcan framework, we make user-friendly software and all pre-computed prediction models publicly available. We also host MetaXcan results for publicly available GWAMA results and make it freely available to the research community. This database lays the groundwork for a comprehensive catalog of phenome-wide consequences of complex molecular processes.

Software and Resources

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

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). w_{lg} are weights to predict expression $T_g = \sum_{l \in \text{Model}_g} w_{lg} X_l$, derived from an independent training set.

We model the phenotype as linear functions of X_l and T_g

$$Y = \alpha_1 + X_l \beta_l + \eta$$

$$Y = \alpha_2 + T_g \gamma_g + \epsilon,$$

where α_1 and α_2 are intercepts, η and ϵ error terms independent of X_l and T_g, respectively. Let $\hat{\gamma}_g$ and $\hat{\beta}_l$ be 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 $\widehat{\text{Var}}$ and $\widehat{\text{Cov}}$ the operators that compute the sample variance and covariance, i.e. $\widehat{\text{Var}}(Y) = \hat{\sigma}_Y^2 = \sum_{i=1,n} (Y_i - \bar{Y})^2/(n-1)$ with $\bar{Y} = \sum_{i=1,n} Y_i/n$. Let $\hat{\sigma}_l^2 = \widehat{\text{Var}}(X_l)$, $\hat{\sigma}_g^2 = \widehat{\text{Var}}(T_g)$ and $\Gamma_g = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})/n$, 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 ($\hat{\beta}_l$ and se), estimated variances of SNPs ($\hat{\sigma}_l^2$), estimated covariances between SNPs in each gene model (Γ_g), and prediction model weights w_{lg}.

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

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

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

and

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

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

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

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

$$\begin{aligned}
\hat{\sigma}_g^2 &= \widehat{\text{Var}}\left(\sum_{l \in \text{Model}_g} w_{lg} X_l\right) \\
&= \widehat{\text{Var}}(\mathbf{W}_g \mathbf{X}_g) && \text{where } \mathbf{W}_g \text{ is the vector of } w_{lg} \text{for SNPs in the model of } g \\
&= \mathbf{W}_g' \widehat{\text{Var}}(\mathbf{X}_g) \mathbf{W}_g && \text{where } \Gamma_g \text{ is the } \widehat{\text{Var}}(\mathbf{X}_g) = \text{ sample covariance matrix of } \mathbf{X}_g \\
&= \mathbf{W}_g' \Gamma_g \mathbf{W}_g
\end{aligned} \tag{4}$$

Calculation of regression coefficient $\hat{\gamma}_g$

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

$$\begin{aligned}
\hat{\gamma}_g &= \frac{\widehat{\mathrm{Cov}}(T_g, Y)}{\hat{\sigma}_g^2} \\
&= \frac{\widehat{\mathrm{Cov}}(\sum_{l \in \mathrm{Model}_g} w_{lg} X_l, Y)}{\hat{\sigma}_g^2} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \widehat{\mathrm{Cov}}(X_l, Y)}{\hat{\sigma}_g^2} && \text{by linearity of } \widehat{\mathrm{Cov}} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg} \hat{\beta}_l \sigma_l^2}{\hat{\sigma}_g^2} && \text{using Eq 3} \qquad (5)
\end{aligned}$$

Calculation of standard error of $\hat{\gamma}_g$

Also from the properties of linear regression we know that

$$\mathrm{se}(\hat{\gamma}_g) = \sqrt{\mathrm{Var}(\hat{\gamma}_g)} = \frac{\hat{\sigma}_\epsilon}{\sqrt{n \hat{\sigma}_g^2}} = \frac{\hat{\sigma}_Y^2 (1 - R_g^2)}{n \hat{\sigma}_g^2} \qquad (6)$$

In this equation, $\hat{\sigma}_Y^2 / n$ is not necessarily known but can be estimated using the analogous equation (6) for beta

$$\mathrm{se}(\hat{\beta}_l) = \frac{\hat{\sigma}_Y^2 (1 - R_l^2)}{n \hat{\sigma}_l^2} \qquad (7)$$

Thus

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

Notice that the right hand side of (8) 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 estimated effect size $\hat{\gamma}_g$ and standard error $\mathrm{se}(\gamma_g)$, or Z score,

$$Z_g = \frac{\hat{\gamma}_g}{\mathrm{se}(\hat{\gamma}_g)} \tag{9}$$

with which we can compute the p value as

$$p = 2\ \mathrm{pnorm}(-|Z_g|) \tag{10}$$

$$\begin{aligned}
Zg &= \frac{\hat{\gamma}_g}{\mathrm{se}(\hat{\gamma}_g)} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg}\hat{\beta}_l\hat{\sigma}_l^2}{\hat{\sigma}_g^2} \sqrt{\frac{n}{\hat{\sigma}_Y^2} \frac{\hat{\sigma}_g^2}{(1-R_g^2)}} && \text{using Eq. 5 and 6} \\
&= \sum_{l \in \mathrm{Model}_g} \frac{w_{lg}\hat{\beta}_l\hat{\sigma}_l^2}{\hat{\sigma}_g} \sqrt{\frac{(1-R_l^2)}{\mathrm{se}(\hat{\beta}_l)^2\hat{\sigma}_l^2}} \sqrt{\frac{1}{(1-R_g^2)}} \\
&= \sum_{l \in \mathrm{Model}_g} w_{lg}\ \frac{\hat{\sigma}_l}{\hat{\sigma}_g}\ \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)}\ \sqrt{\frac{1-R_l^2}{1-R_g^2}} && (11) \\
&\approx \sum_{l \in \mathrm{Model}_g} w_{lg}\ \frac{\hat{\sigma}_l}{\hat{\sigma}_g}\ \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} && (12)
\end{aligned}$$

Based on results with actual and simulated data for realistic effect size ranges, we have found that the last approximation does not affect our ability to identify the association. The approximation becomes inaccurate only when the effect sizes are very large. But in these cases, the small decrease in statistical efficiency induced by the approximation is compensated by the large power to detect the larger effect sizes.

Expression model training

To train our prediction models, we obtained genotype data and normalized gene expression data collected by the GTEx Project. We used 44 different tissues sampled by GTEx and thus generated 44 different

tissue-wide models. Sample sizes for different tissues range from 70 (Uterus) to 361 (Muscle - Skeletal). The models referenced in this paper make use of the GTEx ProjectâĂŹs V6p data, a patch to the version 6 data and makes use of improved gene-level annotation. To avoid foreseeable errors in future predictions, we removed ambiguously stranded SNPs from genotype data, i.e. ref/alt pairs A/T, C/G, T/A, G/C. Genotype data was filtered to include only SNPs with MAF > 0.01. For each tissue, normalized gene expression data was adjusted for covariates such as gender, sequencing platform, the top 3 principal components from genotype data and top PEER Factors. The number of PEER Factors used was determined by sample size: 15 for n $<$ 150, 30 for n between 150 and 250, and 35 for n $>$ 250 Covariate data was provided by GTEx. For our analysis, we used protein-coding genes only.

For each gene-tissue pair for which we had adjusted expression data, we fit an Elastic-Net model based on the genotypes of the samples for the SNPs located within 1 Mb upstream of the geneâĂŹs transcription start site and 1 Mb downstream of the transcription end site. We used the R package glmnet with mixing parameter alpha equal to 0.5, and the penalty parameter lambda was chosen through 10-fold cross-validation.

Once we fit all models, we retained only the models which reached significance at a False Discovery Rate of less than 0.05. For each tissue examined, we created a sqlite database to store the weights of the prediction models, as well as other statistics regarding model training. These databases have been made available for download at PredictDB.org.

Comparison with TWAS

Formal similarity with TWAS can be made more explicit by rewriting MetaXcan formula in matrix form. With the following notation and definitions

$$\begin{aligned}
\tilde{\mathbf{W}}_g &= (\sigma_1 w_{1g}, ..., \sigma_p w_{pg})' \\
\mathbf{Z}_{\text{SNPs}} &= (Z_1, ..., Z_p)' \\
&= (\frac{\hat{\beta}_1}{\text{se}(\beta_1)}, ..., \frac{\hat{\beta}_p}{\text{se}(\beta_p)})'
\end{aligned}$$

and correlation matrix of SNPs in the model for gene g

$$\Sigma_g = \text{diag}(\frac{1}{\sigma_1}, ..., \frac{1}{\sigma_p}) \cdot \Gamma_g \cdot \text{diag}(\frac{1}{\sigma_1}, ..., \frac{1}{\sigma_p})$$

it is quite straightforward to write the numerator in 1 and 12 as

$$\tilde{\mathbf{W}}_g \cdot \mathbf{Z}_{\text{SNPs}}$$

and the denominator, the variance of the predicted expression level of gene g, as

$$\tilde{\mathbf{W}}_g' \cdot \Sigma_g \cdot \tilde{\mathbf{W}}_g$$

thus

$$Z_g = \frac{\tilde{\mathbf{W}}_g \cdot \mathbf{Z}_{\text{SNPs}}}{\tilde{\mathbf{W}}_g' \cdot \Sigma_g \cdot \tilde{\mathbf{W}}_g}$$

This equation has the same form as the TWAS expression but because TWAS was derived as an extension of DIST or ImpG, which impute unmeasured SNPs results, it was restricted to Gaussian imputation type of prediction or equivalently Best Linear Unbiased Prediction (BLUP) also equivalent to ridge regression-based prediction [19]. In our derivation of MetaXcan we demonstrated that there is no need to restrict the prediction models to Gaussian imputation type, i.e. ridge regression prediction, but more flexible and better performing linear models can be used. We have also quantified the approximation underlying this approach, which is given by the factor $\sqrt{\frac{1-R_l^2}{1-R_g^2}}$.

Comparison with SMR

SMR quantifies the strength of the association between expression levels of a gene and complex traits with T_{SMR} using the following function of the eQTL and GWAS Z-score statistics.

$$T_{\text{SMR}} = \frac{Z_{\text{eqtl}}^2 Z_{\text{GWAS}}^2}{Z_{\text{eqtl}}^2 + Z_{\text{GWAS}}^2} \tag{13}$$

which can be expressed as

$$\frac{1}{T_{\text{SMR}}} = \frac{1}{Z_{\text{eqtl}}^2} + \frac{1}{Z_{\text{GWAS}}^2} \tag{14}$$

Here 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 χ^2 (Z is asymptotically normally, Z2 is chisquare, 1/Z2 is inverse chisquare). This approximation only holds if the significance of the eQTL is very large so that only the GWAS Z-score contributes to the statistics. So within the range where this approximation is valid

$$\frac{1}{T_{\text{SMR}}} \approx \frac{1}{Z_{\text{GWAS}}^2}$$

thus

$$T_{\text{SMR}} \approx Z_{\text{GWAS}}^2$$

On the other hand the MetaXcan formula when only the top eQTL is used to predict the expression level of a gene is

$$\begin{aligned} Z_{\text{MetaXcan}} &= \sum_{l \in \text{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)} && \text{using 12} \\ &= w_{1g} \frac{\hat{\sigma}_1}{\sqrt{w_{1g}^2 \sigma_1^2}} Z_1 && \text{only top eQTL is in the model} \\ &= Z_1 \end{aligned}$$

where Z_1 is the GWAS Z-score of the top eQTL in the model for gene. Thus

$$Z_{\text{MetaXcan}}^2 = Z_{\text{GWAS}}^2$$

SMR will work well if there is no allelic heterogeneity.

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.

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

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

4. 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.

5. 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.

6. 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`.

7. 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`.

8. 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).

9. 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`.

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

11. 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.

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

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

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

15. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature genetics. 2016;48(5):481–7. Available from: http://www.nature.com/doifinder/10.1038/ng.3538$\delimiter"026E30F$nhttp://www.ncbi.nlm.nih.gov/pubmed/27019110.

16. Casella G, Berger R. Statistical Inference. 2nd ed. Imprint Australia; Pacific Grove, CA : Thomson Learning, c2002.; 2002.

17. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, GTEx Consortium, 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.

18. 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).

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

20. 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.

21. Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu SA. DIST: Direct imputation of summary statistics for unmeasured SNPs. Bioinformatics. 2013;29(22):2925–2927.

22. Wen X, Stephens M. Using linear predictors to impute allele frequencies from summary or pooled genotype data. The Annals of Applied Statistics. 2010;4(3):1158–1182. Available from: `http://projecteuclid.org/euclid.aoas/1287409368`.

23. Zou H, Hastie T. Regularization and variable selection via the elastic-net. Journal of the Royal Statistical Society. 2005;67:301–320.

24. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010;26(8):1112–1118.

25. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. Nucleic Acids Research. 2014;42(D1).

26. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research. 2015;44(D1):D862–8. Available from: `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4702865{&}tool=pmcentrez{&}rendertype=abstract`.

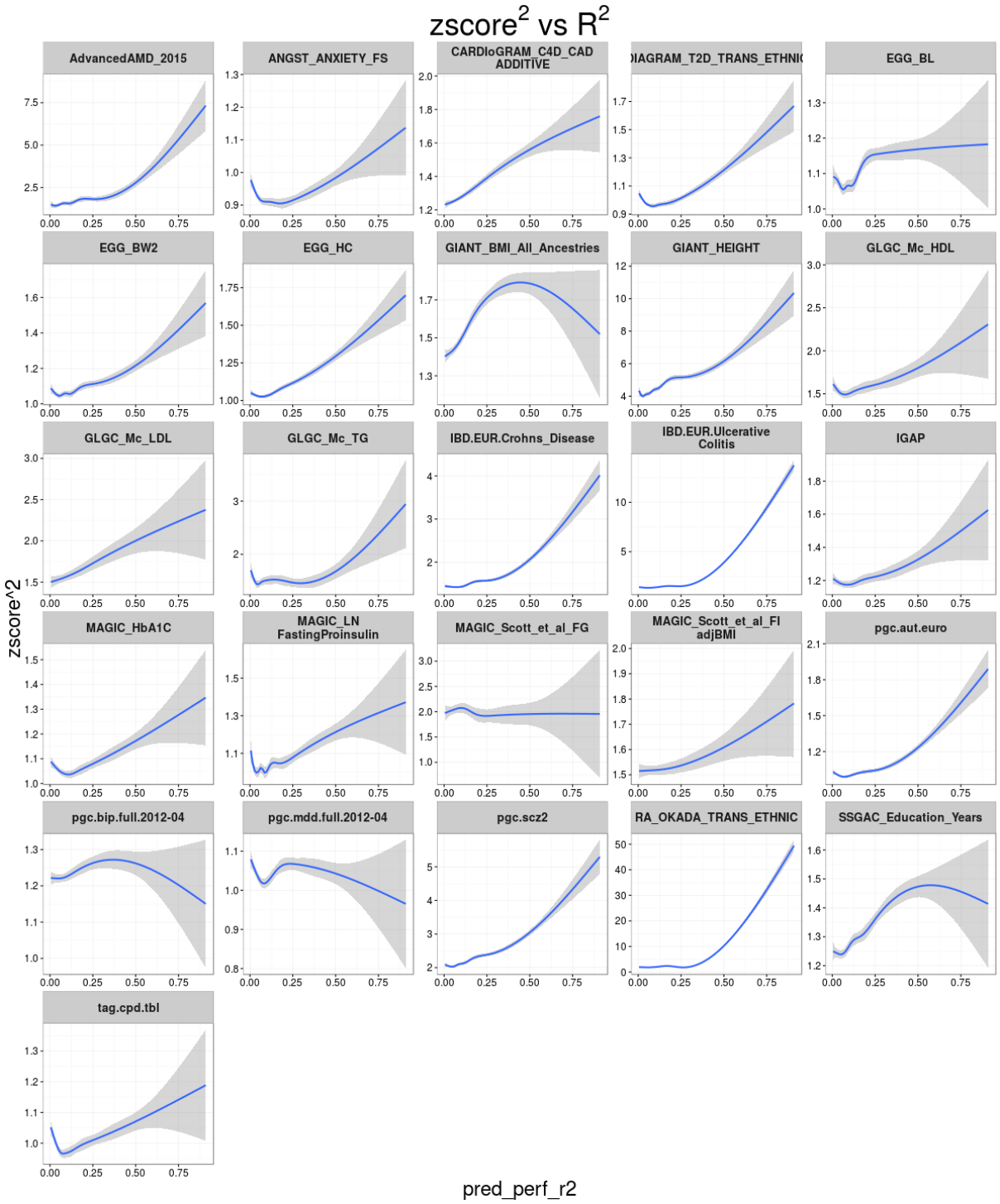
27. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016 Jan;530(7589):1–17. Available from: `http://dx.doi.org/10.1038/nature16549`.

28. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature. 2010 May;466(7307):714–719. Available from: `http://dx.doi.org/10.1038/nature09266`.

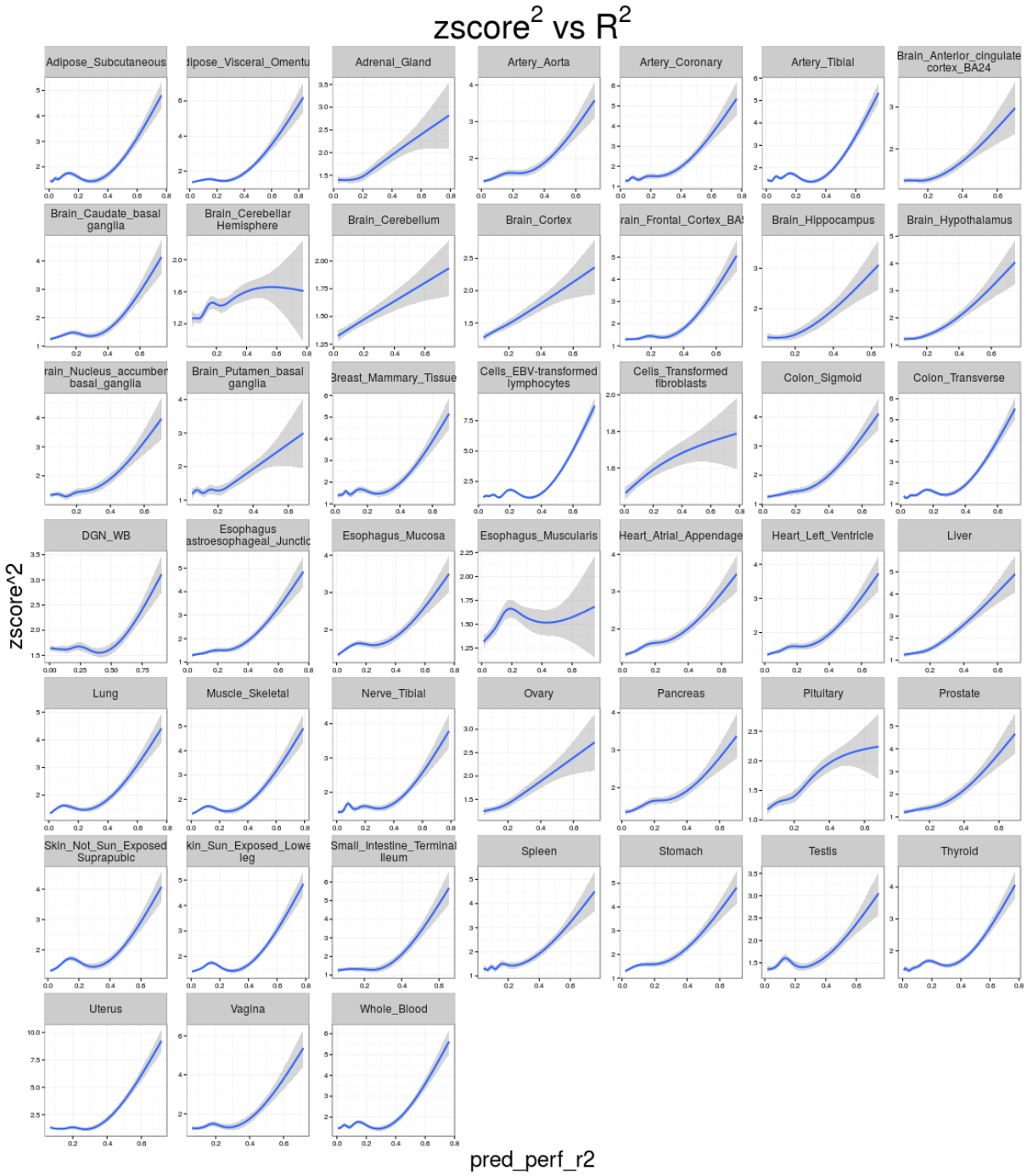
29. Dadu RT, Ballantyne CM. Lipid lowering with PCSK9 inhibitors. Nature Publishing Group. 2014 Jun;11(10):563–575. Available from: http://dx.doi.org/10.1038/nrcardio.2014.84.

30. Franzén O, Ermel R, Cohain A, Akers NK, Di Narzo A, Talukdar HA, et al. Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. Science. 2016 Aug;353(6301):827–830. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.aad6970.

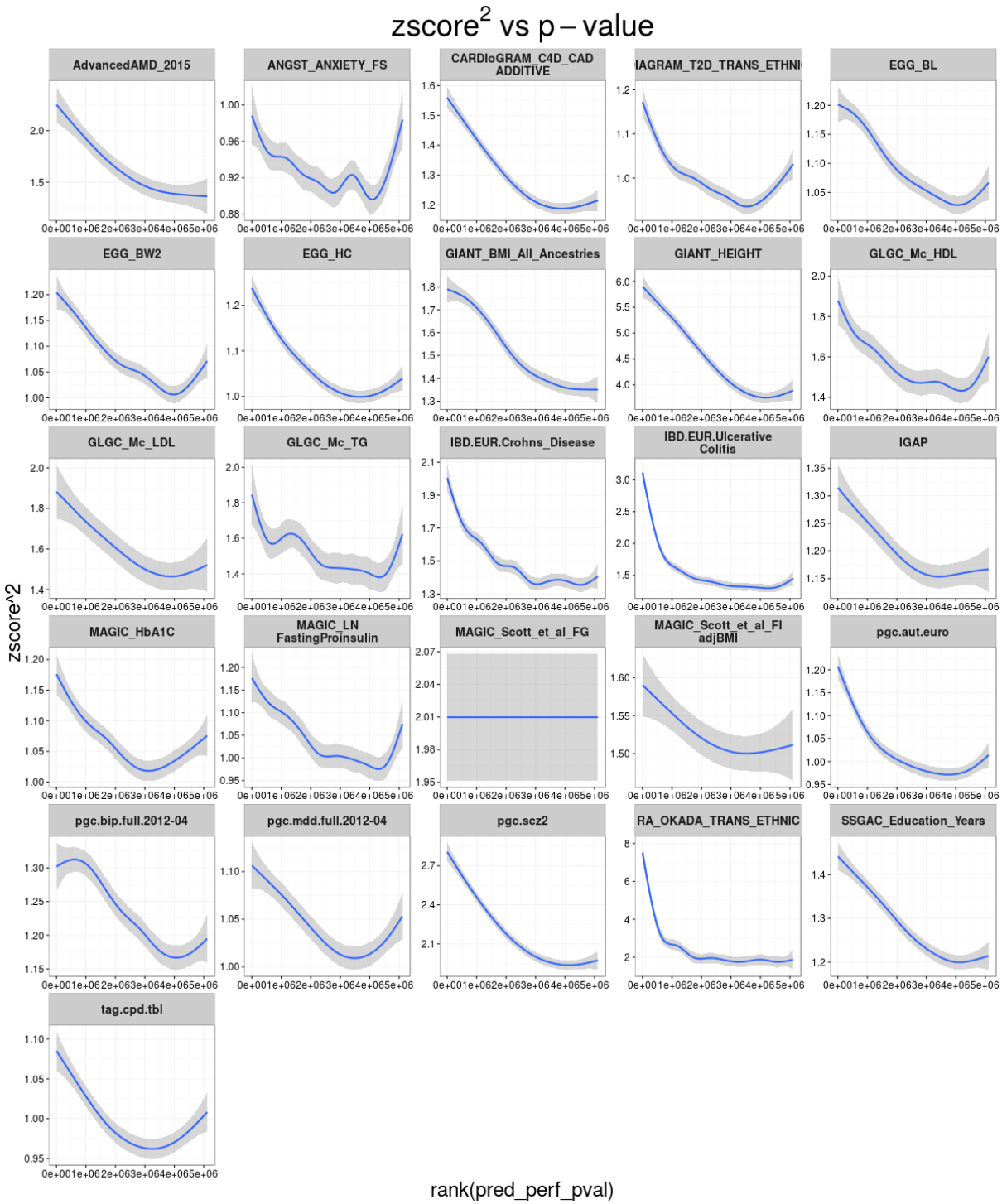
Supplementary Material



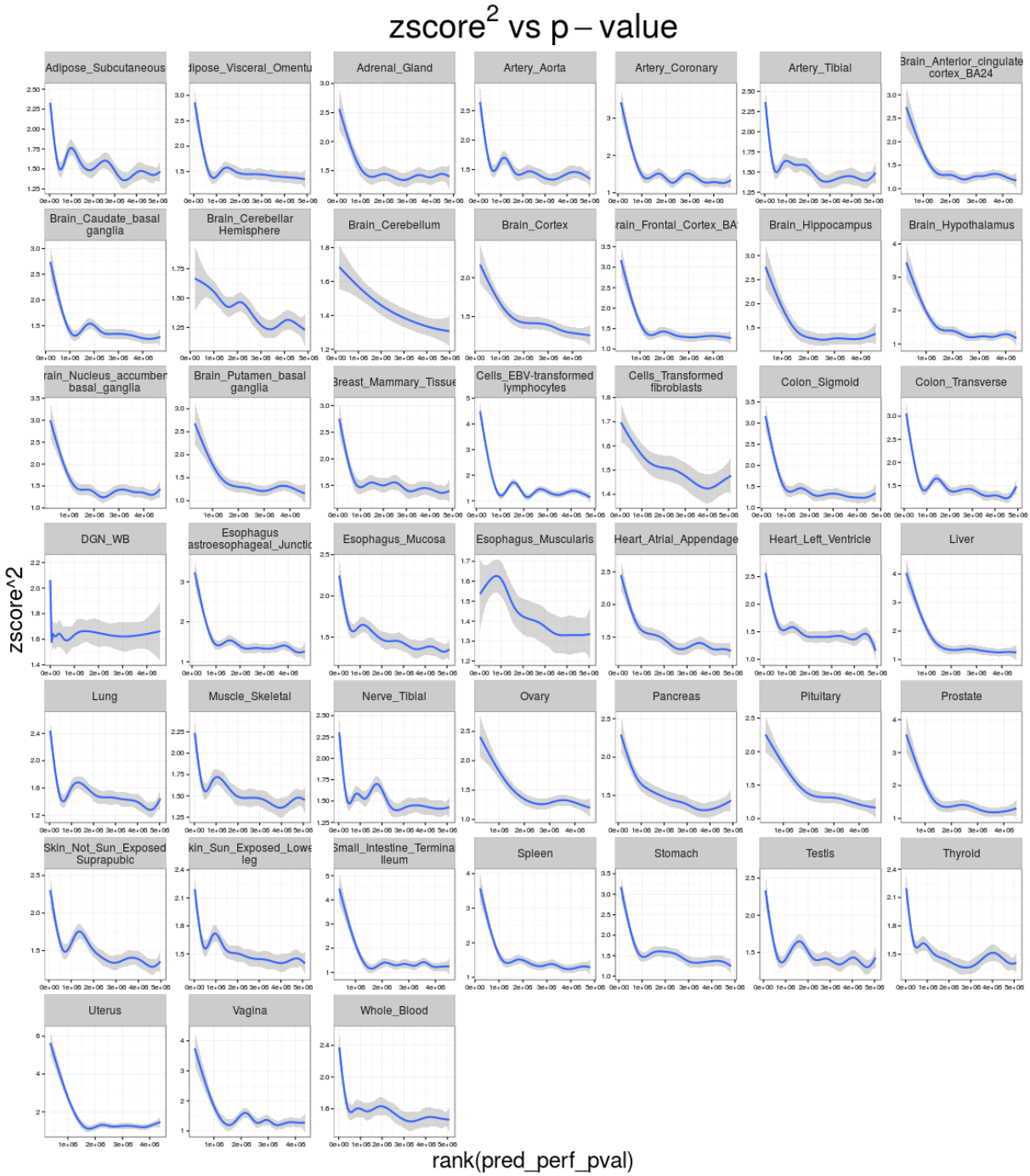
Supplementary Figure 1. Z-score^2 vs predicted performance R^2 by phenotype. When averaged across all genes and tissues within each phenotype the significance of the association tends to be more pronounced as the genetic component is larger.



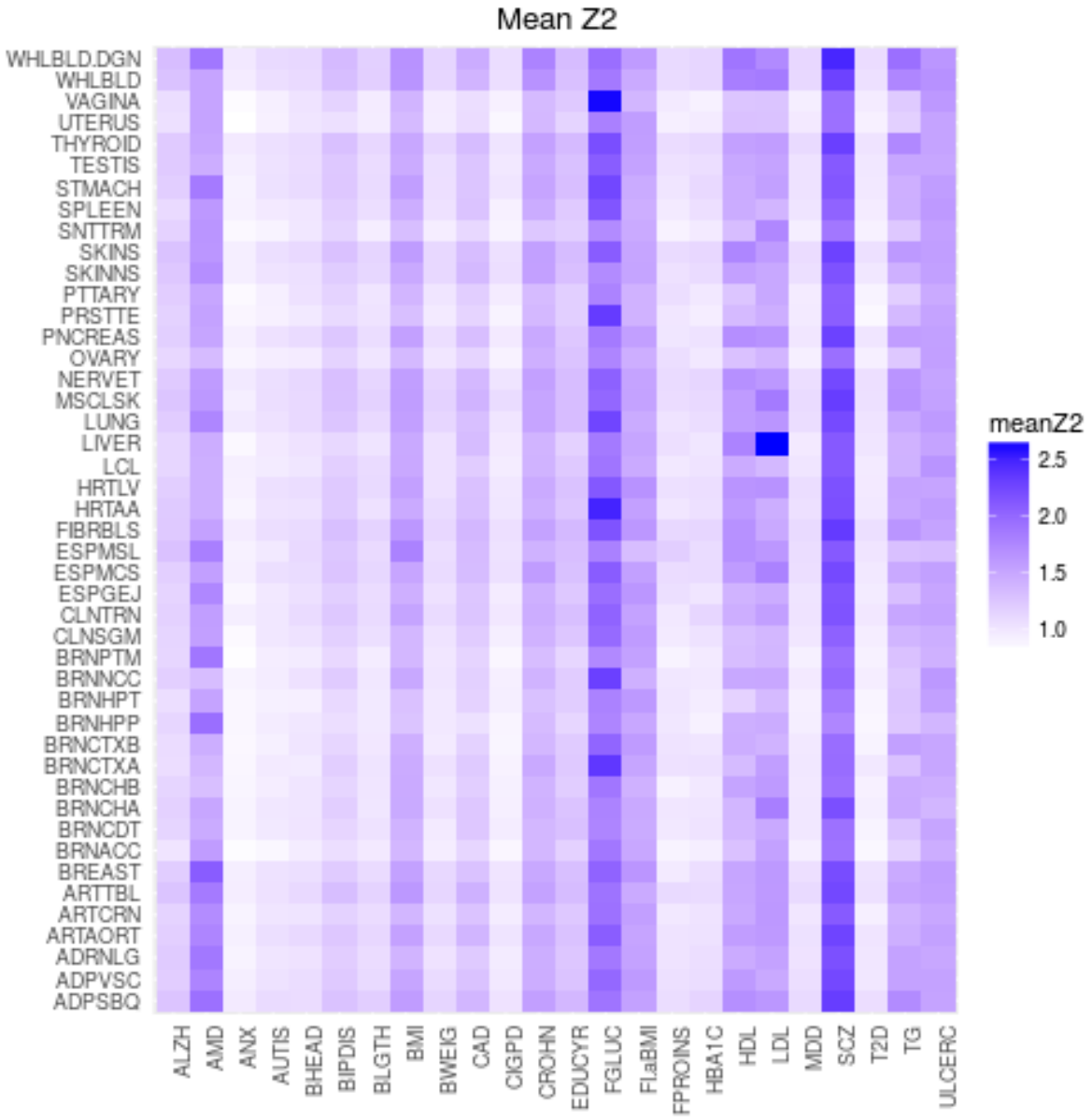
Supplementary Figure 2. Z-score^2 vs predicted performance R^2 by tissue. When averaged across all genes and phenotypes within each tissue the significance of the association tends to be more pronounced as the genetic component is larger.



Supplementary Figure 3. Z-score^2 vs predicted performance p-value by phenotype. When averaged across all genes and tissues within each phenotype the significance of the association tends to be more pronounced as the cross validated prediction is more significantly associated with the observed expression.



Supplementary Figure 4. Z-score^2 vs predicted performance p-value by tissue. When averaged across all genes and phenotypes within each tissue the significance of the association tends to be more pronounced as the cross validated prediction is more significantly associated with the observed expression.



Supplementary Figure 5. Average enrichment of significant genes by tissues. This figure shows the average square of the zscores (effect size/standard error) of the association between the genetic component of gene expression levels and phenotype. CIGPD (cigarettes per day), BMI (body mass index), FGLUC (fasting glucose), T2D (type 2 diabetes), CAD (coronary artery disease), LDL (low-density lipoprotein cholesterol), TG (triglycerides), RA (rheumatoid arthritis), ALZH (alzheimer's disease), HDL (high-density lipoprotein cholesterol), CROHN (Crohn's disease), ULCERC (ulcerative colitis), HEIGHT, BHEAD (birth head), BLGTH (birth length), BWEIG (birth weight), AUTIS (autism), EDUCYR (education years), SCZ (schizophrenia), AMD (age-related macular degeneration), ANX (anxiety), HBA1C (Hemogoblin A1C), FPROINS (fasting proinsulin), FI.aBMI (fasting insuline adjusted for BMI), MDD (major depressive disorder), BIPDIS (bipolar disorder).

Supplementary Tables

Supplementary Table 1. ClinVar genes with significant association in MetaXcan. Data included in clinvar_enrichment.txt

Supplementary Table 2. MetaXcan Association yields results more significant than Top SNPs. Genes associated by MetaXcan to Coronary Artery Disease GWAS where MetaXcan outperforms individual SNPs in a 2 Mb window around the gene.

Gene Name	Tissue	Pval	Top SNP in Region	Top SNP PValue
TUBG2	Adipose Visceral Omentum	2.34E-07	rs72823056	1.50E-06
UTP11L	Artery Tibial	1.58E-07	rs28470722	9.84E-07
FHL3	Skin Sun Exposed Lower leg	1.99E-07	rs28470722	9.84E-07
IP6K2	Adipose Subcutaneous	2.14E-07	rs7623687	5.22E-07
SNF8	Thyroid	2.20E-07	rs35895680	3.76E-07
CCDC97	Adipose Subcutaneous	1.31E-09	chr19:41790086:D	1.75E-07
FURIN	Artery Aorta	1.27E-08	rs2521501	5.01E-08
FES	Cells Transformed fibroblasts	1.23E-08	rs2521501	5.01E-08
PCSK9	Nerve Tibial	1.04E-08	rs11206510	2.34E-08
SWAP70	Spleen	1.01E-08	rs10840293	1.28E-08
NT5C2	Testis	3.79E-09	rs11191416	4.65E-09
IL6R	Colon Transverse	2.31E-10	rs6689306	2.60E-09
TCF21	Nerve Tibial	7.19E-12	rs12202017	1.98E-11
LIPA	Whole Blood	1.67E-14	rs1412444	5.15E-12

Supplementary Table 3. List of Genome-wide Association Meta Analysis (GWAMA) Consortia and phenotypes. # Signif. column lists the number of gene/tissue pairs that were found to be significant after bonferroni correction for total number of tests ($\approx 2.5\ 10^{-7}$). # Unique Sig. column lists the number of unique genes among the significant gene/tissue pairs.

Consortium	Phenotype	Sample Size	# Signif	# Unique Sig.
PGC	Attention Deficit/Hyperactivity Disorder	5415	0	0
PGC	Bipolar Disorder	16731	13	8
PGC	Major Depressive Disorder	18759	1	1
PGC	Schizophrenia	150064	1122	305
CIAC	Clozapine-Induced Agranulocytosis	1352	0	0
CONVERGE	Major Depressive Disorder	11670	0	0
IGAP	Alzheimer	54162	124	55
TAG	Tobacco Cigarettes per Day	38181	23	13
IBD	Inflammatory Bowel Disease	20833	1052	230
IBD	Ulcerative Colitis	27432	565	123
IBD	Chron's Disease	34652	607	164
GIANT	Body Mass Index	339224	508	129
GIANT	Waist-to-Hip Ratio	142762	18	15
GIANT	Waist Circumference	224459	102	30
GIANT	Hip Circumference	224459	373	83
GIANT	Height	253288	5840	1685
CARDIoGRAM C4D	Coronary Artery Disease (additive model)	184305	136	57
CARDIoGRAM C4D	Myocardial Infarction	184305	80	32
MAGIC	Fasting Glucose	133010	542	262
MAGIC	Fasting Insulin /Adjusted for BMI)	108557	102	72
MAGIC	Fasting Glucose and BMI interaction	58074	10	2
MAGIC	Fasting Proinsulin	10701	187	43
MAGIC	Glycated Hemogoblin	46368	69	21
SSGAC	College Completion	101069	5	3
SSGAC	Education Years	95427	20	10
EGG	Birth Length	28459	7	2
EGG	Birth Weight	26836	9	8
EGG	Childhood Obesity	13848	13	6
EGG	Head Circumference	10678	0	0
GLGC	Low-density Lipoprotein Cholesterol	188578	825	296
GLGC	High-density Lipoprotein Cholesterol	188578	822	262
GLGC	Triglicerids	188578	709	245
GPC	Five Factor Model Conscientiousness	17375	0	0
ANGST	Anxiety Disorder (Factor Score)	18186	0	0
AMD	Age-related Macular Degeneration	34009	754	152
DIAGRAM	Type II diabetes	149461	33	19
RA	Rheumatoid Arthritis	80799	1567	242

Supplementary Table 4. MetaXcan association results for *SORT1*. Association with LDL cholesterol, coronary artery disease, and myocardial infarction are shown for available tissue models. Liver shows the most significant association with all three phenotypes. Also liver is the tissue with the most active regulation of *SORT1* expression, with 49% of the expression explained by our genetic prediction model. This is expected given the importance of this tissue in liver metabolism and its mediating effect on cardiovascular disease. Pvalue is the significance of the association between predicted expression levels and the phenotype. Effect size is the change in the phenotype when there is a change of 1 standard deviation in the predicted expression. Pred.Perf.R2 column is the cross validated R^2 in the training set between observed and predicted expression level. This can also be interpreted as a lower bound of the heritability of the expression trait. Pred.Perf.Pvalues is the p values of the correlation between predicted and observed expression. Note that tissue models will be available only when regulation was sufficiently active to yield a significant genetic component for the gene. Full set of results can be queried in `gene2pheno.org`

Gene	Phenotype	Effect Size	Pvalue	Tissue	Pred.Perf.R2	Pred.Perf.Pvalue
SORT1	CAD	-0.058	1.28E-17	Liver	0.492	1.18E-15
		-0.035	3.58E-07	Pancreas	0.114	2.52E-05
		-0.023	9.29E-04	DGN WB	0.017	8.27E-05
		-0.018	8.73E-03	Esophagus Mucosa	0.051	4.08E-04
		0.014	0.06	Small Intestine Terminal Ileum	0.173	1.70E-04
		-0.014	0.15	Spleen	0.092	3.89E-03
		-0.007	0.24	Testis	0.178	3.83E-08
		0.005	0.54	Brain Hippocampus	0.087	7.48E-03
		-0.002	0.58	Brain Anterior cingulate cortex BA24	0.172	2.93E-04
		0.001	0.89	Breast Mammary Tissue	0.028	0.024
		0.001	0.99	Pituitary	0.064	0.018
SORT1	LDL-C	-0.085	7.45E-183	Liver	0.492	1.18E-15
		-0.058	6.48E-96	Pancreas	0.114	2.52E-05
		-0.030	2.92E-31	Esophagus Mucosa	0.051	4.08E-04
		-0.031	2.76E-27	DGN WB	0.017	8.27E-05
		0.020	5.91E-11	Brain Hippocampus	0.087	7.48E-03
		-0.017	3.56E-06	Spleen	0.092	3.89E-03
		-0.011	5.16E-04	Testis	0.178	3.83E-08
		0.008	0.03	Small Intestine Terminal Ileum	0.173	1.70E-04
		-0.003	0.15	Brain Anterior cingulate cortex BA24	0.172	2.93E-04
		-0.003	0.20	Pituitary	0.064	1.79E-02
		-0.001	0.84	Breast Mammary Tissue	0.028	0.02
SORT1	Myocardial Infarction	-0.051	5.17E-12	Liver	0.492	1.18E-15
		-0.031	4.41E-05	Pancreas	0.114	2.52E-05
		-0.020	0.01	DGN WB	0.017	8.27E-05
		-0.016	0.03	Esophagus Mucosa	0.051	4.08E-04
		0.012	0.12	Small Intestine Terminal Ileum	0.173	1.70E-04
		-0.008	0.24	Testis	0.178	3.83E-08
		-0.007	0.29	Brain Anterior cingulate cortex BA24	0.172	2.93E-04
		-0.012	0.33	Spleen	0.092	3.89E-03
		0.006	0.53	Pituitary	0.064	0.018
		0.002	0.82	Brain Hippocampus	0.087	7.48E-03
		-0.001	0.93	Breast Mammary Tissue	0.028	0.024

Supplementary Table 5. MetaXcan association between *C4A* and schizophrenia MetaXcan association with schizophrenia for available tissue models. *C4A* is actively regulated across all tissues, with prediction R^2 ranging from 8% to 39%. Predicted expression levels of *C4A* are also significantly associated with schizophrenia risk uniformly across all tissues. Pvalue is the significance of the association between predicted expression levels and the phenotype. Effect size is the change in the phenotype when there is a change of 1 standard deviation in the predicted expression. Pred.Perf.R2 column is the cross validated R^2 in the training set between observed and predicted expression level. This can also be interpreted as a lower bound of the heritability of the expression trait. Pred.Perf.Pvalues is the p values of the correlation between predicted and observed expression. Note that tissue models will be available only when regulation was sufficiently active to yield a significant genetic component for the gene. Full set of results can be queried in `gene2pheno.org`

Gene	Phenotype	Effect Size	Pvalue	Tissue	Pred.Perf.R2	Pred.Perf.Pvalue
C4A	Schizophrenia	0.072	2.29E-20	Pancreas	0.266	1.66E-11
		0.069	7.72E-20	Artery Aorta	0.234	6.11E-13
		0.060	1.46E-19	Testis	0.347	4.61E-16
		0.065	2.61E-19	Thyroid	0.276	3.65E-21
		0.066	6.78E-19	Heart Atrial Appendage	0.394	8.55E-19
		0.062	8.45E-19	Adipose Subcutaneous	0.221	8.78E-18
		0.064	9.25E-19	Colon Sigmoid	0.165	2.94E-06
		0.068	1.01E-18	Heart Left Ventricle	0.260	5.60E-14
		0.070	1.15E-18	Liver	0.380	1.86E-11
		0.069	1.98E-18	Cells EBV-transformed lymphocytes	0.229	7.29E-08
		0.062	2.23E-18	Stomach	0.305	6.18E-15
		0.067	3.54E-18	Brain Hypothalamus	0.094	5.25E-03
		0.058	1.02E-17	Lung	0.196	8.29E-15
		0.061	2.70E-17	Colon Transverse	0.202	8.55E-10
		0.062	3.73E-17	Muscle Skeletal	0.184	1.29E-17
		0.056	4.74E-17	Nerve Tibial	0.327	1.20E-23
		0.061	9.07E-17	Adipose Visceral Omentum	0.219	1.81E-11
		0.053	3.64E-16	Brain Putamen basal ganglia	0.188	4.61E-05
		0.057	3.99E-16	Artery Coronary	0.099	5.20E-04
		0.063	1.16E-15	Brain Frontal Cortex BA9	0.176	3.10E-05
		0.068	1.74E-15	Esophagus Gastroesophageal Junction	0.224	2.00E-08
		0.050	4.11E-15	Prostate	0.215	6.22E-06
		0.057	6.16E-15	Esophagus Mucosa	0.263	1.48E-17
		0.059	1.53E-14	Breast Mammary Tissue	0.257	2.45E-13
		0.054	2.55E-14	Skin Sun Exposed Lower leg	0.237	2.33E-19
		0.056	1.18E-13	Brain Cerebellum	0.231	2.76E-07
		0.051	6.32E-13	Whole Blood	0.205	1.89E-18
		0.050	2.20E-12	Brain Cerebellar Hemisphere	0.166	7.33E-05
		0.054	4.34E-12	Skin Not Sun Exposed Suprapubic	0.350	7.11E-20
		0.046	1.01E-11	Cells Transformed fibroblasts	0.233	2.81E-17
		0.051	6.82E-11	Adrenal Gland	0.169	1.76E-06
		0.038	3.19E-10	Artery Tibial	0.169	5.16E-13
		0.042	5.71E-10	Brain Caudate basal ganglia	0.126	2.92E-04
		0.045	1.09E-09	Uterus	0.101	7.20E-03
		0.045	5.37E-09	Spleen	0.280	1.00E-07
		0.037	2.95E-08	Brain Anterior cingulate cortex BA24	0.246	9.32E-06
		0.032	1.09E-04	Small Intestine Terminal Ileum	0.083	1.09E-02
		0.023	2.61E-04	Pituitary	0.145	2.76E-04
		0.015	0.02	Brain Hippocampus	0.125	1.21E-03
		0.016	0.03	Brain Cortex	0.099	1.80E-03

Supplementary Table 6. MetaXcan association results for *PCSK9*. Association with LDL cholesterol, coronary artery disease, and myocardial infarction are shown for available tissue models. The significant association between LDL-C and PCSK9 in visceral fat is consistent with other reports [30] but the most significant association is found in tibial nerve. Tibial nerve was the most actively regulated tissue with 18% of the expression level of the gene being explained by our genetic prediction model (cross validated). Pvalue is the significance of the association between predicted expression levels and the phenotype. Effect size is the change in the phenotype when there is a change of 1 standard deviation in the predicted expression. Pred.Perf.R2 column is the cross validated R^2 in the training set between observed and predicted expression level. This can also be interpreted as a lower bound of the heritability of the expression trait. Pred.Perf.Pvalues is the p values of the correlation between predicted and observed expression. Note that tissue models will be available only when regulation was sufficiently active to yield a significant genetic component for the gene. Full set of results can be queried in `gene2pheno.org`

Gene	Phenotype	Effect Size	Pvalue	Tissue	Pred.Perf.R2	Pred.Perf.Pvalue
PCSK9	CAD	0.039	1.04E-08	Nerve Tibial	0.179	1.48E-12
		0.037	4.14E-07	Lung	0.010	0.102
		0.029	4.60E-05	Whole Blood	0.007	0.119
		0.020	4.53E-03	Testis	0.043	9.05E-03
		-0.017	0.01	Colon Transverse	0.022	0.054
		0.015	0.02	Adipose Visceral Omentum	0.056	1.13E-03
		0.014	0.04	Brain Cerebellum	0.072	6.09E-03
		-0.008	0.23	Skin Sun Exposed Lower leg	0.010	0.080
		0.007	0.35	Artery Tibial	0.023	0.010
		-0.006	0.48	Vagina	0.079	0.012
		-0.003	0.60	Artery Coronary	0.036	0.038
		-0.001	0.95	Brain Cortex	0.044	0.039
PCSK9	LDL-C	0.039	1.45E-27	Nerve Tibial	0.179	1.48E-12
		-0.027	5.05E-21	Colon Transverse	0.022	0.054
		0.032	2.19E-13	Lung	0.010	0.102
		0.030	2.41E-13	Adipose Visceral Omentum	0.056	1.13E-03
		-0.021	4.27E-12	Skin Sun Exposed Lower leg	0.010	0.080
		0.019	1.05E-10	Whole Blood	0.007	0.119
		0.020	2.48E-10	Brain Cerebellum	0.072	6.09E-03
		0.010	3.65E-03	Brain Cortex	0.044	0.039
		-0.005	0.44	Vagina	0.079	0.012
		0.004	0.47	Testis	0.043	9.05E-03
		-0.002	0.51	Artery Coronary	0.036	0.038
		0.002	0.99	Artery Tibial	0.023	0.010
PCSK9	Myocardial	0.037	8.62E-07	Nerve Tibial	0.179	1.48E-12
		0.035	7.89E-06	Lung	0.010	0.102
		0.029	3.05E-04	Whole Blood	0.007	0.119
		0.018	0.02	Testis	0.043	9.05E-03
		-0.013	0.08	Skin Sun Exposed Lower leg	0.010	0.080
		0.012	0.09	Adipose Visceral Omentum	0.056	1.13E-03
		-0.011	0.15	Colon Transverse	0.022	0.054
		0.009	0.25	Brain Cerebellum	0.072	6.09E-03
		-0.004	0.58	Brain Cortex	0.044	0.039
		0.003	0.72	Artery Tibial	0.023	0.010
		-0.003	0.85	Vagina	0.079	0.012
		0.001	0.90	Artery Coronary	0.036	0.038

Supplementary Table 7. List of Tissue Models. # protein cod. lists the number of genes in the training set for the tissue, # **samples** is the samples available with expression and genotype data, # **signif. models** lists the number of models that achieved cross validated prediction significance FDR lower than 5%.

Tissue	# protein cod.	# samples	# signif. models (FDR <.05)
Adipose Subcutaneous	15935	298	7249
Adipose Visceral Omentum	15790	185	4568
Adrenal Gland	15370	126	4174
Artery Aorta	15401	197	6182
Artery Coronary	15437	118	3222
Artery Tibial	15388	285	7121
Brain Anterior cingulate cortex BA24	15385	72	2559
Brain Caudate basal ganglia	15658	100	3544
Brain Cerebellar Hemisphere	15202	89	4068
Brain Cerebellum	15456	103	4995
Brain Cortex	15652	96	3558
Brain Frontal Cortex BA9	15547	92	3258
Brain Hippocampus	15628	81	2566
Brain Hypothalamus	15818	81	2451
Brain Nucleus accumbens basal ganglia	15636	93	3057
Brain Putamen basal ganglia	15374	82	2749
Breast Mammary Tissue	16188	183	4648
Cells EBV-transformed lymphocytes	13905	114	3660
Cells Transformed fibroblasts	14556	272	7609
Colon Sigmoid	15599	124	3720
Colon Transverse	16010	169	4788
Esophagus Gastroesophageal Junction	15364	127	3601
Esophagus Mucosa	15741	241	6889
Esophagus Muscularis	15556	218	6533
Heart Atrial Appendage	15242	159	4565
Heart Left Ventricle	14834	190	4858
Liver	14767	97	2759
Lung	16336	278	6564
Muscle Skeletal	14959	361	6563
Nerve Tibial	15998	256	8113
Ovary	15238	85	2880
Pancreas	15335	149	4931
Pituitary	16131	87	3335
Prostate	15994	87	2614
Skin Not Sun Exposed Suprapubic	16110	196	5633
Skin Sun Exposed Lower leg	16259	302	7567
Small Intestine Terminal Ileum	15872	77	2613
Spleen	15371	89	3715
Stomach	15989	170	4096
Testis	17683	157	7043
Thyroid	16193	278	8026
Uterus	15164	70	2159
Vagina	15715	79	2041
Whole Blood	14858	338	6650