Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics

Alvaro N. Barbeira[1], Scott P. Dickinson[1], Jason M. Torres[2], Jiamao Zheng[1], Eric S. Torstenson[3], Heather E. Wheeler[4], Kaanan P. Shah[1], Rodrigo Bonazzola[1], Tzintzuni Garcia[5], 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

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

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

Abstract

Scalable, integrative methods to understand mechanisms that link genetic variants with phenotypes are needed. Here we derive a mathematical expression to compute PrediXcan (a gene mapping approach) results using summary data (S-PrediXcan) and show its accuracy and general robustness to misspecified reference sets. We apply this framework to 44 GTEx tissues and 100+ phenotypes from GWAS and meta-analysis studies, creating a growing public catalog of associations that seeks to capture the effects of gene expression variation on human phenotypes. Replication in an independent cohort is shown. Most of the associations were tissue specific, suggesting context specificity of the trait etiology. Colocalized significant associations in unexpected tissues underscore the need for an agnostic scanning of multiple contexts to improve our ability to detect causal regulatory mechanisms. Monogenic disease genes are enriched among significant associations for related traits, suggesting that smaller alterations of these genes may cause a spectrum of milder phenotypes.

Introduction

Over the last decade, GWAS have been successful in robustly associating genetic loci to human complex traits. However, the mechanistic understanding of these discoveries is still limited, hampering the translation of the associations 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. Functional class quantification 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 examined tissues, is the Genotype-Tissue Expression Project (GTEx): a large-scale effort where DNA and RNA were collected from multiple tissue samples from nearly 1000 individuals and sequenced to high coverage [9, 10]. 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 [11], a gene-level association approach that tests the mediating effects of gene expression levels on phenotypes. PrediXcan is implemented on GWAS or sequencing studies (i.e. studies with genome-wide interrogation of DNA variation and phenotypes). It imputes transcriptome levels with models trained in measured transcriptome datasets (e.g. GTEx). These predicted expression levels are then correlated with the phenotype in a gene association test that addresses some of the key limitations of GWAS [11].

Meta-analysis efforts that aggregate results from multiple GWAS 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 use summary level data rather than individual level data are needed.

Methods similar to PrediXcan that estimate the association between intermediate gene expression levels and phenotypes, but use summary statistics have been reported: TWAS (summary version) [15] and SMR (Summary Mendelian Randomization) [16]. Another class of methods that integrate eQTL

information with GWAS results are based on colocalization of eQTL and GWAS signals. Colocalized signals provide evidence of possible causal relationship between the target gene of an eQTL and the complex trait. These include RTC [1], Sherlock [17], COLOC [18], and more recently eCAVIAR [19] and ENLOC [20].

Here we derive a mathematical expression that allows us to compute the results of PrediXcan without the need to use individual-level data, greatly expanding its applicability. We compare with existing methods and outline a best practices framework to perform integrative gene mapping studies, which we term MetaXcan.

We apply the MetaXcan framework by first training over 1 million elastic net prediction models of gene expression traits, covering protein coding genes across 44 human tissues from GTEx, and then performing gene-level association tests over 100 phenotypes from 40 large meta-analysis consortia and dbGaP.

Results

Computing PrediXcan results using summary statistics

We have derived an analytic expression 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-A, we illustrate the mechanics of Summary-PrediXcan (S-PrediXcan) in relation to traditional GWAS and the individual-level PrediXcan method [11].

We find high concordance between PrediXcan and S-PrediXcan results indicating that in most cases, we can use the summary version without loss of power to detect associations. Figure 2 shows the comparison of PrediXcan and S-PrediXcan Z-scores for a simulated phenotype (under the null hypothesis), a cellular growth phenotype and two disease phenotypes: type 1 diabetes and bipolar disorder from the WTCCC Consortium [21]. For the simulated phenotype, the study sets (in which GWAS is performed) and the reference set (in which LD between SNPs is computed) were African, East Asian, and European from 1000 Genomes. The training set (in which prediction models are trained) was European (DGN Cohort [5]) in all cases. The high correlation between PrediXcan and S-PrediXcan demonstrates the robustness of our method to mismatches between reference and study sets. Despite the generally good concordance between the summary and individual level methods, there were a handful of false positive

results with S-PrediXcan much more significant than PrediXcan. This underscores the need to use closely matched LD information whenever possible. Supplementary Figure 11 shows S-PrediXcan's performance on a phenotype simulated under the alternative hypothesis.

Notice that we are not testing here whether PrediXcan itself is robust to population differences between training and study sets. Robustness of the prediction across populations has been previously reported [22]. We further corroborated this in Supplementary Figure 10.

Next we compare with other summary result-based methods such as S-TWAS, SMR, and COLOC.

Colocalization estimates complement PrediXcan results

One class of methods seeks to determine whether eQTL and GWAS signals are colocalized or are distinct although linked by LD. This class includes COLOC [18], Sherlock [17], and RTC [1], and more recently eCAVIAR [19], and ENLOC [20]. Thorough comparison between these methods can be found in [18, 19]. HEIDI, the post filtering step in SMR that estimates heterogeneity of GWAS and eQTL signals, can be included in this class. We focus here on COLOC, whose quantification of the probability of five configurations complements well with S-PrediXcan results.

COLOC provides the probability of 5 hypotheses: H0 corresponds to no eQTL and no GWAS association, H1 and H2 correspond to association with eQTL but no GWAS or vice-versa, H3 corresponds to eQTL and GWAS association but independent signals, and finally H4 corresponds to shared eQTL and GWAS association. P0, P1, P2, P3, and P4 are the corresponding probabilities for each configuration. The sum of the five probabilities is 1. The authors [18] recommend to interpret H0, H1, and H2 as limited power; we will aggregate these three hypothesis into one event with probability 1-P3-P4 for convenience.

Figure 3 shows ternary plots [23] with P3, P4, and 1-P3-P4 as vertices. The blue region, top sub-triangle, corresponds to high probability of colocalized eQTL and GWAS signals (P4). The orange region, bottom left, corresponds to high probability of distinct eQTL and GWAS signals (P3). The gray region, center and bottom right, corresponds to low probability of both colocalization and independent signals.

Figure 3-B shows association results for all gene-tissue pairs with the height phenotype. We find that most associations fall in the gray, 'undetermined', region. When we restrict the plot to S-PrediXcan Bonferroni-significant genes (Figure 3-C), three distinct peaks emerge in the high P4 region ('colocalized signals'), high P3 region ('independent signals' or 'non-colocalized signals'), and 'undetermined' region. Moreover, when genes with low prediction performance are excluded (Supplementary Figure 6-D) the

'undetermined' peak significantly diminishes.

These clusters provide a natural way to classify significant genes and complement S-PrediXcan results. Depending on false positive/false negative trade-off choices, genes in the 'independent signals' or both 'independent signals' and 'undetermined' can be filtered out. The proportion of colocalized associations (P4>0.5) ranged from 5 to 100% depending on phenotype with a median of 27.6%. The proportion of 'non-colocalized' associations ranged from 0 to 77% with a median of 27.0%. Supplementary Table 2 summarizes the percentages of significant associations that fall into the different colocalization regions.

This post-filtering idea was first implemented in the SMR approach using HEIDI. Comparison of COLOC results with HEIDI is shown in Figure 6-E to -F.

Comparison of S-PrediXcan to S-TWAS

Gusev et al. have proposed Transcriptome-Wide Association Study based on summary statistics (S-TWAS), which imputes the SNP level Z-scores into gene level Z-scores. This is not the same as computing the results of individual level TWAS. We show (in Methods section) that the difference between the individual level and summary level TWAS is given by the factor $\sqrt{\frac{1-R_l^2}{1-R_g^2}}$, where R_l is the proportion of variance in the phenotype explained by a SNP's allelic dosage, and R_g is the proportion explained by gene expression (see Methods section). For most practical purposes we have found that this factor is very close to 1 so that if the same prediction models were used, no substantial difference between S-TWAS and S-PrediXcan should be expected.

Figure 4-A shows a diagram of S-PrediXcan and S-TWAS. Both use SNP to phenotype associations results ($Z_{X,Y}$) and prediction weights (w_{X,T_g}) to infer the association between the gene expression level (T_g) and phenotype (Y).

Figure 4-B compares S-TWAS significance (as reported in [24]) to S-PrediXcan significance. The difference between the two approaches is mostly driven by the different prediction models: TWAS uses BSLMM [25] whereas PrediXcan uses elastic net [26]. BSLMM allows two components: one sparse (small set of large effect predictors) and one polygenic (all variants contribute some marginal effect to the prediction). For PrediXcan we have chosen to use a sparse model (elastic net) based on the finding that the genetic component of gene expression levels is mostly sparse [27].

Figure 4-C shows that the proportion of non-colocalized (independent) GWAS and eQTL signals is larger among TWAS significant genes than among S-PrediXcan significant ones. This is likely due to

the polygenic component of BSLMM models, a wider set of SNPs increasing the chance of capturing LD-contaminated (non-colocalized) association. Figure 4-D shows that, for most traits, the proportion of colocalized signals is larger among S-PrediXcan significant genes than S-TWAS significant genes.

Comparison of S-PrediXcan to SMR

Zhu et al. have proposed Summary Mendelian Randomization (SMR) [16], 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 [16]) for the mediating effect of the target gene expression on the phenotype.

Unfortunately, the derived statistic is mis-calibrated. A QQ plot comparing the SMR statistic (under the null hypothesis of no eQTL or GWAS association) is off by a factor of 1/4. The sample mean of the statistic is 0.25 instead of 1, the mean of a χ^2_1 random variable. See Figure 5 (B and C). The χ^2 approximation is only valid in two extreme cases: when the eQTL association is much stronger than the GWAS association or vice versa, when the GWAS association is much stronger than the eQTL association. See Methods for details.

One limitation is that the significance of the SMR statistic is the lower of the top eQTL association (genotype to expression) or the GWAS association (genotype to phenotype) as shown in Figure 5 (E and F). Given the much larger sample sizes of GWAS studies, for most genes, the combined significance will be determined by the eQTL association. The combined statistic forces us to apply multiple testing correction for all genes, even those that are distant to GWAS associated loci, which is unnecessarily conservative. Keep in mind that currently both SMR and PrediXcan only use cis associations. An example may clarify this further. Let us suppose that for a given phenotype there is only one causal SNP and that the GWAS yielded a highly significant p-value, say 10^{-50}. Let us also suppose that there is only one gene (gene A) in the vicinity (we are only using cis predictors) associated with the causal SNP with p=10^{-5}. SMR would compute the p-values of all genes and yield a p-value $\approx 10^{-5}$ for gene A (the less significant p-value). However, after multiple correction this gene would not be significantly associated with the phenotype. Here it is clear that we should not be adjusting for testing of all genes when we know a priori that only one is likely to produce a gene level association. In contrast, the PrediXcan p-value would be $\approx 10^{-50}$ for gene A and would be distributed uniformly from 0 to 1 for the remaining genes. Most likely only gene A (or perhaps a handful of genes, just by chance) would be significant after Bonferroni correction.

If we further correct for prediction uncertainty (here =eQTL association), a p-value of $\approx 10^{-5}$ would remain significant since we only need to correct for the (at most) handful of genes that were Bonferroni significant for the PrediXcan p-value.

Another potential disadvantage of this method is that only top-eQTLs are used for testing the gene level association. This does not allow to aggregate the effect on the gene across multiple variants.

Figure 5-D compares S-PrediXcan (elastic net) and SMR association results. As expected, SMR p-values tend to be less significant than S-PrediXcan's in large part due to the additional adjustment for the uncertainty in the eQTL association. Figures 5-E and -F show that the SMR significance is bounded by the eQTL and GWAS association strengths of the top eQTL.

SMR introduces a post filtering step via an approach called HEIDI, which is compared to COLOC in Figure 3 and Supplementary Figure 6.

MetaXcan framework

Building on S-PrediXcan and existing approaches, we define a general framework (MetaXcan) to integrate eQTL information with GWAS results and map disease-associated genes. This evolving framework can incorporate models and methods to increase the power to detect causal genes and filter out false positives. Existing methods fit within this general framework as instances or components (Figure 6).

The framework starts with the training of prediction models for gene expression traits followed by a selection of high-performing models. Next, a mathematical operation is performed to compute the association between each gene and the downstream complex trait. Additional adjustment for the uncertainty in the prediction model can be added. To avoid capturing LD-contaminated associations, which can occur when expression predictor SNPs and phenotype causal SNPs are different but in LD, we use colocalization methods that estimate the probability of shared or independent signals.

PrediXcan implementations use elastic net models motivated by our observation that gene expression variation is mostly driven by sparse components [27]. TWAS implementations have used Bayesian Sparse Linear Mixed Models [25] (BSLMM). SMR fits into this scheme with prediction models consisting solely of the top eQTL for each gene (weights are not necessary here since only one SNP is used at a time).

For the last step, we chose COLOC to estimate the probability of colocalization of GWAS and eQTL signals. COLOC probabilities cluster more distinctly into different classes and thus, unlike other methods, suggests a natural cut off threshold at P=0.5. Another advantage of COLOC is that for genes with low

probability of colocalization, it further distinguishes distinct GWAS and eQTL signals from low power. This is a useful feature that future development of colocalization methods should also offer. SMR, on the other hand, uses its own estimate of 'heterogeneity' of signals calculated by HEIDI.

Suggested association analysis pipeline

1. Perform PrediXcan or S-PrediXcan using all tissues. Use Bonferroni correction for all gene-tissue pairs: keep p$<$ 0.05/ number of gene-tissue pairs tested.

2. Keep associations with significant prediction performance adjusting for number of PrediXcan significant gene-tissue pairs: keep prediction performance p-values $<$ 0.05/(number of significant associations from previous step).

3. Filter out LD-contaminated associations, i.e. gene-tissue pairs in the 'independent signal' (='non-colocalized') region of the ternary plot (See Figure 3-A): keep COLOC P3$<$ 0.5 (Blue and gray regions in Figure 3-A).

4. If further reduction of number of genes to be taken to replication or validation is desired, keep only hits with explicit evidence of colocalization: P4$>$ 0.5 (Blue region in Figure 3-A).

Any choice of thresholds has some level of arbitrariness. Depending on the false positive and negative trade off, these number may be changed.

Gene expression variation in humans is associated to diverse phenotypes

We downloaded summary statistics from meta analyses of over 100 phenotypes from 40 consortia. The full list of consortia and phenotypes is shown in Supplementary Table 4. 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 Methods section, and a whole blood model from the DGN cohort presented in [11].

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 pairs. Height had the largest number of significantly associated unique genes at 1,686 (based on a

GWAMA of 250K individuals). Other polygenic diseases with a large number of associations include schizophrenia with 305 unique significant genes ($n = 150$K individuals), low-density lipoprotein cholesterol (LDL-C) levels with 296 unique significant genes ($n = 188$K), other lipid levels, glycemic traits, and immune/inflammatory disorders such as rheumatoid arthritis and inflammatory bowel disease. For other psychiatric phenotypes, a much smaller number of significant associations was found, with 8 significant genes for bipolar disorder ($n = 16{,}731$) and one for major depressive disorder ($n = 18{,}759$), probably due to smaller sample sizes, but also smaller effect sizes.

When step 2 from the suggested pipeline is applied, keeping only reliably predicted genes, we are left with 739 genes for height, 150 for schizophrenia, 117 for LDL-C levels.

After step 3, which keeps genes that are without strong evidence of LD-contamination, these numbers dropped to 264 for height, 58 for schizophrenia, and 60 for LDL-C levels. After step 4, which keeps only genes with strong evidence of colocalization, we find 215 genes for height, 49 for schizophrenia, and 35 for LDL-C. The counts for the full set of phenotypes can be found in Supplementary Table 4.

Mostly, genome-wide significant genes tend to cluster around known SNP-level genome-wide significant loci or sub-genome-wide significant loci. Regions with sub-genome-wide significant SNPs can yield genome-wide significant results in S-PrediXcan, because of the reduction in multiple testing and the increase in power arising from the combined effects of multiple variants. Supplementary Table 3 lists a few examples where this occurs.

The proportion of colocalized associations (P4>0.5) ranged from 5 to 100% depending on phenotype with a median of 27.6%. The proportion of 'non colocalized' associations ranged from 0 to 77% with a median of 27.0%.

See full set of results in our online catalog (gene2pheno.org). Significant gene-tissue pairs are included in Supplementary Table 5. To facilitate comparison, the catalog contains all SMR results we generated and the S-TWAS results reported by [24] for 30 GWAS traits and GTEx BSLMM models. Note that SMR application to 28 phenotypes was reported by [28] using whole blood eQTL results from [29].

Moderate changes in ClinVar gene expression is associated with milder phenotypes

We reasoned that if complete knock out of monogenic disease genes cause severe forms of the disease, more moderate alterations of gene expression levels (as effected by regulatory variation in the population) could cause more moderate forms of the disease. Thus moderate alterations in expression levels of monogenic

disease genes (such as those driven by eQTLs) may have an effect on related complex traits, and this effect could be captured by S-PrediXcan association statistics. To test this hypothesis, we obtained genes listed in the ClinVar database [30] for obesity, rheumatoid arthritis, diabetes, Alzheimer's, Crohn's disease, ulcerative colitis, age-related macular degeneration, schizophrenia, and autism. Figure 8 displays the QQ plot for all associations and those in ClinVar database. As postulated, we found enrichment of significant S-PrediXcan associations for ClinVar genes for all tested phenotypes except for autism and schizophrenia. The lack of significance for autism is probably due to insufficient power: the distribution of p-values is close to the null distribution. In contrast, for schizophrenia, many genes were found to be significant in the S-PrediXcan analysis. There are several reasons that may explain this lack of enrichment: genes identified with GWAS and subsequently with S-PrediXcan have rather small effect sizes, so that it would not be surprising that they were missed until very large sample sizes were aggregated; ClinVar genes may originate from rare mutations that are not well covered by our prediction models, which are based on common variation (due to limited sample sizes of eQTL studies and the minor allele frequency -MAF-filter used in GWAS studies); or the mechanism of action of the schizophrenia linked ClinVar genes may be different than the alteration of expression levels. Also, the pathogenicity of some of the ClinVar entries has been questioned [31]. 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 improves discovery

Most genes were found to be significantly associated in a handful of tissues as illustrated in Figure 9-B. For example, for LDL-C levels, liver was the most enriched tissue in significant associations as expected given known biology of this trait. (See Supplementary Figure 5). 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).

However, in general, tissues expected to stand out as more enriched for diseases given currently known biology did not consistently do so when we looked at the average across all (significant) genes, using various measures of enrichment. For example, the enrichment in liver was less apparent for high-density lipoprotein cholesterol (HDL-C) or triglyceride levels. We find for many significant associations that the evidence is present across multiple tissues. This may be caused by a combination of context specificity and sharing of regulatory mechanism across tissues.

Next, we illustrate the challenges of identifying disease relevant tissues based on eQTL information using three genes with well established biology: *C4A* for schizophrenia [32] and *SORT1* [33] and *PCSK9* both for LDL-C and cardiovascular disease. S-PrediXcan 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 9-A. Representative results are shown in Supplementary Tables 6, 7, and 8. Supplementary Table 9 contains the full set MetaXcan results (i.e. association, colocalization, HEIDI) for these genes.

SORT1 is a gene with strong evidence for a causal role in LDL-C levels, and as a consequence, is likely to affect risk for cardiovascular disease [33]. 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 S-PrediXcan association in liver (p-value ≈ 0, $Z = -28.8$), consistent with our prior knowledge of lipid metabolism. In this example, tissue specific results suggest a causal role of *SORT1* in liver.

However, in the following example, association results across multiple tissues do not allow us to discriminate the tissue of action. *C4A* is a gene with strong evidence of causal effect on schizophrenia risk via excessive synaptic pruning in the brain during development [32]. Our results show that *C4A* is associated with schizophrenia risk in all tissues ($p < 2.5 \times 10^{-7}$ in 36 tissue models and $p<0.05$ for the remaining 4 tissue models).

PCSK9 is a target of several LDL-C lowering drugs currently under trial to reduce cardiovascular events [34]. The STARNET study [35] profiled gene expression levels in cardiometabolic disease patients and showed tag SNP rs12740374 to be a strong eQTL for *PCSK9* in visceral fat but not in liver. Consistent with this, our S-PrediXcan results also show a highly significant association between *PCSK9* and LDL-C ($p \approx 10^{-13}$) 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 ($p \approx 10^{-27}$). Accordingly, in our training set (GTEx), there is much stronger evidence of regulation of this gene in tibial nerve compared to visceral fat.

Most associations highlighted here have high colocalization probabilities. See Supplementary tables 6, 7, and 8. However, visceral fat association shows evidence of non colocalization (probability of independent signals P3=0.69 in LDL-C). It is possible that the relevant regulatory activity in visceral adipose tissue was not detected in the GTEx samples for various reasons but it was detected in tibial nerve. Thus by looking into all tissues' results we increase the window of opportunities where we can detect the

association.

PCSK9 yields colocalized signals for LDL-C levels in Tibial Nerve, Lung and Whole blood. *SORT1* shows colocalization with LDL-C in liver (P4≈1) and pancreas (P4=0.90). *C4A* is colocalized with schizophrenia risk for the majority of the tissues (29/40) with a median colocalization probability of 0.82.

These examples demonstrate the power of studying 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. Based on these results, we recommend to scan all tissues' models to increase the chances to detect the relevant regulatory mechanism that mediates the phenotypic association. False positives can be controlled by Bonferroni correcting for the additional tests.

Replication in an independent cohort, GERA

We used data from the Resource for Genetic Epidemiology Research on Adult Health and Aging study (GERA, phs000674.v1.p1) [36, 37]. This is a study led by the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics with over 100,000 participants. We downloaded the data from dbGaP and performed GWAS followed by S-PrediXcan analysis of 22 conditions available in the European subset of the cohort.

For replication, we chose Coronary Artery Disease (CAD), LDL cholesterol levels, Triglyceride levels, and schizophrenia, which had closely related phenotypes in the GERA study and had a sufficiently large number of Bonferroni significant associations in the discovery set. Analysis and replication of the type 2 diabetes phenotype can be found in [38]. Coronary artery disease hits were compared with 'Any cardiac event', LDL cholesterol and triglyceride level signals were compared with 'Dyslipidemia', and schizophrenia was compared to 'Any psychiatric event' in GERA.

High concordance between discovery and replication is shown in Figure 10 where dyslipidemia association Z-scores are compared to LDL cholesterol Z-scores. The majority of gene-tissue pairs (92%, among the ones with Z-score magnitude greater than 2 in both sets) have concordant direction of effects in the discovery and replication sets. The high level of concordance is supportive of an omnigenic trait architecture [39]

Following standard practice in meta-analysis, we consider a gene to be replicated when the following three conditions are met: the p-value in the replication set is <0.05, the direction of discovery and replication effects are the same, and the meta analyzed p-value is Bonferroni significant with the discovery

Discovery phenotype	Replication phenotype	# signif genes in disc set	# replicated genes	π_1(all) in repl	π_1(sig) in repl	% replicated genes	# replicated coloc or undeterm
Coronary artery disease	Any cardiac event	56	6	0.4%	49.1%	10.7%	6
LDL cholesterol	Dyslipidemia	282	219	5.8%	90.8%	78.5%	184
Triglycerides	Dyslipidemia	233	100	5.8%	73.1%	43.5%	69
Schizophrenia	Any psychiatric event	285	60	1.2%	47.6%	21.1%	51

Table 1. Replication of results in GERA. Significant genes/tissue pairs were replicated using a closely matched phenotype in an independent dataset from the GERA cohort [36]. The criteria consisted in significance threshold for replication at $p < 0.05$, concordant directions of effect, and meta analysis p-value less than the Bonferroni threshold in the discovery set. π_1 is an estimate of proportion of true positives in the replication set. π_1(all) uses all gene-tissue pairs whereas π_1(sig) is computed using only gene-tissue pairs that were significant in the discovery set. The column '# replicated genes coloc or undeterm' is the number of replicated genes excluding the ones for which there was strong evidence of independent GWAS and eQTL signals.

threshold.

Among the 56 genes significantly associated with CAD in the discovery set, 6 (11%) were significantly associated with 'Any cardiac event' in GERA. Using 'Dyslipidemia' as the closest matching phenotype, 78.5% and 43.5% of LDL and triglyceride genes replicated, respectively. Among the 285 genes associated with schizophrenia in the discovery set, 51 (21%) replicated. The low replication rate for CAD and Schizophrenia is likely due to the broad phenotype definitions in the replication.

We found no consistent replication pattern difference between colocalized and non-colocalized genes. This is not unexpected if the LD pattern is similar between discovery and replication sets.

The full list of significant genes can be queried in `gene2pheno.org`.

Discussion

Here we derive a mathematical expression to compute PrediXcan results without using individual level data, which greatly expands its applicability and is robust to study and reference set mismatches. This has not been done before. TWAS, which for the individual level approach only differs from PrediXcan on the prediction model used in the implementation, has been extended to use summary level data. When Gaussian imputation is used, the relationship between individual level and summary versions of TWAS is clear. This is not the case when extended to general weights (such as BSLMM). Our mathematical

derivation shows the analytic difference between them explicitly.

The larger proportion of non-colocalized signals from TWAS suggests that by using BSLMM (with polygenic component), it could be more susceptible to LD-contamination than PrediXcan, which uses elastic net (a sparser model). Improved colocalization methods without single causal variant assumption may be needed to strengthen this argument. But the predominantly sparse genetic architecture of gene expression traits [27] supports the benefit of elastic net over BSLMM predictors. We show that SMR statistics needs to be calibrated and argue that by combining the eQTL and GWAS uncertainties into one statistic, it forces the user to apply multiple correction that may be unnecessarily conservative.

We also add a post filtering step, to mitigate issues with LD-contamination. Based on consistency with PrediXcan and interpretability of results, we have chosen to use COLOC for filtering. However, colocalization estimation is an active area of research and improved versions or methods will be adopted in the future.

Despite the generally good concordance between the summary and individual level methods, there were a handful of false positive results with S-PrediXcan much more significant than PrediXcan. This underscores the need to use closely matched LD information whenever possible.

We applied our framework to over 100 phenotypes using transcriptome prediction models trained in 44 tissue from the GTEx Consortium and generated a catalog of downstream phenotypic association results of gene expression variation, a growing resource for the community.

The enrichment of monogenic disease genes among related phenotype associations suggests that moderate alteration of expression levels as effected by common genetic variation may cause a continuum of phenotypic changes. Alternatively, a more complex interplay between common and rare variation could be taking place such as higher tolerance to loss of function mutations in lower expressing haplotypes which could induce association with predicted expression.

We are finding that most trait associations are tissue specific; i.e. they are detected in a handful of tissues. However, we also find that expected tissues given known biology do not necessarily rank among the top enriched tissues. This suggests context specificity of the pathogenic mechanism; specific developmental stage or environmental conditions may be necessary to detect the regulatory event. On the other hand, we are detecting associations in unexpected tissues which suggests a sharing of regulation across multiple tissues/contexts or perhaps novel biology that takes place in these tissues. In either case, agnostic scanning of a broad set of tissues is necessary to discover these mechanisms.

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. 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

Summary-PrediXcan formula

Figure 1-B shows the main analytic expression used by Summary-PrediXcan 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, 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 phenotype variance and sample size. Dropping this factor from the formula does not affect the accuracy of the results as demonstrated in the close to perfect concordance between PrediXcan and Summary-PrediXcan results on the diagonal of Figure 2-A.

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,

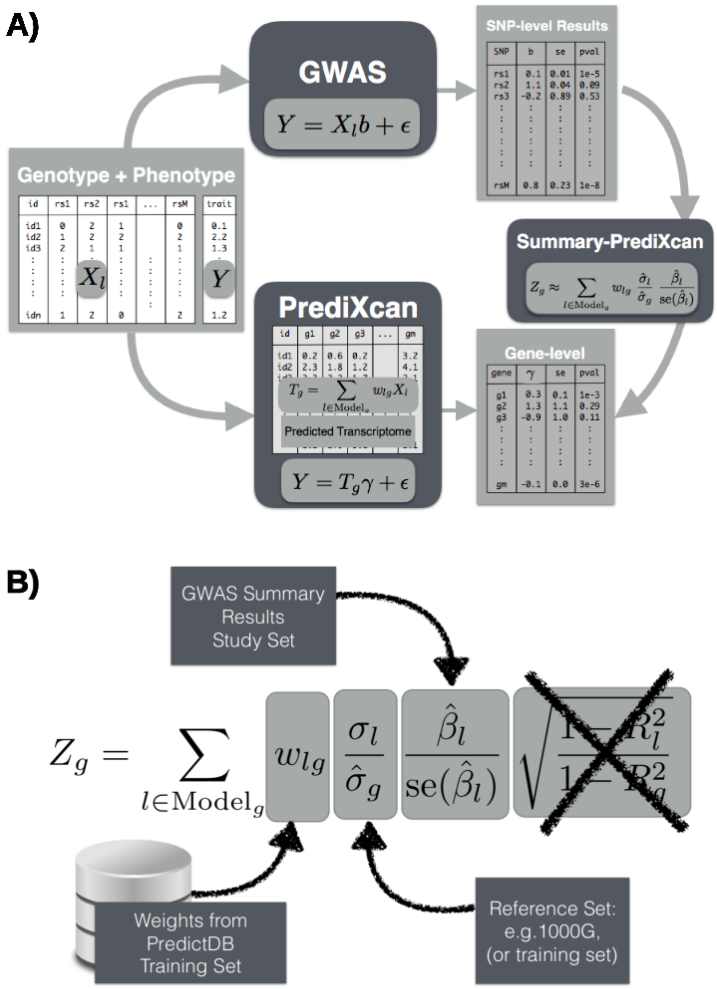


Figure 1. Panel A: Comparison between GWAS, PrediXcan, and Summary-PrediXcan. Both GWAS and PrediXcan take genotype and phenotype data as input. GWAS computes the regression coefficients of Y on X_l using the model $Y = a + X_l b + \epsilon$, where Y is the phenotype and X_l the individual SNP dosage. The output is a 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. Summary-PrediXcan directly computes the gene-level association results using the output from GWAS. **Panel B: Components of the S-PrediXcan formula.** This figure shows the components of the formula to calculate PrediXcan gene-level association results using summary statistics. The different sets involved as 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 (1000G, or training set having some advantages) 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. We found this approximation to have negligible impact on the results.

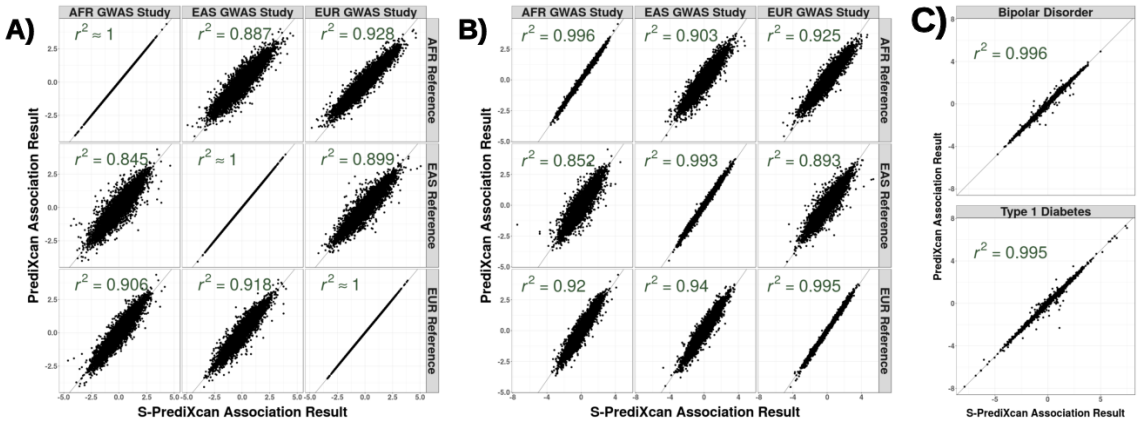


Figure 2. Comparison of PrediXcan vs. S-PrediXcan for **A)** a simulated phenotype under null hypothesis of no genetic component; **B)** a cellular phenotype (=intrinsic growth); and **C)** bipolar disorder and type 1 diabetes studies from Wellcome Trust Case Control Consortium (WTCCC). Gene expression prediction models were based on the DGN cohort presented in [11]. For the simulated phenotype, study sets (GWAS set) and reference sets (LD calculation set) consisted of African (661), East Asian (504) and European (503) individuals from the 1000 Genomes Project. When the same study set is used as reference set, we obtained a high correlation: $r^2 > 0.99999$. For the intrinsic growth phenotype, study sets were a subset of 140 individuals from each of the African, Asian an European groups from 1000 Genomes Project. The reference set was the same as for the simulated phenotype. For the disease phenotypes, the study set consisted of British individuals, and the LD calculation set was the European population subset of the 1000 Genomes Project.

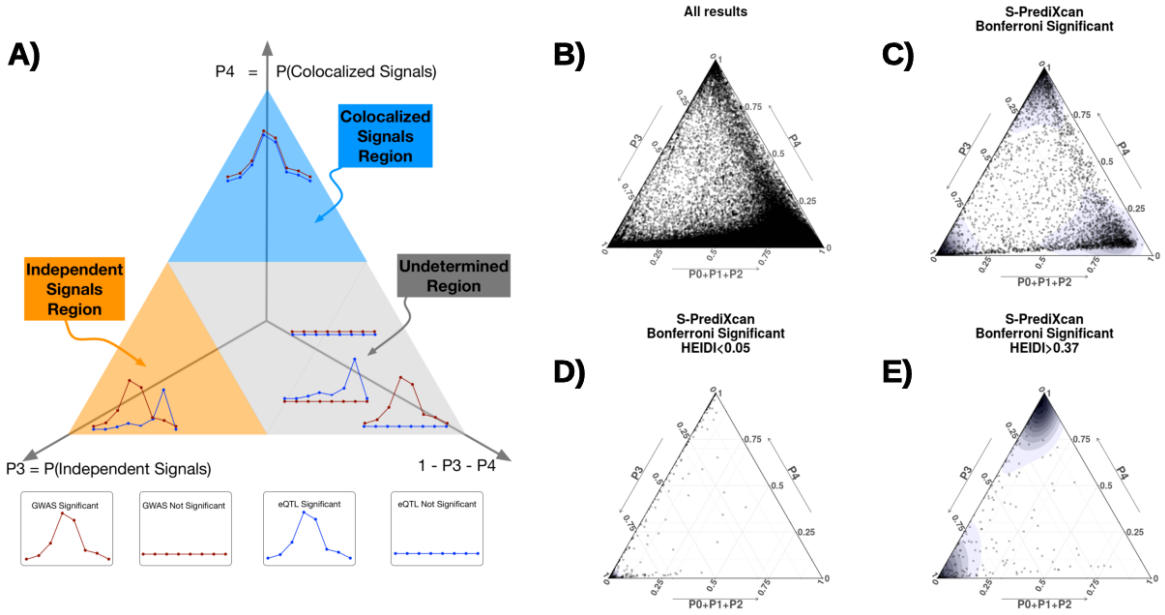


Figure 3. Colocalization status of S-PrediXcan results.
Panel A shows a ternary plot that represents the probabilities of various configurations from COLOC. This plot conveniently constrains the values such that the sum of the probabilities is 1. All points in a horizontal line have the same probability of 'colocalized' GWAS and eQTL signals (P4), points on a line parallel to the right side of the triangle (NW to SE) have the same probability of 'Independent signals' (P3), and lines parallel to the left side of the triangle (NE to SW) correspond to constant P1+P2+P3. Top sub-triangle in blue corresponds to high probability of colocalization (P4>0.5), lower left sub-triangle in orange corresponds to probability of independent signals (P3>0.5), and lower right parallelogram corresponds to genes without enough power to determine or reject colocalization. The following panels present ternary plots of COLOC probabilities with a density overlay for S-PrediXcan results of the Height phenotype.
Panel B shows the colocalization probabilities for all gene-tissue pairs. Most results fall into the 'undetermined' region.
Panel C shows that if we keep only Bonferroni-significant S-PrediXcan results, associations tend to cluster into three distinct regions: 'independent signals', 'colocalized' and 'undertermined'.
Panel D shows that HEIDI significant genes (to be interpreted as high heterogeneity between GWAS and eQTL signals, i.e. distinct signals) tightly cluster in the 'independent signal' region, in concordance with COLOC. A few genes fall in the 'colocalized' region, in disagreement with COLOC classification. Unlike COLOC results, HEIDI does not partition the genes into distinct clusters and an arbitrary cutoff p-value has to be chosen.
Panel E shows genes with large HEIDI p-value (no evidence of heterogeneity) which fall in large part in the 'colocalized' region. However a substantial number fall in 'independent signal' region, disagreeing with COLOC's classification.

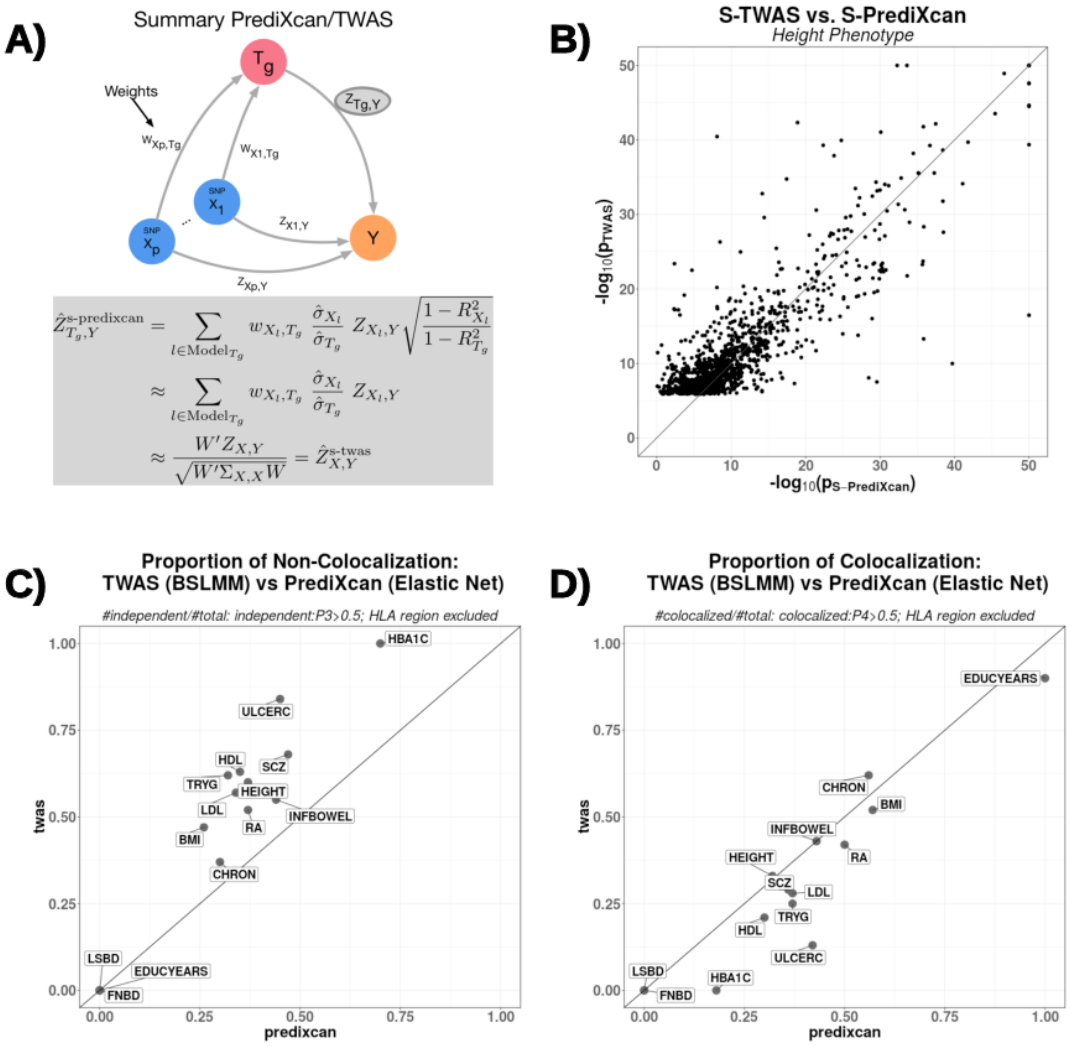


Figure 4. Comparison between S-PrediXcan and S-TWAS.
Panel A depicts how Summary- TWAS and PrediXcan test the mediating role of gene expression level T_g. Multiple SNPs are linked to the expression level of a gene via weights w_{X,T_g}.
Panel B shows the significance of Summary-TWAS (BSLMM) vs. Summary-PrediXcan (elastic net), for the height phenotype across 44 GTEx tissues. There is a small bias caused by using S-TWAS results available from [24], which only lists significant hits. S-PrediXcan tends to yield a larger number of significant associations (see Supplementary Figure 13). P-values were thresholded at 10^{-50} for visualization purposes.
Panel C shows the proportion of non-colocalized associations (distinct eQTL and GWAS signals) from S-TWAS significant vs S-PrediXcan significant results. For all phenotypes, S-TWAS has a higher proportion of LD-contaminated signals compared to S-PrediXcan, as estimated via COLOC.
Panel D shows the proportion of colocalized associations (shared eQTL and GWAS signals) from S-TWAS significant vs S-PrediXcan significant results. For most phenotypes, TWAS has lower proportion of colocalized signals compared to S-PrediXcan, as estimated via COLOC.
Phenotype Abbreviation: Femoral Neck Bone Density (FNBD), Lumbar Spine Bone Density (LSBD), Body Mass Index (BMI), Height (HEIGHT), Low-Density Lipoprotein Cholesterol (LDL), High-Density Lipoprotein Cholesterol (HDL), Tryglicerides (TRYG), Chron's Disease (CHRON), Inflammatory Bowel's Disease (INFBOWEL), Ulcerative Colitis (ULCERC), Hemogoblin Levels (HBA1C) HOMA Insulin Response (HOMA-IR) Schizophrenia (SCZ), Rheumatoid Arthritis (RA), College Completion (COLLEGE), Education Years (EDUCYEARS)

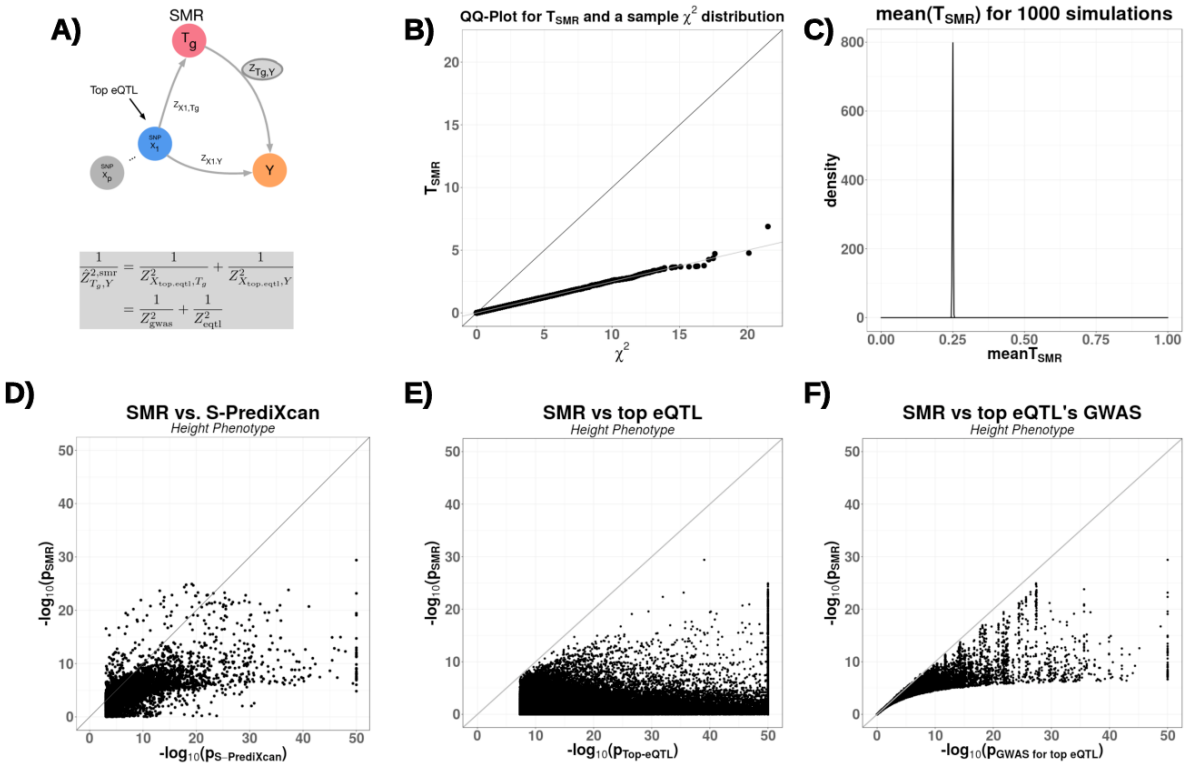


Figure 5. Comparison between Summary-PrediXcan and SMR.
Panel A depicts how SMR tests the mediating role of gene expression level T_g. The top eQTL is linked to the phenotype as an instrumental variable in a Mendelian Randomization approach.
Panel B shows a QQ plot for simulated values of T_{SMR}. Under the null hypothesis of no genetic effect, 10^5 values for Z^2_{GWAS} and Z^2_{eQTL} were sampled from a χ^2_1 distribution, and then T_{SMR} statistic was calculated. This statistic is compared to a χ^2_1 distribution to illustrate the mis-calibration of T_{SMR}. A factor of $1/4$ seems necessary to match the desired distribution.
Panel C shows the sample mean of T_{SMR} from 1000 simulations, centered on 0.25 instead the expected value of 1 for a χ^2 distribution.
Panel D shows the significance of SMR vs. the significance of Summary-PrediXcan. As expected, SMR associations tend to be smaller than S-PrediXcan ones.
Panels E and **F** show that the SMR statistics significance is bounded by GWAS and eQTL p-values. The p-values (-log10) of the SMR statistics are plotted against the GWAS p-value of the top eQTL SNP (**panel E**), and the gene's top eQTL p-value (**panel F**).
Some of the associations, GWAS and eQTL p-values were more significant than shown since they were thresholded at 10^{-50} to improve visualization.

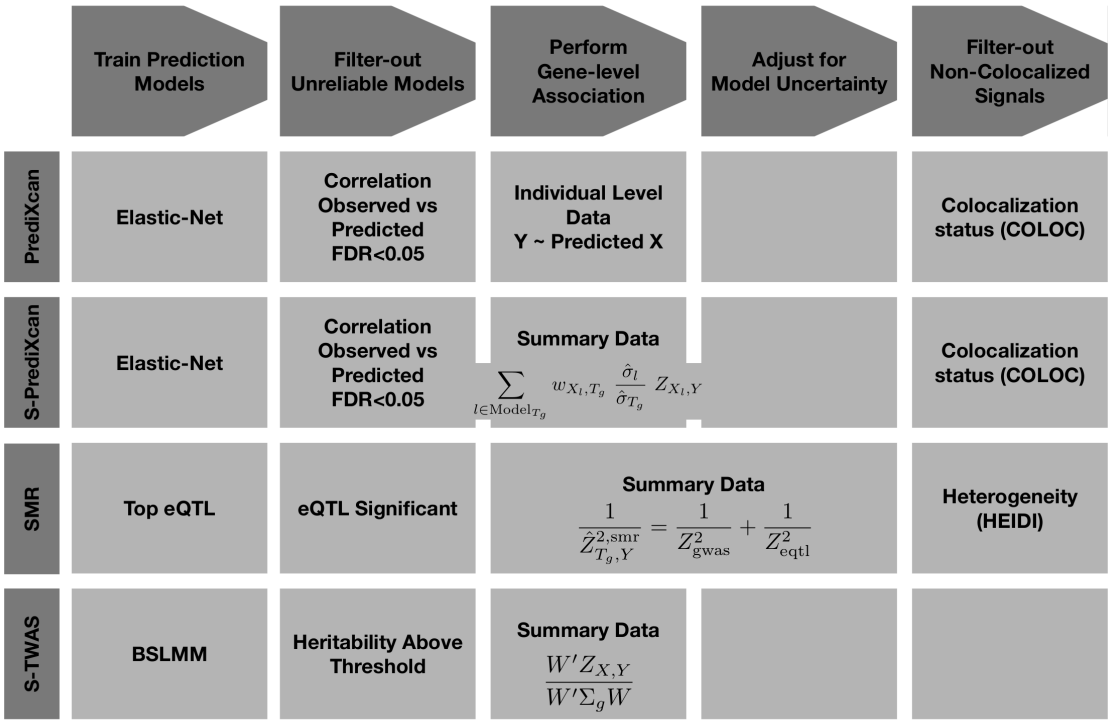


Figure 6. MetaXcan Framework The figure shows a general framework (MetaXcan) which encompasses methods such as PrediXcan, TWAS, SMR, COLOC among others.

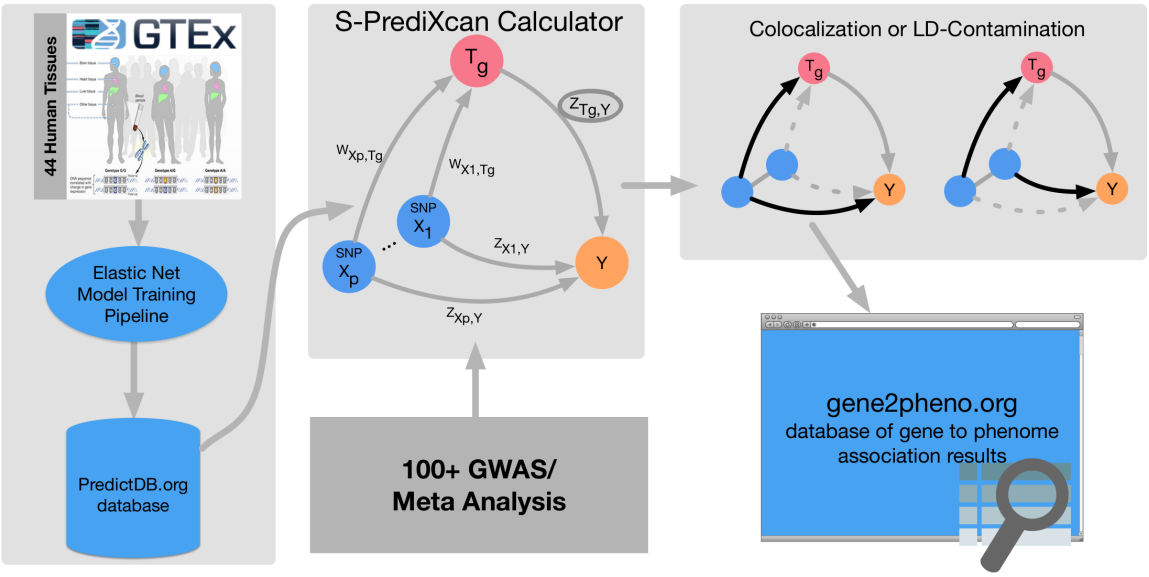


Figure 7. MetaXcan Framework application. This figure summarizes the application of the MetaXcan framework with S-PrediXcan using 44 GTEx tissue transcriptomes and over 100 GWAS and meta analysis results. We trained prediction models using elastic-net [26] and deposited the weights and SNP covariances in the publicly available resource (`http://predictdb.org/`). The weights, covariances and over 100 GWAS summary results were processed with S-PrediXcan. Colocalization status was computed and the full set of results was deposited in `gene2pheno.org`.

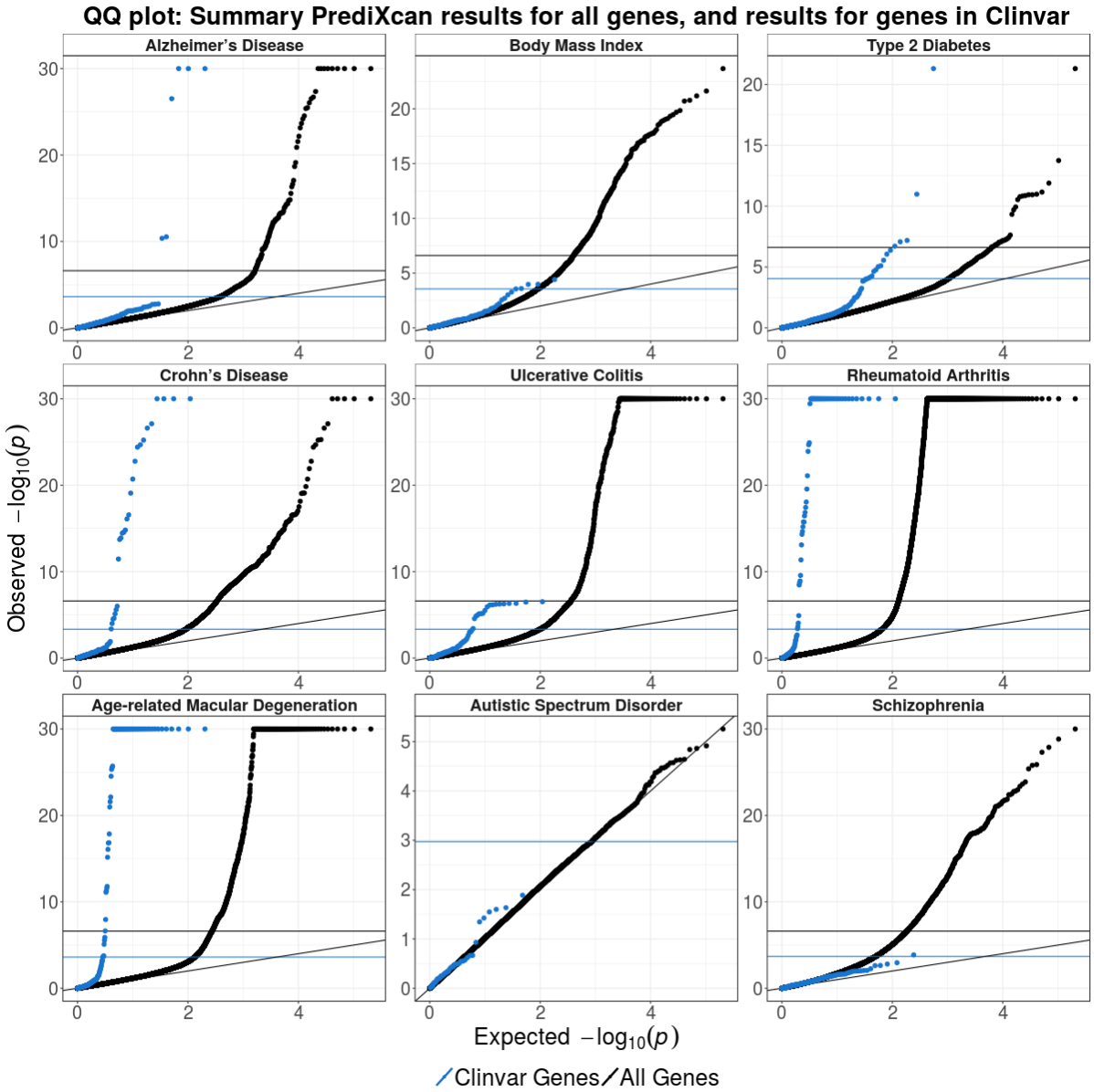


Figure 8. ClinVar genes show significant S-PrediXcan associations. Genes implicated in ClinVar tended to be more significant in S-PrediXcan for most diseases tested, except for schizophrenia and autism. This suggests that more moderate alteration of monogenic disease genes may contribute in a continuum of more moderate but related phenotypes. Alternatively, a more complex interplay between common and rare variation could be taking place such as higher tolerance to loss of function mutations in lower expressing haplotypes which could induce association with predicted expression. Blue circles correspond to the QQ plot of genes in ClinVar that were annotated with the phenotype and black circles correspond to all genes.

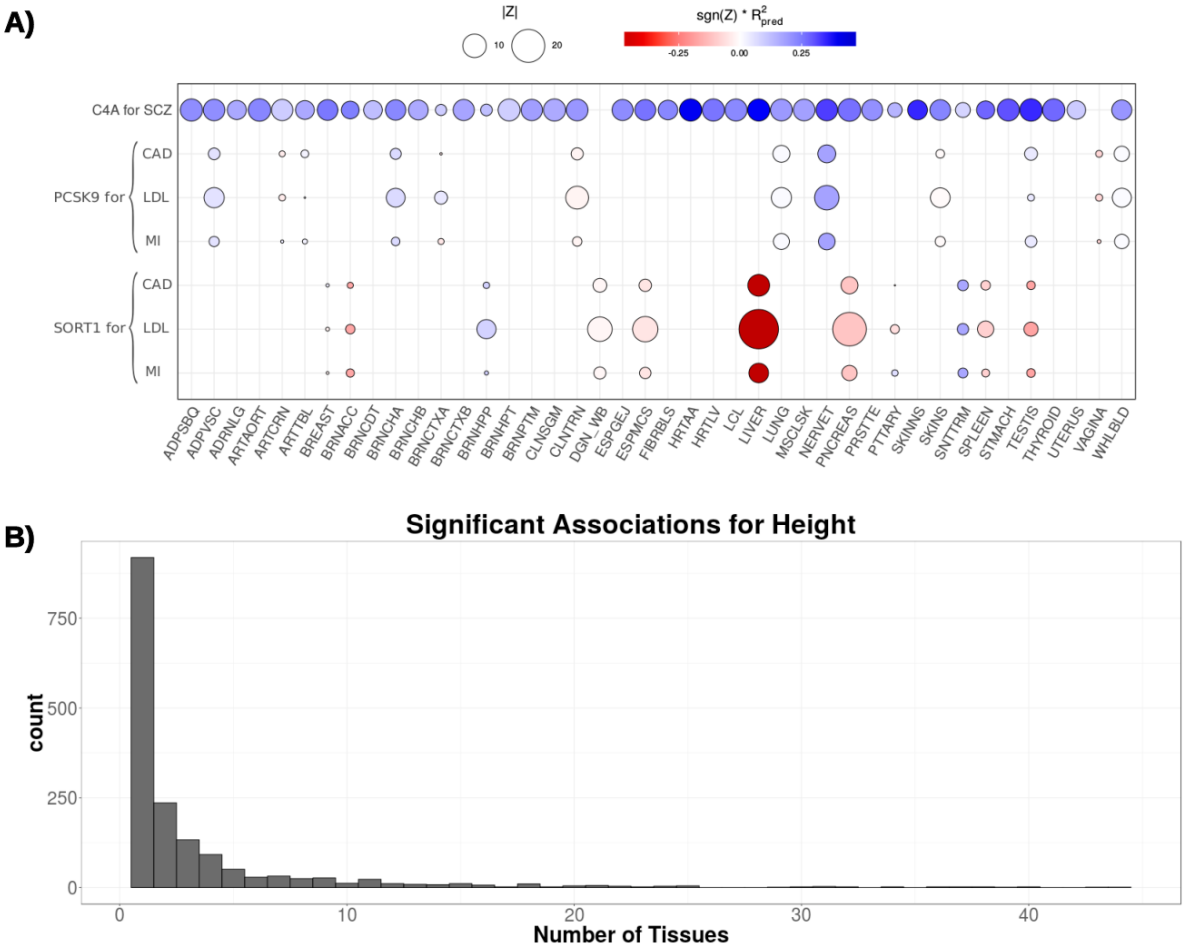


Figure 9. A) S-PrediXcan association for *PCSK9*, *SORT1*, and *C4A* by tissue. This figure shows the association strength between three well studied genes and corresponding phenotypes. *C4A* associations with schizophrenia (SCZ) are significant across most tissues. *SORT1* associations with LDL-C, coronary artery disease (CAD), and myocardial infarction (MI) are most significant in liver. *PCSK9* associations with LDL-C, coronary artery disease (CAD), and myocardial infarction (MI) are most significant in tibial nerve.
The size of the points represent the significance of the association between predicted expression and the traits indicated on the top labels. Red indicates negative correlation whereas blue indicates positive correlation. R^2_{pred} is a performance measure computed as the correlation squared between observed and predicted expression, cross validated in the training set. Darker points indicate larger genetic component and consequently more active regulation in the tissue.
B) Tissue specificity of most trait associations. This figure shows a histogram of the number of tissues for which a gene is significantly associated with height (other phenotypes show similar pattern).
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).

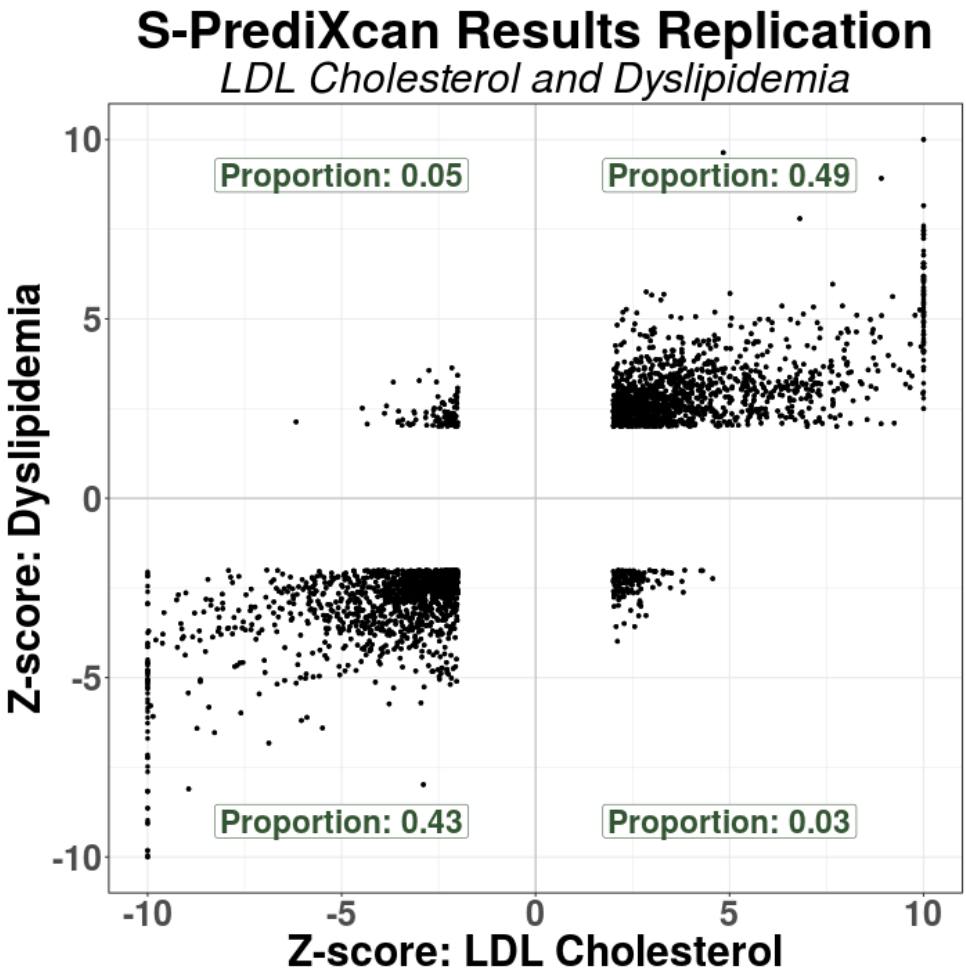


Figure 10. Discovery and replication Z-scores for lipid trait. This figure shows the Z-scores of the association between dyslipidemia (GERA) and predicted gene expression levels on the vertical axis and the Z-scores for LDL cholesterol on the horizontal axis. To facilitate visualization, very large Z-scores where thresholded to 10. Proportions in each quadrant were computed excluding Z-scores with magnitude smaller than 2 to filter out noise.

- dosage and alternate allele are assumed to be the same.

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

- study set (e.g. GWAS study set),
- training set (e.g. GTEx, DGN),
- population reference set (e.g. the training set or 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 can also be 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 `https://github.com/hakyimlab/MetaXcan`.

Association enrichment

We display the enrichment for selected phenotypes in Supplementary Figure 5, measured as mean(Z^2). For visualization purposes, we selected 25 phenotypes from different categories such as anthropometric traits, cardiometabolic traits, autoimmune diseases, and psychiatric conditions (please see figure caption for the list of selected phenotypes). The simple mean of Z^2 for all gene-tissue pairs in a phenotype was taken.

Derivation of Summary-PrediXcan Formula

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

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 $p \times n$ 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, typically $p << n$).

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 their standard error), 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

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

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

Also from the properties of linear regression we know that

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

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

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

Thus:

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

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 $p = 2\Phi(-|Z_g|)$ where $\Phi(.)$ is the normal CDF function.

$$\begin{aligned}
Zg &= \frac{\hat{\gamma}_g}{\mathrm{se}(\hat{\gamma}_g)} \\
&= \sum_{l\in\mathrm{Model}_g} \frac{w_{lg}\hat{\beta}_l\hat{\sigma}_l^2}{\hat{\sigma}_g^2}\sqrt{\frac{n}{\hat{\sigma}_Y^2}\frac{\hat{\sigma}_g^2}{(1-R_g^2)}} && \text{using Eq. 5 and 6} \\
&= \sum_{l\in\mathrm{Model}_g} \frac{w_{lg}\hat{\beta}_l\hat{\sigma}_l^2}{\hat{\sigma}_g}\sqrt{\frac{(1-R_l^2)}{\mathrm{se}(\hat{\beta}_l)^2\hat{\sigma}_l^2}}\sqrt{\frac{1}{(1-R_g^2)}} \\
&= \sum_{l\in\mathrm{Model}_g} w_{lg}\,\frac{\hat{\sigma}_l}{\hat{\sigma}_g}\,\frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)}\sqrt{\frac{1-R_l^2}{1-R_g^2}} && (10) \\
&\approx \sum_{l\in\mathrm{Model}_g} w_{lg}\,\frac{\hat{\sigma}_l}{\hat{\sigma}_g}\,\frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} && (11)
\end{aligned}$$

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.

Calculation of σ_g in reference set

The variance of predicted expression is computed using equation (4) which takes weights for each SNP in the prediction model and the correlation (LD) between the SNPs. The correlation is computed in a reference set such as 1000G or in the training set.

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 (dbGaP Accession phs000424.v6.p1). Sample sizes for different tissues range from 70 (Uterus) to 361 (Muscle - Skeletal). The models referenced in this paper make use of the GTEx Project's V6p data, a patch to the version 6 data and makes use of improved gene-level annotation. 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 gene's 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 q-value less than 0.05 [40]. 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.

Online Catalog and SMR, COLOC, TWAS

Supplementary Table 4 shows the list of GWA/GWAMA studies we considered in this analysis. We applied S-PrediXcan to these studies using the transcriptome models trained on GTEx studies for patched version 6. For simplicity, S-PrediXcan only considers those SNPs that have a matching set of alleles in the prediction model, and adjusts the dosages ($2 - \text{dosage}$) if the alleles are swapped.

To make the results of this study broadly accessible, we built a Postgre SQL relational database to store S-PrediXcan results, and serve them via a web application `http://gene2pheno.org`.

We also applied SMR [16] to the same set of GWAMA studies, using the GTEx eQTL associations. We downloaded version 0.66 of the software from the SMR website, and ran it using the default parameters. We converted the GWAMA and GTEx eQTL studies to SMR input formats. In order to have SMR compute the colocalization test, for those few GWAMA studies where allele frequency was not reported, we filled in with frequencies from the 1000 Genomes Project [41] as an approximation. We also used the 1000 Genomes genotype data as reference panel for SMR.

Next we ran COLOC [18] over the same set of GWAMA and eQTL studies. We used the R package available from CRAN. We used the Approximate Bayes Factor colocalization analysis, with effect sizes, their standard errors, allele frequencies and sample sizes as arguments. When the frequency information was missing from the GWAS, we filled in with data from the 1000 Genomes Project.

For comparison purposes, we have also included the results of the application of Summary-TWAS to 30 traits publicly shared by the authors [24].

Comparison with TWAS

Formal similarity with TWAS can be made more explicit by rewriting S-PrediXcan 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 (11) as

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

and in 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

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

This equation has the same form as the TWAS expression if we use the scaled weight vector $\tilde{W}_g$ instead of W_g. Summary-TWAS imputes the Z-score for the gene-level result assuming that under the null hypothesis, the Z-scores are normally distributed with the same correlation structure as the SNPs; whereas in S-PrediXcan we compute the results of PrediXcan using summary statistics. Thus, S-TWAS and S-PrediXcan yield equivalent mathematical expressions (after setting the factor $\sqrt{\frac{1-R_l^2}{1-R_g^2}} \approx 1$).

Summary-PrediXcan with only top eQTL as predictor

The S-PrediXcan formula when only the top eQTL is used to predict the expression level of a gene can be expressed as

$$\begin{aligned} Z_{\text{s-prediXcan}} &= \sum_{l \in \text{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)} && \text{using 11} \\ &= w_{1g} \frac{\hat{\sigma}_1}{\sqrt{w_{1g}^2 \sigma_1^2}} Z_1 && \text{only top eQTL is in the model} \\ &= Z_1 \end{aligned}$$

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

$$Z^2_{\text{top eqtl s-predixcan}} = Z^2_{\text{GWAS}} \tag{12}$$

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:

$$T_{\text{SMR}} = \frac{Z^2_{\text{eqtl}} Z^2_{\text{gwas}}}{Z^2_{\text{eqtl}} + Z^2_{\text{gwas}}} \tag{13}$$

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.

This SMR statistic is quite different from a χ_1-square random variable as assumed in [16]. A quick simulation shows that the mean of T_{SMR} is 1/4 of the mean of a χ_1-square random variable, as shown in Figure 5-B and -C. Only in two extreme cases, the chi-square approximation holds: when $Z_{\text{eqtl}} >> Z_{\text{gwas}}$ or $Z_{\text{eqtl}} << Z_{\text{gwas}}$. In these extremes, we can apply Taylor expansions to find an interpretable form of the SMR statistic.

If $Z_{\text{eqtl}} >> Z_{\text{gwas}}$, i.e. if the eQTL association is much more significant than the GWAS association,

$$\begin{aligned} T_{\text{SMR}} &= \frac{Z^2_{\text{gwas}}}{1 + \frac{Z^2_{\text{gwas}}}{Z^2_{\text{eqtl}}}} \\ &\approx Z^2_{\text{gwas}} \left(1 - \frac{Z^2_{\text{gwas}}}{Z^2_{\text{eqtl}}}\right) \qquad (14) \\ &\quad \text{so for large enough } Z^2_{\text{eqtl}} \text{ relative to } Z^2_{\text{gwas}} \\ &\approx Z^2_{\text{gwas}} = Z^2_{\text{top eqtl s-predixcan}} \qquad (15) \end{aligned}$$

with the last equality from 12. Thus, in this case, the SMR statistic is slightly smaller than the (top eQTL based) S-PrediXcan χ_1-square. This reduced significance is accounting for the uncertainty in the eQTL association. As the evidence for eQTL association grows, the denominator Z^2_{eqtl} increases and the difference tends to 0.

On the other extreme when the GWAS association is much stronger than the eQTL's, $Z_{\text{eqtl}} << Z_{\text{gwas}}$,

$$
\begin{aligned}
T_{\text{SMR}} &= \frac{Z_{\text{eqtl}}^2}{1 + \frac{Z_{\text{eqtl}}^2}{Z_{\text{gwas}}^2}} \\
&\approx Z_{\text{eqtl}}^2 \left(1 - \frac{Z_{\text{eqtl}}^2}{Z_{\text{gwas}}^2}\right) \qquad (16) \\
&\quad \text{so for large enough } Z_{\text{gwas}}^2 \text{ relative to } Z_{\text{eqtl}}^2 \\
&\approx Z_{\text{eqtl}}^2 \qquad (17)
\end{aligned}
$$

In both extremes, the SMR statistic significance is approximately equal to the less significant of the two statistics GWAS or eQTL, albeit strictly smaller.

In between the two extremes, the right distribution must be computed using numerical methods. When we look at the empirical distribution of the SMR statistic's p-value against the GWAS and eQTL (top eQTL for the gene) p-values, we find the ceiling of the SMR statistic is maintained as shown in Figure 5-E and -F. Supplementary Figure 12 shows a comparison of colocalization proportions between SMR and PrediXcan.

GERA imputation

Genotype files were obtained from dbGaP, and updated to release 35 of the probe annotations published by Affymetrix via PLINK [42]. Probes were filtered out that had a minor allele frequency of <0.01, were missing in >10% of subjects, or did not fit Hardy-Weinberg equilibrium. Subjects were dropped that had an unexpected level of heterozygosity (F >0.05). Finally the HRC-1000G-check-bim.pl script (from `http://www.well.ox.ac.uk/~wrayner/tools/`) was used to perform some final filtering and split data by chromosome. Phasing (via eagle v2.3 [43]) and imputation against the HRC r1.1 2016 panel [44] (via minimac3) were carried out by the Michigan Imputation Server [45].

GERA GWAS and MetaXcan Application

European samples had been split into ten groups during imputation to ease the computational burden on the Michigan server, so after obtaining the imputed .vcf files, we used the software PLINK [42] to convert the genotype files into the PLINK binary file format and merge the ten groups of samples together, while dropping any variants not found in all sample groups. For the association analysis, we performed

a logistic regression using PLINK, and following QC practices from [14] we filtered out individuals with genotype missingness > 0.03 and filtered out variants with minor allele frequency < 0.01, missingness > 0.05, out of Hardy-Weinberg equilibrium significant at 1e-6, or had imputation quality < 0.8. We used gender and the first ten genetic principal components as obtained from dbGaP as covariates. Following all filtering, our analysis included 61,444 European samples with 7,120,064 variants. MetaXcan was then applied to these GWAS results, using the 45 prediction models (GTEx and DGN).

Acknowledgments

Grants

We acknowledge the following US National Institutes of Health grants: R01MH107666 (H.K.I.), T32 MH020065 (K.P.S.), R01 MH101820 (GTEx), P30 DK20595 (Diabetes Research and Training Center), F31 DK101202 (J.M.T.), 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.

This work was completed in part with resources provided by the University of Chicago Research Computing Center, Bionimbus [46], and the Center for Research Informatics. The Center for Research Informatics is funded by the Biological Sciences Division at the University of Chicago with additional funding provided by the Institute for Translational Medicine, CTSA grant number UL1 TR000430 from the National Institutes of Health.

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. Aguet F, Brown AA, Castel S, Davis JR, Mohammadi P, Segre AV, et al. Local genetic effects on gene expression across 44 human tissues. bioRxiv. 2016;Available from: http://biorxiv.org/content/early/2016/09/09/074450.

11. 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.

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. 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.

16. 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.

17. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. The American Journal of Human Genetics. 2013 May;92(5):667–680. Available from: http://dx.doi.org/10.1016/j.ajhg.2013.03.022.

18. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genetics. 2014 May;10(5):e1004383. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24830394&retmode=ref&cmd=prlinks.

19. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al.; Los Angeles Los Angeles CA 90095 USA. Department of Computer Science. Colocalization of GWAS and eQTL Signals Detects Target Genes. Am J Hum Genet. 2016;Available from: http:://dx.doi.org/10.1016/j.ajhg.2016.10.003.

20. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genetics. 2017 Mar;13(3):e1006646. Available from: http://dx.plos.org/10.1371/journal.pgen.1006646.

21. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007;447(7145):661–78. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17554300.

22. Manor O, Segal E. Robust Prediction of Expression Differences among Human Individuals Using Only Genotype Information. PLoS Genetics. 2013 Mar;9(3):e1003396. Available from: http://dx.plos.org/10.1371/journal.pgen.1003396.

23. Hamilton N. ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams; 2016. R package version 2.2.0. Available from: https://CRAN.R-project.org/package=ggtern.

24. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American Journal of Human Genetics. 2017 Mar;100(3):473–487. Available from: http://dx.doi.org/10.1016/j.ajhg.2017.01.031.

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

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

27. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, Cox NJ, et al. Survey of the Heritability and Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS Genetics. 2016;12(11).

28. Pavlides JMW, Zhu Z, Gratten J, Mcrae AF, Wray NR, Yang J. Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. Genome medicine. 2016 Aug;8(1):1–6. Available from: http://dx.doi.org/10.1186/s13073-016-0338-4.

29. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics. 2013 Sep;45(10):1238–1243. Available from: http://www.nature.com/doifinder/10.1038/ng.2756.

30. 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.

31. Shah N, Hou YCC, Yu HC, Sainger R, Dec E, Perkins B, et al. Identification of misclassified ClinVar variants using disease population prevalence. 2016 Sep;p. 1–23. Available from: http://biorxiv.org/lookup/doi/10.1101/075416.

32. 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;530(7589):177–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26814963{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4752392.

33. 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;466(7307):714–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20686566{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3062476.

34. 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.

35. 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.

36. Hoffmann TJ, Ehret GB, Nandakumar P, Ranatunga D, Schaefer C, Kwok PY, et al. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nature Genetics. 2016 Nov;49(1):54–64. Available from: http://www.nature.com/doifinder/10.1038/ng.3715.

37. Cook JP, Morris AP. Multi-ethnic genome-wide association study identifies novel locus for type 2 diabetes susceptibility. European Journal of Human Genetics. 2016 Aug;24(8):1175–1180. Available from: http://www.nature.com/doifinder/10.1038/ejhg.2016.17.

38. Torres JM, Barbeira AN, Bonazzola R, Morris AP, Shah KP, Wheeler HE, et al. Integrative cross tissue analysis of gene expression identifies novel type 2 diabetes genes. bioRxiv. 2017;Available from: http://biorxiv.org/content/early/2017/02/27/108134.

39. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017 Jun;169(7):1177–1186. Available from: http://dx.doi.org/10.1016/j.cell.2017.05.038.

40. Storey JD. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003 Jul;100(16):9440–9445. Available from: http://scholar.google.com.proxy.uchicago.edu/scholar?hl=en&lr=&q=info:eSXwkHMI-nQJ:scholar.google.com/&output=search.

41. Auton A, Altshuler DM, Durbin RMJA, Wang J, Yang H, Auton A, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. Available from: http://www.nature.com/nature/journal/v526/n7571/fig{_}tab/nature15393{_}SF1.html{%}5Cnhttp://dx.doi.org/10.1038/nature15393{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26432245{%}5Cnhttp://www.nature.com/doifinder/10.1038/nature15393.

42. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015 dec;4(1):7. Available from: http://gigascience.biomedcentral.com/articles/10.1186/s13742-015-0047-8.

43. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics. 2016 oct;48(11):1443–1448. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27694958http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5096458http://www.nature.com/doifinder/10.1038/ng.3679.

44. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics. 2016 aug;48(10):1279–1283. Available from: http://www.nature.com/doifinder/10.1038/ng.3643.

45. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nature Genetics. 2016 aug;48(10):1284–1287. Available from: http://www.nature.com/doifinder/10.1038/ng.3656.

46. Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, et al. Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets. Journal of the American Medical

Informatics Association : JAMIA. 2014 Nov;21(6):969–975. Available from: https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-002155.

47. 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).

Supplementary Material

S-PrediXcan vs. PrediXcan in simulated data

We measure performance as the ability of S-PrediXcan to infer accurate PrediXcan results. Differences arise mostly because of LD differences between reference and study sets. We use genotype data from different populations from 1000G to assess the robustness to large differences between reference and study sets.

First we simulated normally distributed phenotype (under the null hypothesis of no genetic effect). We use prediction models trained on Depression Genes and Network's (DGN) Whole Blood data set [5,27] 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-A. For each population combination, we computed PrediXcan association results for the simulated phenotype and compared them with results generated using S-PrediXcan 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 PrediXcan and S-PrediXcan is almost 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 substantial differences between study and reference sets.

S-PrediXcan vs. PrediXcan in real data (cellular phenotype)

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 [47]. 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 S-PrediXcan for different combinations of reference and study sets, using whole blood prediction models trained in the DGN cohort. The results are shown in Figure 2-B. As with our simulation study, the S-PrediXcan 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

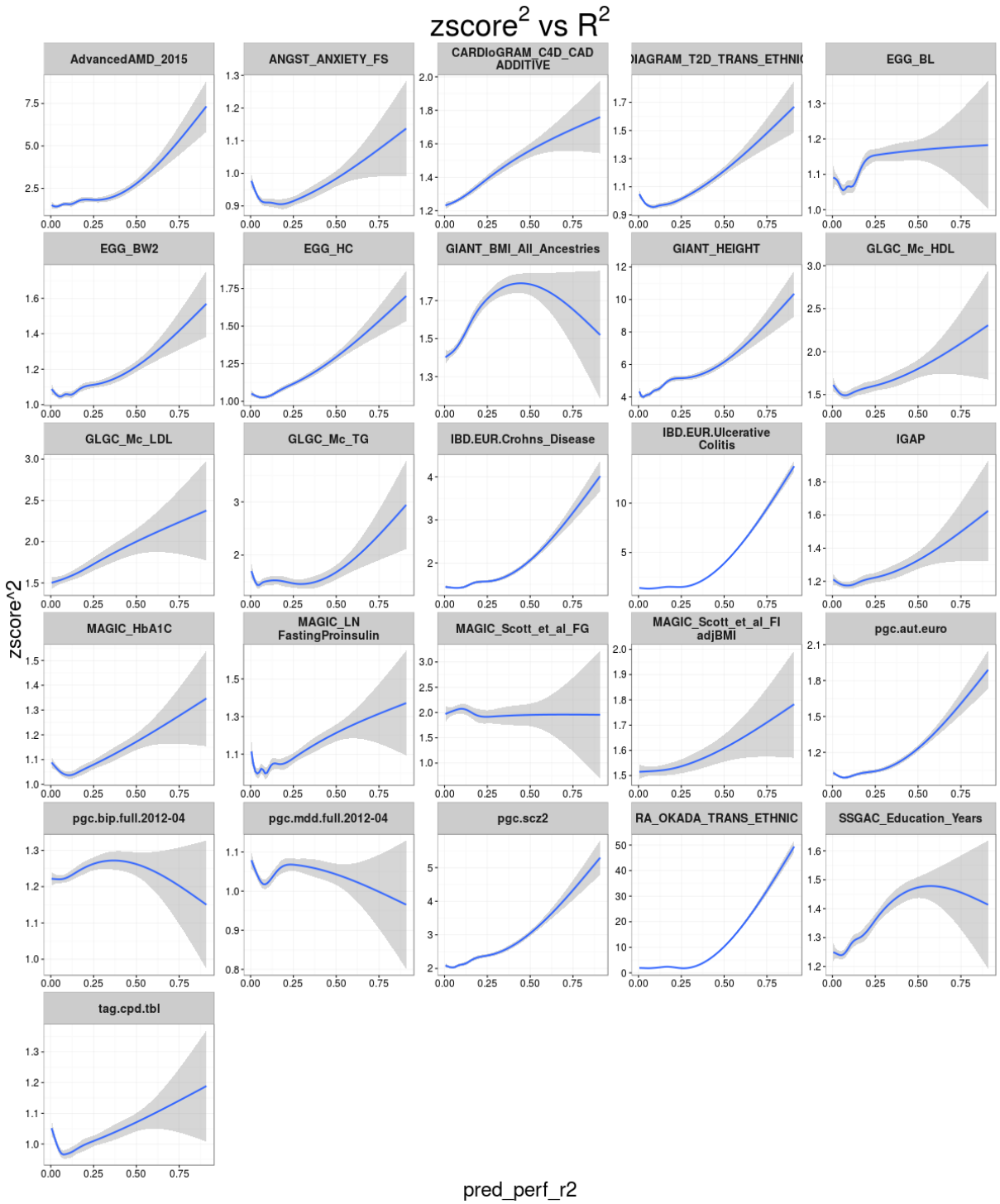
to the plots for the simulated phenotypes, the diagonal concordance is slightly lower than 1. This is due to the fact that more individuals were included in the reference set than in the study set, thus the study and reference sets were not identical for S-PrediXcan.

S-PrediXcan vs. PrediXcan in disease phenotypes from WTCCC

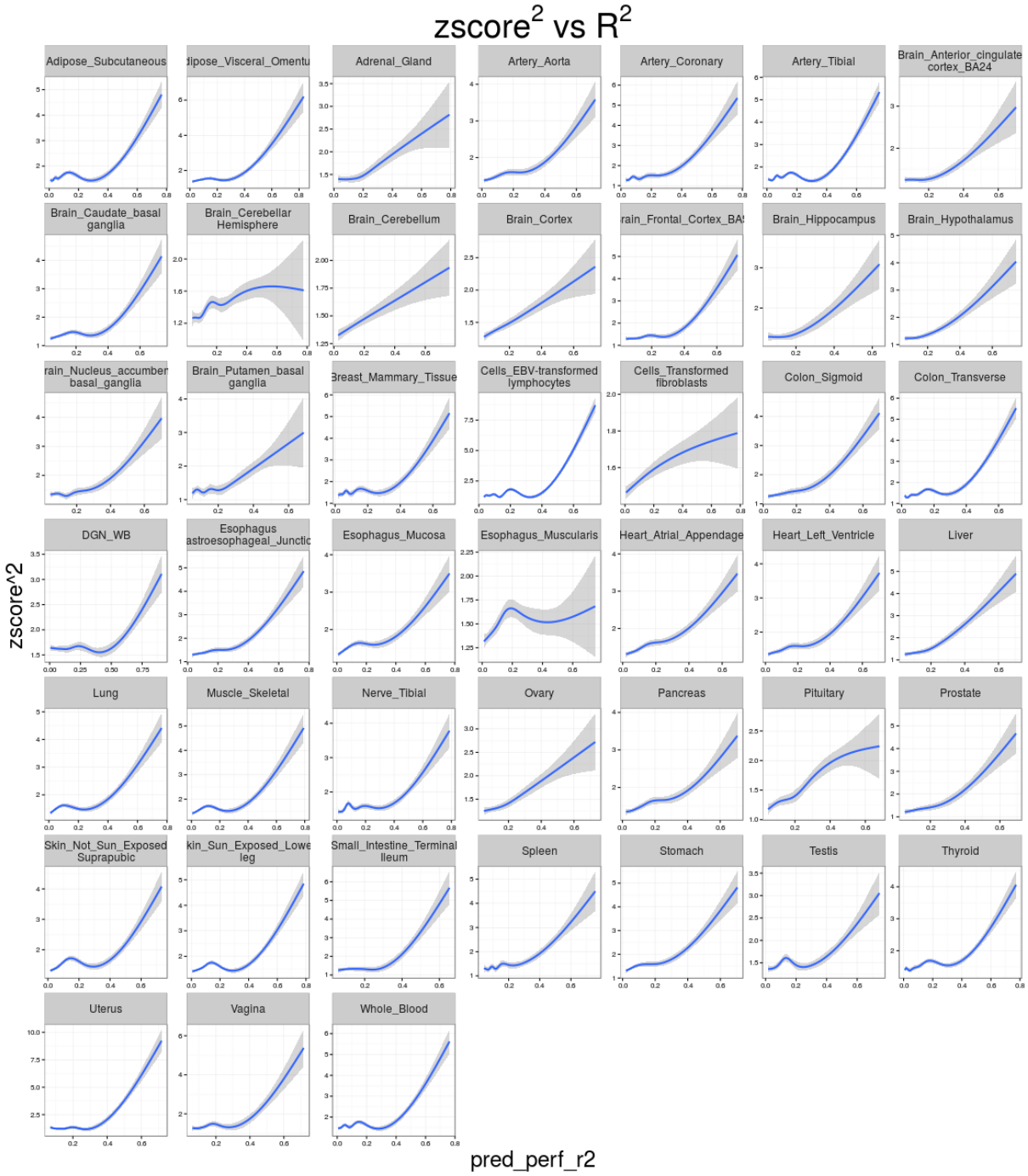
We show the comparison of PrediXcan and summary-PrediXcan results for two diseases: Bipolar Disorder (BD) and Type 1 Diabetes (T1D) from the WTCCC in Figure 2-C. Other diseases exhibited similar performance (data not shown). Concordance between PrediXcan and Summary-PrediXcan is over 99% for both diseases (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 Summary-PrediXcan formula is based on linear regression. As observed before [25], 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.

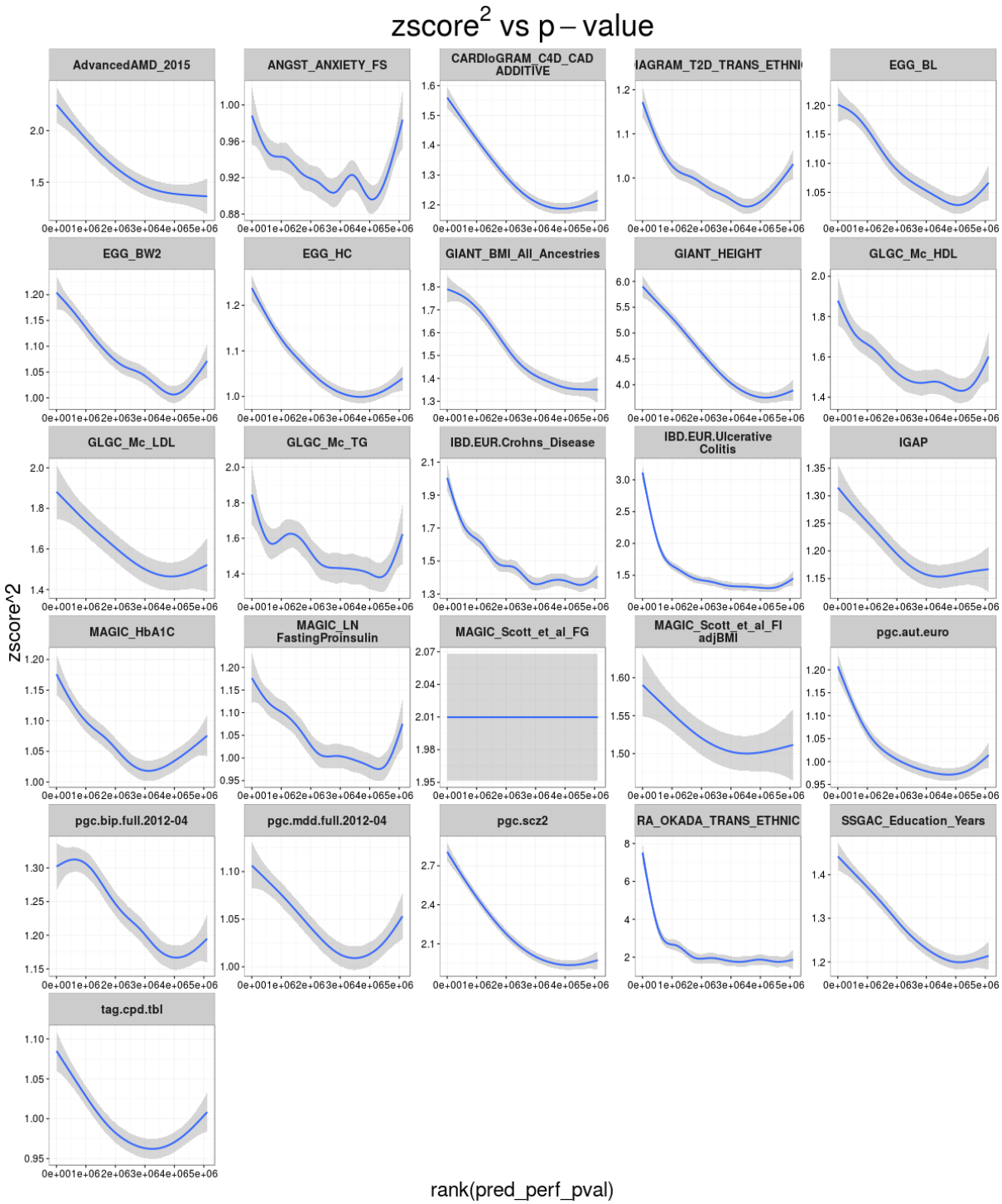
Supplementary Figures



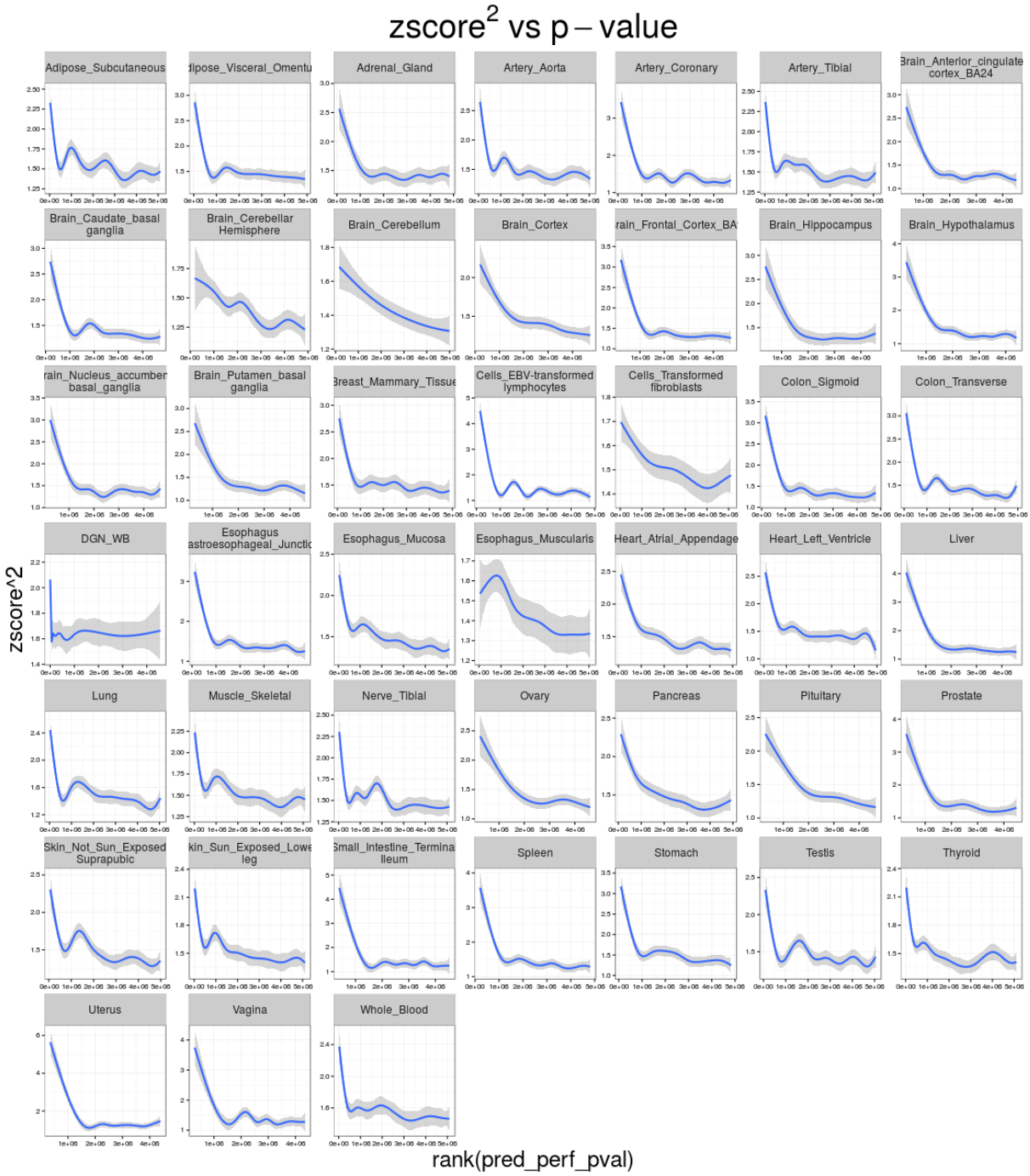
Supplementary Figure 1. Z-score^2 vs predicted performance R^2 by phenotype. When averaged across all genes and tissues within each phenotype the significance of the association tends to be more pronounced as R^2 (a measure of the genetic component) is larger. R^2 is the square of the correlation between predicted and observed expression levels in the training set, evaluated in a cross validated manner.



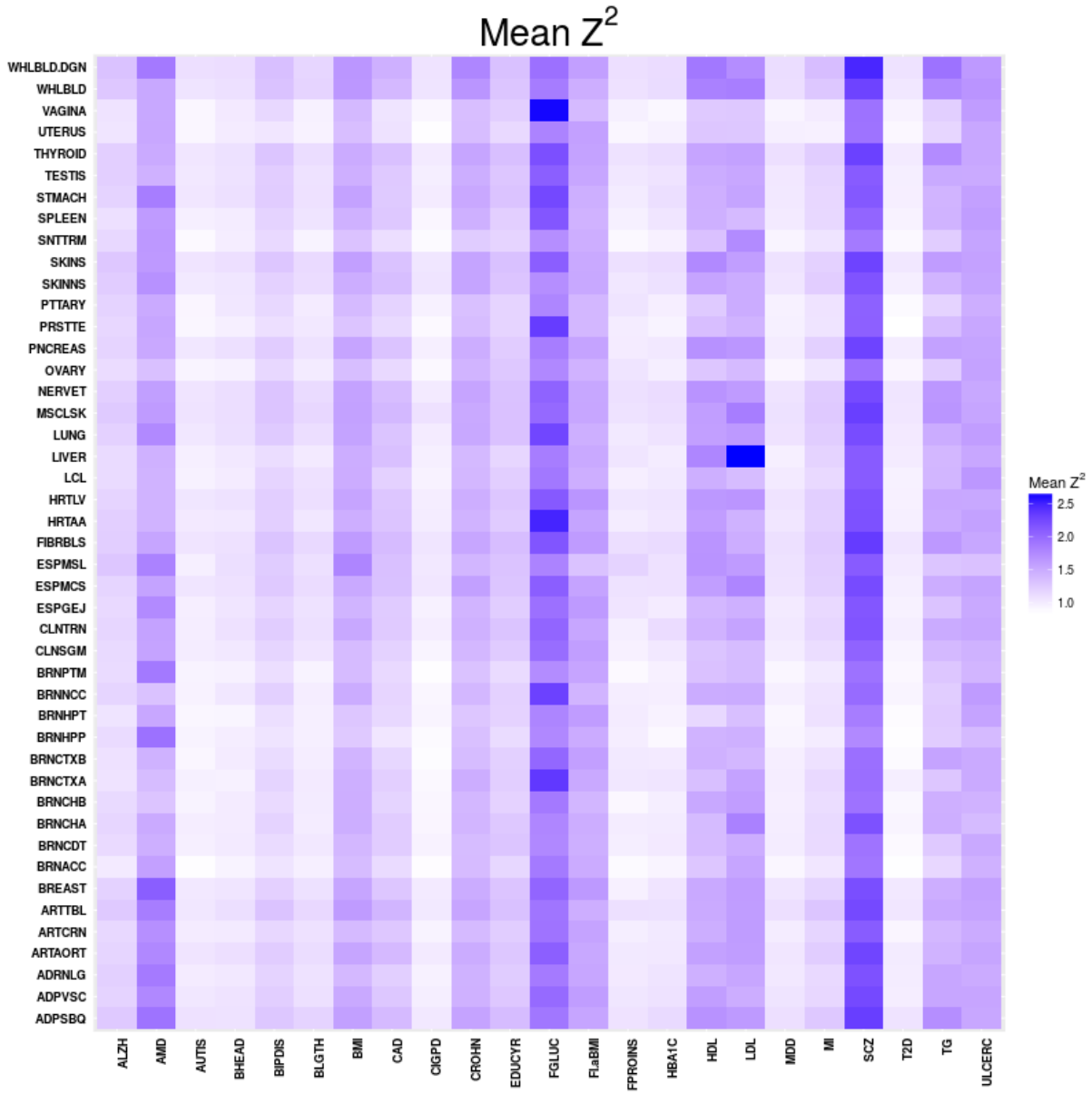
Supplementary Figure 2. Z-score^2 vs predicted performance R^2 by tissue. When averaged across all genes and phenotypes within each tissue the significance of the association tends to be more pronounced as R^2 (a measure of the genetic component) is larger. R^2 is the square of the correlation between predicted and observed expression levels in the training set, evaluated in a cross validated manner.



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. Prediction p-values (or prediction performance p-values) are computed (cross validated) as the p-values of the correlation between predicted and observed expression levels in the training set under the null hypothesis of no correlation.

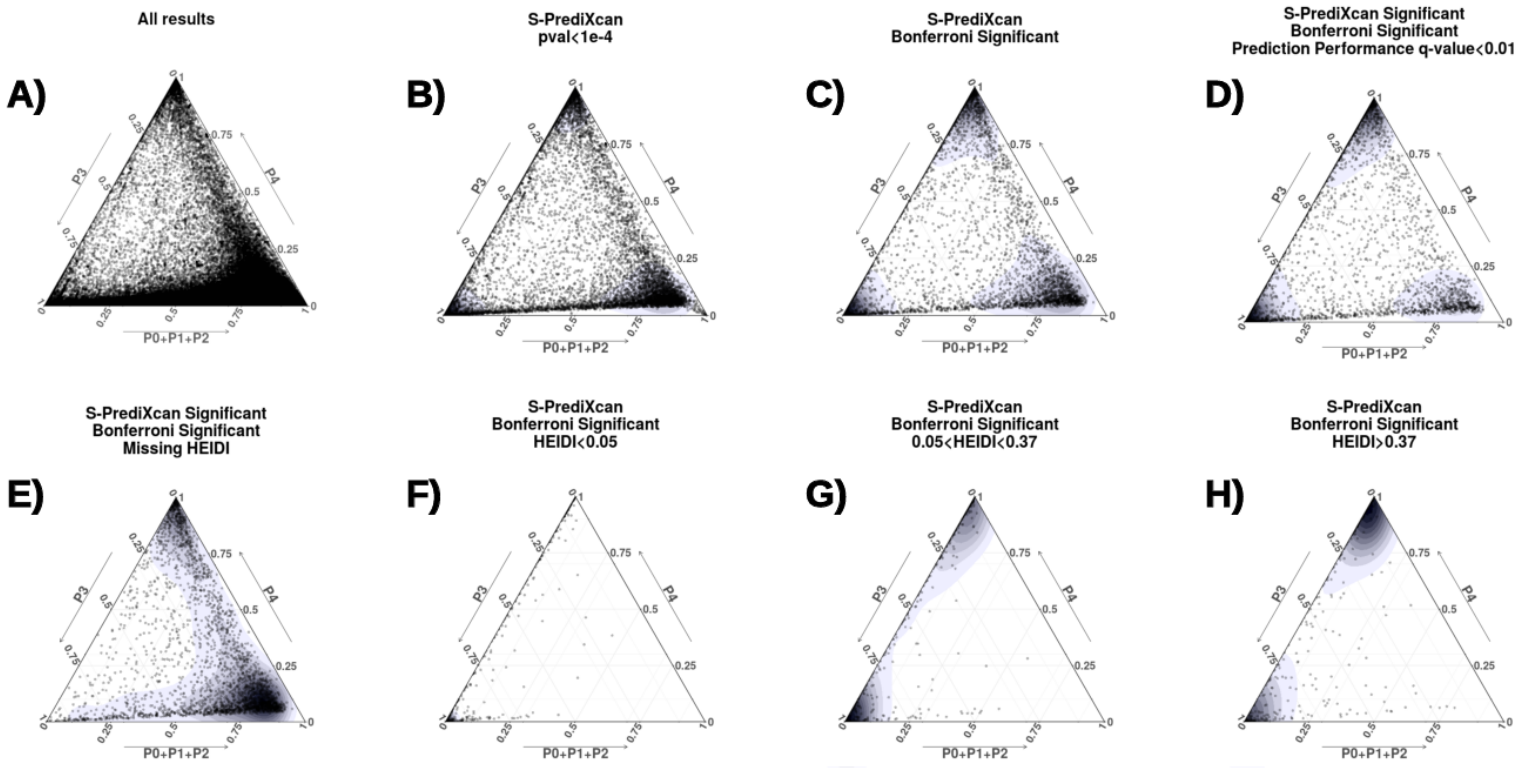


Supplementary Figure 4. Z-score^2 vs predicted performance p-value by tissue. When averaged across all genes and phenotypes within each tissue the significance of the association tends to be more pronounced as the cross validated prediction is more significantly associated with the observed expression. Prediction p-values (or prediction performance p-values) are computed (cross validated) as the p-values of the correlation between predicted and observed expression levels in the training set under the null hypothesis of no correlation.

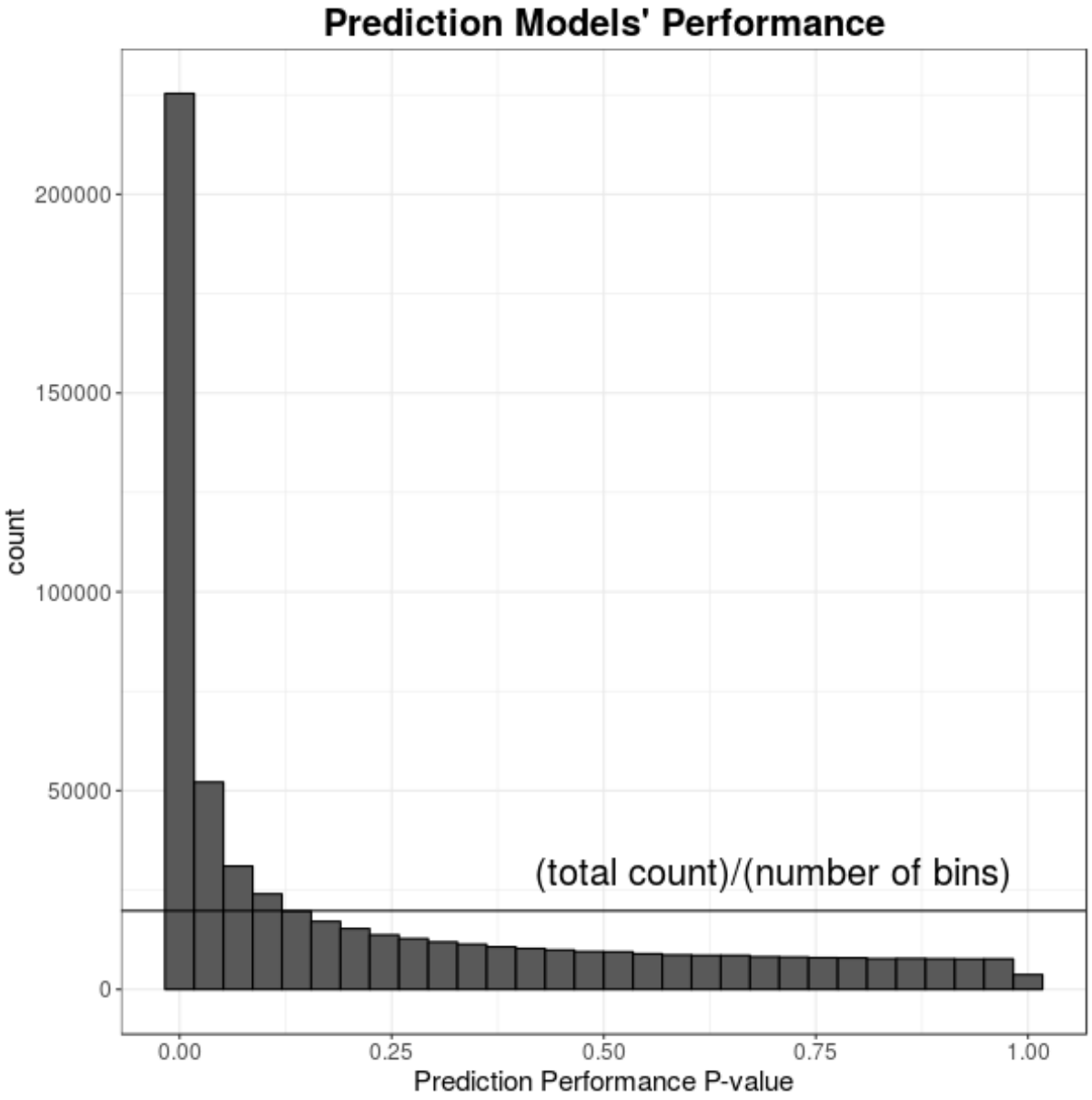


Supplementary Figure 5. Average enrichment of significant genes by tissues. This figure shows the average square of the Z-scores (effect size/standard error) of the association between the genetic component of gene expression levels and phenotype.

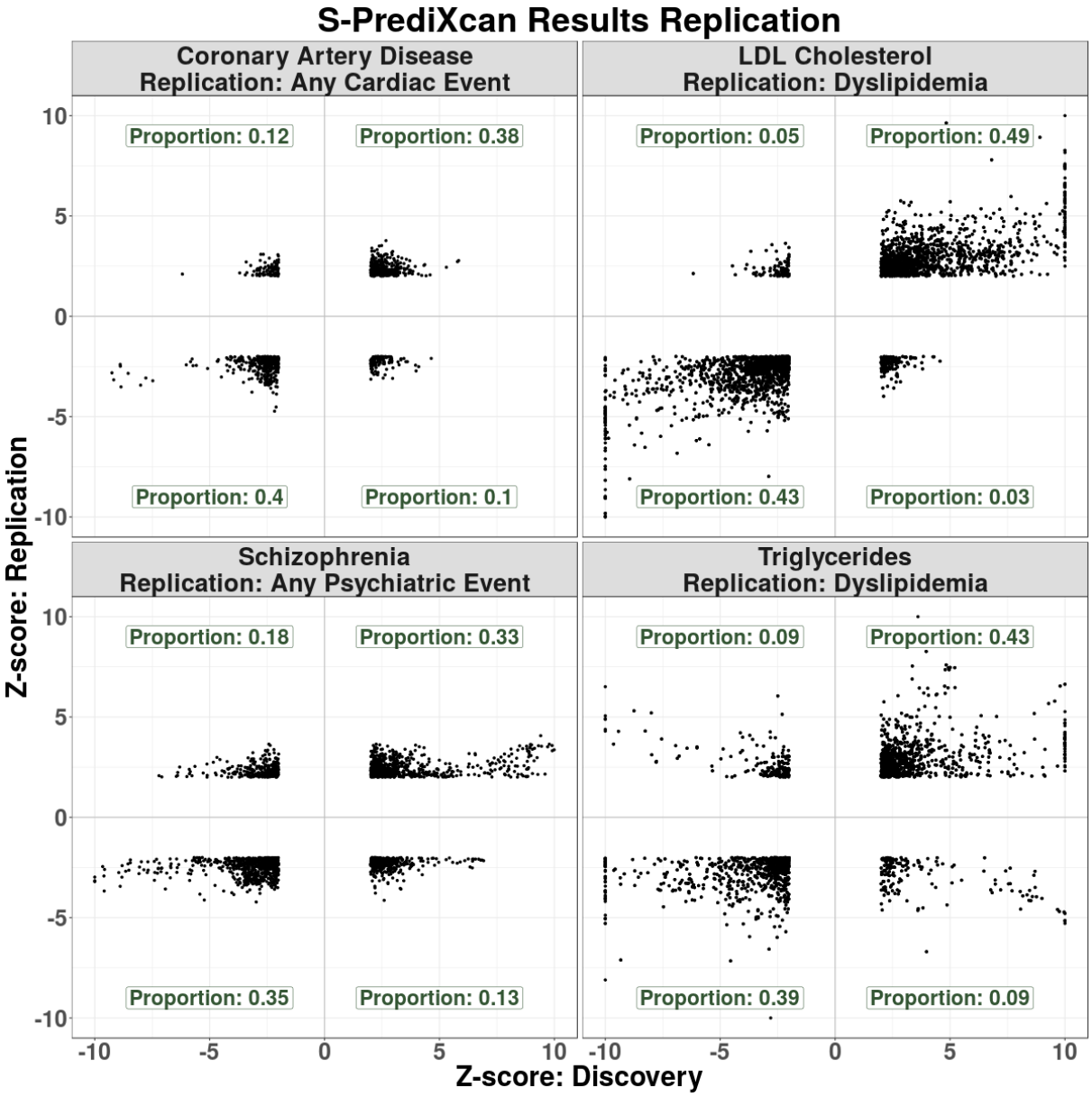
Phenotype abbreviations: CIGPD (cigarettes per day), BMI (body mass index), FGLUC (fasting glucose), T2D (type 2 diabetes), CAD (coronary artery disease), LDL (low-density lipoprotein cholesterol), TG (triglycerides), RA (rheumatoid arthritis), ALZH (alzheimer's disease), HDL (high-density lipoprotein cholesterol), CROHN (Crohn's disease), ULCERC (ulcerative colitis), HEIGHT, BHEAD (birth head circumference), 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).



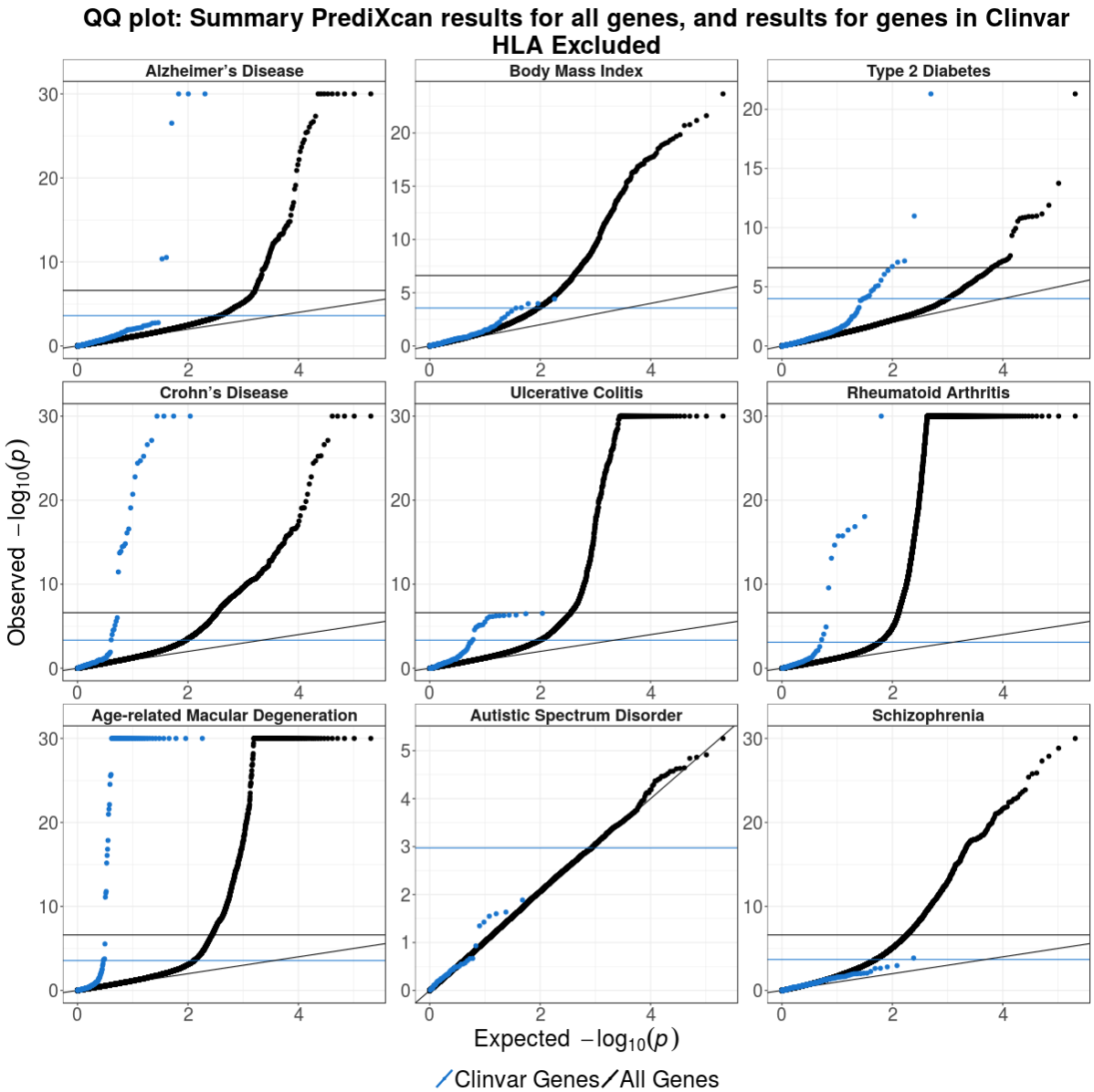
Supplementary Figure 6. Colocalization status of S-PrediXcan results for Height phenotype across all tissues. These ternary plots [23] constrain the values such that the sum of the probabilities is 1. All points in a horizontal line have the same probability of 'colocalized' GWAS and eQTL signals (P4), points on a line parallel to the right side of the triangle (NW to SE) have the same probability of'Independent signals' (P3), and lines parallel to the left side of the triangle (NE to SW) correspond to constant P1+P2+P3. Within each triangle: Top vertex corresponds to high probability of colocalization (P4>0.5), lower left vertex to probability of independent signals (P3>0.5), and lower right vertex corresponds to genes without enough power to determine or reject colocalization. **Panel A** shows that most the genes fall in the 'undetermined' region. When only significant S-PrediXcan associated genes are shown (**Panel B**:p$< 1e-4$ & **Panel C**:p$< 1e-6$), three peaks in each of the regions emerge (interpreted as 'colocalized', 'distinct', 'undetermined'). **Panel D** shows that when genes with low prediction performance are excluded, the 'undetermined' peak significantly diminishes. **Panel E** shows the COLOC probabilities for genes for which HEIDI returned no values. There is a significant peak in the undetermined region, but the density is still significant in other regions. **Panel F** shows genes that have significant HEIDI p-values, evidence of heterogeneity. As expected genes cluster mostly near probability of independent signals. **Panel H** shows genes that have non significant HEIDI p-value. Overall, HEIDI and COLOC tend to agree, although there is a sizable number of cases where the two methods will disagree. Unlike COLOC results, HEIDI does not partition the genes into distinct clusters and an arbitrary cutoff p-value has to be chosen.



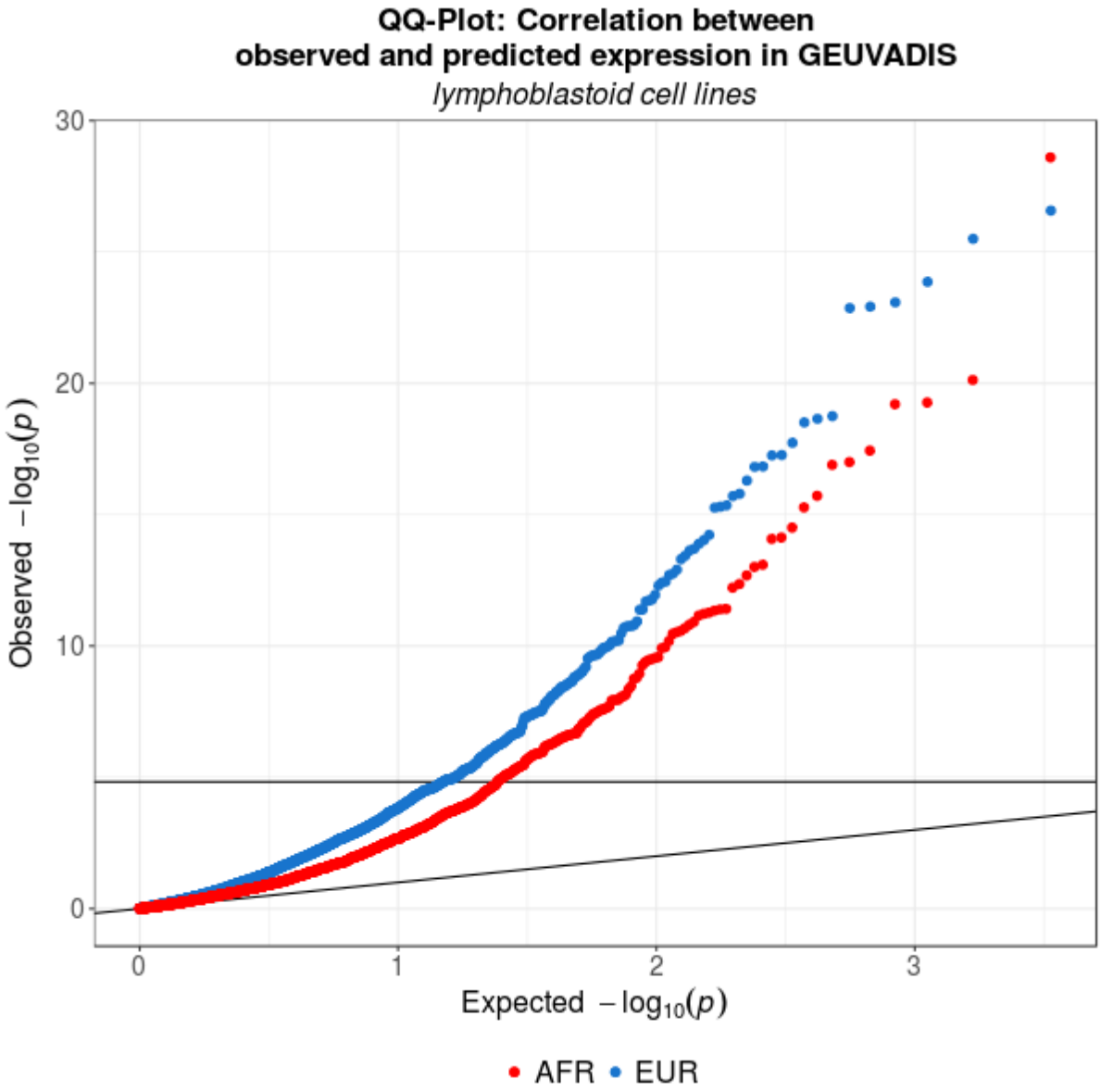
Supplementary Figure 7. Histogram of prediction performance p-value. This figure shows p-values of the correlation between predicted and observed expression levels in the training set. Prediction p-values (or prediction performance p-values) are computed (cross validated) as the p-values of the correlation between predicted and observed expression levels in the training set under the null hypothesis of no correlation.



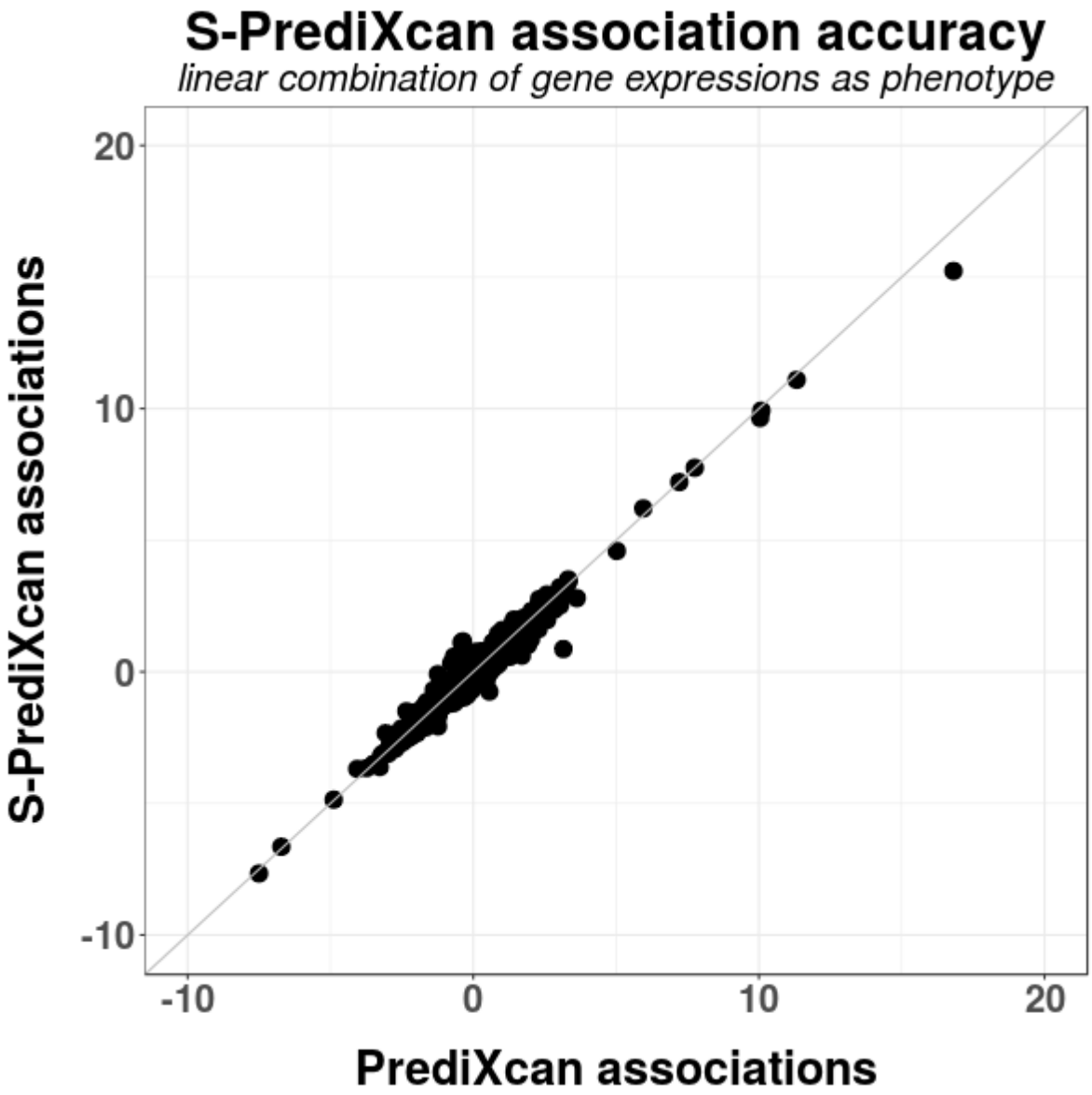
Supplementary Figure 8. Comparison of discovery and replication Z-scores. This figure shows the Z-scores of the discovery phenotype and the matched replication phenotype in GERA. The proportion of concordant direction of effects far exceeds the one with discordant direction of effects. Coronary artery disease has 77% of gene-tissue associations in the same direction of effects as 'Any cardiac event' in GERA. LDL cholesterol shows 92% concordance, TG shows 81%, and schizophrenia shows 67% concordance. Large Z-scores were thresholded to 10 to ease visualization. Proportions in each quadrant were computed excluding Z-scores with magnitude smaller than 2 to filter out noise.



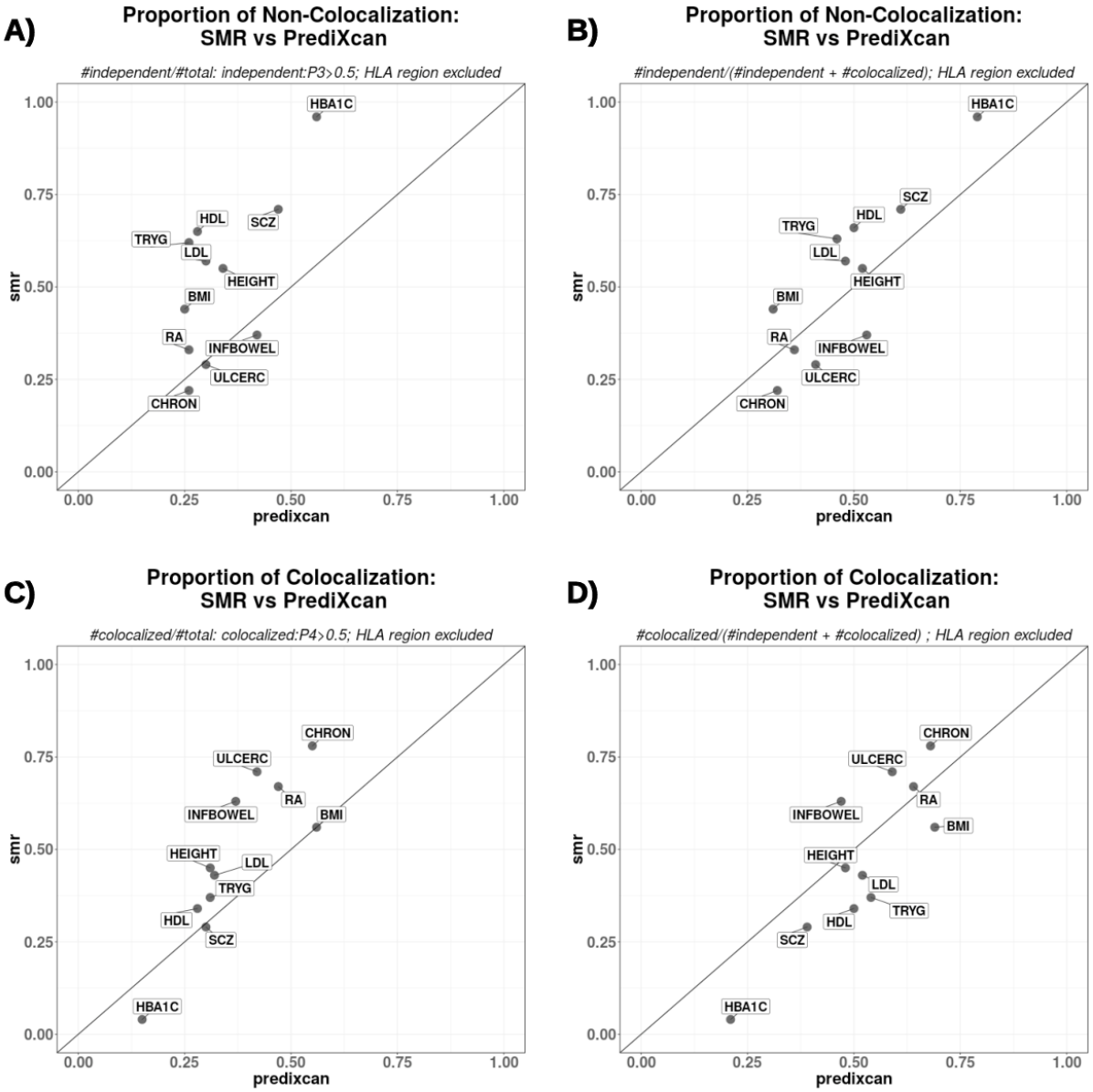
Supplementary Figure 9. ClinVar enrichment of S-PrediXcan associations, excluding genes in the HLA region. Blue circles correspond to the QQ plot of genes in ClinVar that were annotated with the phenotype and black circles correspond to all genes. Genes in the HLA region were excluded because of their complex LD structure, to verify the enrichment robustness. Rheumatoid Arthritis is the only phenotype that experienced a noticeable change, but still displayed significant enrichment.



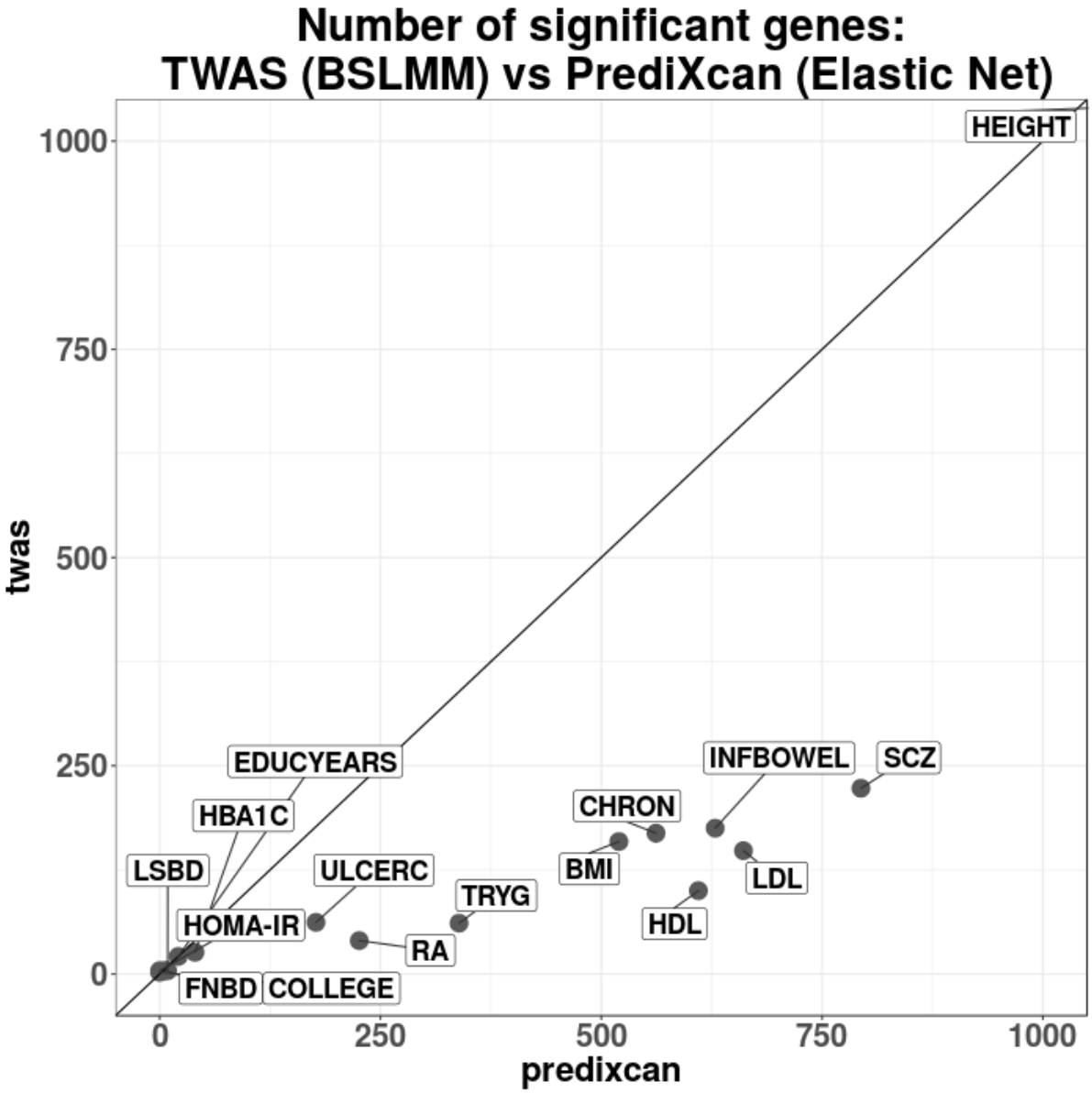
Supplementary Figure 10. Robustness of prediction across populations. Expression was predicted using prediction models trained on GTEx EBV transformed lymphocytes, with mostly European samples. Blue dots display the QQ plot of p-values of the correlation between predicted and observed gene expression levels in 77 European individuals from GEUVADIS [6]. Red dots correspond to the p-values of the correlation for 77 African individuals from GEUVADIS. There is only a small decrease in prediction performance in Africans compared to Europeans. Prediction with other tissue models showed entirely similar behavior.



Supplementary Figure 11. S-PrediXcan Associations for a simulated phenotype under the alternative hypothesis. This figure illustrates S-PrediXcan's performance when the alternative hypothesis is true (i.e. the trait depends linearly on gene expression). We predicted gene expression on European individuals from the 1000 Genomes Project, using a model trained on GTEx Whole Blood study. We selected three genes (*SCYL3*, *MUSTN1*, *GCLC*) and built a phenotype according to $Y = 6T_{SCYL3} + 4T_{GLGC} + 3T_{MUSTN1} + \epsilon$, where T_X is predicted expression for gene X and ϵ is random noise sampled from a normal $N(0, 1)$ distribution. The predicted expression component had standard deviation 1.46, so the noise is comparable to the signal.



Supplementary Figure 12. Comparison of significant results colocalization between PrediXcan and SMR. In these figures, 'colocalized' means high probability of shared eQTL and GWAS signal (P4>0.5), and 'independent' to high probability of distinct eQTL and GWAS signals (P3>0.5).
Panel **A)** shows the proportion of non-colocalized significant associations to total significant associations in PrediXcan and SMR. Panel **B)** shows the proportion to number of colocalized and independent associations.
Panel **C)** shows the proportion of colocalized significant associations (shared eQTL and GWAS signals).
Panel **D)** shows the proportion to number of colocalized and independent associations.
As expected, SMR shows a higher proportion of colocalized and non-colocalized associations than PrediXcan. This can be explained by SMR associations being driven mostly by the top eQTL, whereas PrediXcan can aggregate many small contributions into a significant association that can not be assessed by COLOC. In the case of proportions to total number of colocalized and independent associations, SMR tends to exhibit a slightly larger proportion of independent results and a slightly smaller number of colocalized results.



Supplementary Figure 13. Comparison of number of significant results between PrediXcan and TWAS. This figure contains results from TWAS and PrediXcan for the phenotypes and tissues reported in [24]. Notice that Mancuso et al filtered out genes with low GCTA heritability, which we have shown to underestimate h^2 [27]. This results in much smaller number of genes tested with TWAS than with PrediXcan. This in turn explains the smaller number of significant genes in TWAS despite the fact that when genes are tested, the significance of the two methods is similar as seen in Figure 4-B. Height trait is not shown for visualization purposes, but also exhibited this behavior.

Supplementary Tables

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

Supplementary Table 2. Summary of Colocalization for S-PrediXcan Associations for selected phenotypes. Column 'P4' lists the number of gene/tissue pairs that fall in the 'colocalized' region ($P(H4) > 0.5$, blue in Figure 3-A, 'P3' corresponds to 'non colocalized' or 'independent signal' region ($P(H3) > 0.5$, orange in Figure 3-A), 'undetermined' corresponds to region without strong evidence of either colocalization or non colocalization (gray in Figure 3-A), and the 'missing' column lists gene/tissue pairs for which colocalization yielded NA.

phenotype	total	P4	%	P3	%	undetermined	%	missing	%
Alzheimer's Disease	124	7	5.6%	44	35.5%	65	52.4%	8	6.5%
Bipolar Disorder	13	12	92.3%	0	0.0%	1	7.7%	0	0.0%
Birth Length	7	6	85.7%	0	0.0%	0	0.0%	1	14.3%
Body Mass Index	508	281	55.3%	122	24.0%	79	15.6%	26	5.1%
Cigarettes per Day	23	4	17.4%	7	30.4%	10	43.5%	2	8.7%
Coronary Artery Disease	136	93	68.4%	14	10.3%	20	14.7%	9	6.6%
Crohn's Disease	607	314	51.7%	166	27.3%	84	13.8%	43	7.1%
Education Years	20	19	95.0%	0	0.0%	0	0.0%	1	5.0%
Fasting Glucose	542	91	16.8%	64	11.8%	350	64.6%	37	6.8%
Fasting Insulin adjusted for BMI	102	25	24.5%	8	7.8%	63	61.8%	6	5.9%
Fasting Proinsulin	187	10	5.3%	74	39.6%	91	48.7%	12	6.4%
HDL Cholesterol	821	264	32.2%	236	28.7%	251	30.6%	70	8.5%
Height	5840	1672	28.6%	2063	35.3%	1724	29.5%	381	6.5%
Hemogoblin Levels	69	13	18.8%	36	52.2%	17	24.6%	3	4.3%
LDL Cholesterol	825	219	26.5%	221	26.8%	342	41.5%	43	5.2%
Major Depressive Disorder	1	1	100.0%	0	0.0%	0	0.0%	0	0.0%
Myocardial Infarction	80	69	86.2%	1	1.2%	3	3.8%	7	8.7%
Rheumatoid Arthritis	1580	159	10.1%	1219	77.2%	103	6.5%	99	6.3%
Schizophrenia	1122	283	25.2%	515	45.9%	254	22.6%	70	6.2%
Triglicerids	709	161	22.7%	242	34.1%	255	36.0%	51	7.2%
Type 2 Diabetes	33	19	57.6%	4	12.1%	6	18.2%	4	12.1%
Ulcerative Colitis	565	74	13.1%	371	65.7%	96	17.0%	24	4.2%

Supplementary Table 3. S-PrediXcan Association yields results more significant than Top SNPs. Genes associated by S-PrediXcan to Coronary Artery Disease GWAS where S-PrediXcan outperforms individual SNPs in a 2 Mb window around the gene.

Gene Name	Tissue	P-value	Top SNP in Region	Top SNP P-value
FES	Cells Transformed fibroblasts	1.23E-08	rs2521501	5.0E-08
FHL3	Skin Sun Exposed Lower leg	1.99E-07	rs28470722	9.84E-07
IP6K2	Adipose Subcutaneous	2.14-07	rs7623687	5.22E-07
LIPA	Lung	1.84-12	rs1412444	5.15E-12
LIPA	Whole Blood	1.67-14	rs1412444	5.15E-12
NT5C2	Testis	3.79-09	rs11191416	4.65E-09
TCF21	Adrenal Gland	1.93E-11	rs12202017	1.98E-11
TCF21	Nerve Tibial	7.19E-12	rs12202017	1.98E-11
TUBG2	Adipose Visceral Omentum	2.34E-07	rs72823056	1.5E-06
IL6R	Colon Transverse	2.31E-10	rs6689306	2.6E-09
PCSK9	Nerve Tibial	1.04E-08	rs11206510	2.34E-08
SNF8	Thyroid	2.20E-07	rs35895680	3.76E-07
SWAP70	Spleen	1.00-08	rs10840293	1.28E-08
FURIN	Artery Aorta	1.27E-08	rs2521501	5.01E-08
UTP11L	Artery Tibial	1.58E-07	rs28470722	9.84E-07

Supplementary Table 4. List of Genome-wide Association Meta Analysis (GWAMA) Consortia and phenotypes. Data included in **supplementary_table_consortia.csv**. Columns are consortium name, study name, gene2pheno.org display name, link to pubmed entry if available, study sample size, study population, # of significant gene-tissue pairs, # of significant unique genes, # remaining after filtering prediction performance, # remaining after discarding high non-shared signals probability, # remaining after discarding undetermined results from last column.

Supplementary Table 5. List of Significant Gene Association Results. Data included in **supplementary_table_significant_genes.csv**. Columns are consortium name, gene2pheno.org display name, study name, gene name, tissue, zscore, p-value, probability of 'undetermined' colocalization, probability of non-shared signals, probability of colocalized signals.

Supplementary Table 6. S-PrediXcan 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. P-value 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`. See more details in Supplementary Table 9

Gene	Phenotype	Effect Size	Pvalue	Tissue	Pred.Perf.R2	Pred.Perf.Pvalue	P3	P4
SORT1	CAD	-0.09	1.3e-17	Liver	0.49	1.2e-15	0.00	1.00
		-0.14	3.6e-07	Pancreas	0.11	2.5e-05	0.07	0.88
		-0.25	9.3e-04	DGN WB	0.02	8.3e-05		
		-0.06	8.7e-03	Esophagus Mucosa	0.05	4.1e-04		
		0.03	5.6e-02	Small Intestine Terminal Ileum	0.17	1.7e-04		
		-0.05	1.5e-01	Spleen	0.09	3.9e-03		
		-0.02	2.4e-01	Testis	0.18	3.8e-08		
		0.08	5.4e-01	Brain Hippocampus	0.09	7.5e-03		
		-0.00	5.8e-01	Brain Anterior cingulate cortex BA24	0.17	2.9e-04		
		0.01	8.9e-01	Breast Mammary Tissue	0.03	2.4e-02		
SORT1	Myocardial Infarction	-0.08	5.2e-12	Liver	0.49	1.2e-15	0.00	1.00
		-0.12	4.4e-05	Pancreas	0.11	2.5e-05	0.07	0.88
		-0.21	1.2e-02	DGN WB	0.02	8.3e-05		
		-0.05	2.8e-02	Esophagus Mucosa	0.05	4.1e-04		
		0.03	1.2e-01	Small Intestine Terminal Ileum	0.17	1.7e-04		
		-0.02	2.4e-01	Testis	0.18	3.8e-08		
		-0.01	2.9e-01	Brain Anterior cingulate cortex BA24	0.17	2.9e-04		
		-0.04	3.3e-01	Spleen	0.09	3.9e-03		
		0.01	5.3e-01	Pituitary	0.06	1.8e-02		
		0.04	8.2e-01	Brain Hippocampus	0.09	7.5e-03		
SORT1	LDL-C	-0.14	7.4e-183	Liver	0.49	1.2e-15	0.00	1.00
		-0.24	6.5e-96	Pancreas	0.11	2.5e-05	0.05	0.90
		-0.11	2.9e-31	Esophagus Mucosa	0.05	4.1e-04	0.28	0.41
		-0.34	2.8e-27	DGN WB	0.02	8.3e-05		
		0.36	5.9e-11	Brain Hippocampus	0.09	7.5e-03	0.10	0.06
		-0.08	3.6e-06	Spleen	0.09	3.9e-03	0.11	0.06
		-0.03	5.5e-04	Testis	0.18	3.8e-08	1.00	0.00
		0.02	2.9e-02	Small Intestine Terminal Ileum	0.17	1.7e-04		
		-0.01	1.5e-01	Brain Anterior cingulate cortex BA24	0.17	2.9e-04		
		-0.01	2.0e-01	Pituitary	0.06	1.8e-02		

Supplementary Table 7. S-PrediXcan association between *C4A* and schizophrenia S-PrediXcan 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. P-value 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. P-values of 0.02 and 0.03 for the Brain Hippocampus and Cortex results should not be interpreted as not associated. Brain tissues have limited sample size which could be one of the reasons why this association is less significant than in other tissues. For example there is no significant eQTL for this gene in Brain Hippocampus and Cortex. By using a multi snp model we obtain significant models even when single eQTL analysis does not produce significant results. *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`. See more details in Supplementary Table 9

Gene	Phenotype	Effect Size	Pvalue	Tissue	Pred.Perf.R2	Pred.Perf.Pvalue	P3	P4
C4A	Schizophrenia	0.15	2.3e-20	Pancreas	0.27	1.7e-11	0.06	0.94
		0.16	7.7e-20	Artery Aorta	0.23	6.1e-13	0.44	0.56
		0.12	1.5e-19	Testis	0.35	4.6e-16	0.06	0.94
		0.13	2.6e-19	Thyroid	0.28	3.6e-21	0.46	0.54
		0.12	6.8e-19	Heart Atrial Appendage	0.39	8.6e-19	0.69	0.31
		0.15	8.5e-19	Adipose Subcutaneous	0.22	8.8e-18	0.10	0.90
		0.22	9.3e-19	Colon Sigmoid	0.16	2.9e-06	0.15	0.83
		0.15	1.0e-18	Heart Left Ventricle	0.26	5.6e-14	0.12	0.88
		0.13	1.2e-18	Liver	0.38	1.9e-11	0.33	0.67
		0.19	2.0e-18	Cells EBV-transformed lymphocytes	0.23	7.3e-08	0.07	0.92
		0.15	2.2e-18	Stomach	0.30	6.2e-15	0.16	0.84
		0.34	3.5e-18	Brain Hypothalamus	0.09	5.3e-03	0.27	0.39
		0.15	1.0e-17	Lung	0.20	8.3e-15	0.07	0.93
		0.16	2.7e-17	Colon Transverse	0.20	8.6e-10	0.07	0.93
		0.18	3.7e-17	Muscle Skeletal	0.18	1.3e-17	0.10	0.90
		0.11	4.7e-17	Nerve Tibial	0.33	1.2e-23	0.19	0.81
		0.18	9.1e-17	Adipose Visceral Omentum	0.22	1.8e-11	0.08	0.92
		0.13	3.6e-16	Brain Putamen basal ganglia	0.19	4.6e-05	0.17	0.62
		0.21	4.0e-16	Artery Coronary	0.10	5.2e-04	0.37	0.48
		0.15	1.2e-15	Brain Frontal Cortex BA9	0.18	3.1e-05	0.12	0.87
		0.16	1.7e-15	Esophagus Gastroesophageal Junction	0.22	2.0e-08	0.13	0.87
		0.12	4.1e-15	Prostate	0.21	6.2e-06	0.42	0.34
		0.12	6.2e-15	Esophagus Mucosa	0.26	1.5e-17	0.06	0.94
		0.13	1.5e-14	Breast Mammary Tissue	0.26	2.4e-13	0.26	0.74
		0.14	2.5e-14	Skin Sun Exposed Lower leg	0.24	2.3e-19	0.11	0.89
		0.14	1.2e-13	Brain Cerebellum	0.23	2.8e-07	0.10	0.88
		0.12	6.3e-13	Whole Blood	0.20	1.9e-18	0.05	0.95
		0.10	2.2e-12	Brain Cerebellar Hemisphere	0.17	7.3e-05	0.23	0.76
		0.08	4.3e-12	Skin Not Sun Exposed Suprapubic	0.35	7.1e-20	0.22	0.78
		0.10	1.0e-11	Cells Transformed fibroblasts	0.23	2.8e-17	0.12	0.88
		0.14	6.8e-11	Adrenal Gland	0.17	1.8e-06	0.19	0.78
		0.09	3.2e-10	Artery Tibial	0.17	5.2e-13	0.18	0.82
		0.08	5.7e-10	Brain Caudate basal ganglia	0.13	2.9e-04	0.30	0.34
		0.11	1.1e-09	Uterus	0.10	7.2e-03	0.34	0.07
		0.08	5.4e-09	Spleen	0.28	1.0e-07	0.07	0.93
		0.06	3.0e-08	Brain Anterior cingulate cortex BA24	0.25	9.3e-06	0.31	0.42
		0.12	1.1e-04	Small Intestine Terminal Ileum	0.08	1.1e-02	0.39	0.16
		0.05	2.6e-04	Pituitary	0.14	2.8e-04	0.30	0.32
		0.03	2.1e-02	Brain Hippocampus	0.12	1.2e-03	NA	NA
		0.03	3.4e-02	Brain Cortex	0.10	1.8e-03	NA	NA

Supplementary Table 8. S-PrediXcan 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 [35] 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. Even though some of these p-values are above 0.05, the corresponding FDR was less than 0.05, on account of small value of π_0 (estimated proportion of null associations). Supplementary figure 7 illustrates this point. *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`. See more details in Supplementary Table 9

Gene	Phenotype	Effect Size	Pvalue	Tissue	Pred.Perf.R2	Pred.Perf.Pvalue	P3	P4
PCSK9	CAD	0.13	1.0e-08	Nerve Tibial	0.18	1.5e-12	0.01	0.99
		0.49	4.1e-07	Lung	0.01	1.0e-01	0.01	0.98
		0.35	4.6e-05	Whole Blood	0.01	1.2e-01	0.09	0.84
		0.10	4.5e-03	Testis	0.04	9.1e-03		
		-0.17	1.2e-02	Colon Transverse	0.02	5.4e-02		
		0.07	2.0e-02	Adipose Visceral Omentum	0.06	1.1e-03		
		0.04	3.9e-02	Brain Cerebellum	0.07	6.1e-03		
		-0.06	2.3e-01	Skin Sun Exposed Lower leg	0.01	8.0e-02		
		0.07	3.5e-01	Artery Tibial	0.02	9.6e-03		
		-0.02	4.8e-01	Vagina	0.08	1.2e-02		
		-0.15	6.0e-01	Artery Coronary	0.04	3.8e-02		
PCSK9	Myocardial Infarction	0.12	8.6e-07	Nerve Tibial	0.18	1.5e-12	0.01	0.98
		0.48	7.9e-06	Lung	0.01	1.0e-01	0.01	0.97
		0.34	3.0e-04	Whole Blood	0.01	1.2e-01	0.07	0.75
		0.09	2.1e-02	Testis	0.04	9.1e-03		
		-0.10	8.5e-02	Skin Sun Exposed Lower leg	0.01	8.0e-02		
		0.05	8.8e-02	Adipose Visceral Omentum	0.06	1.1e-03		
		-0.11	1.5e-01	Colon Transverse	0.02	5.4e-02		
		0.03	2.5e-01	Brain Cerebellum	0.07	6.1e-03		
		-0.01	5.8e-01	Brain Cortex	0.04	3.9e-02		
		0.03	7.2e-01	Artery Tibial	0.02	9.6e-03		
		-0.01	8.5e-01	Vagina	0.08	1.2e-02		
PCSK9	LDL-C	0.13	1.4e-27	Nerve Tibial	0.18	1.5e-12	0.15	0.85
		-0.28	5.1e-21	Colon Transverse	0.02	5.4e-02	0.13	0.11
		0.43	2.2e-13	Lung	0.01	1.0e-01	0.02	0.91
		0.13	2.4e-13	Adipose Visceral Omentum	0.06	1.1e-03	0.69	0.05
		-0.16	4.3e-12	Skin Sun Exposed Lower leg	0.01	8.0e-02	0.15	0.10
		0.39	1.1e-10	Whole Blood	0.01	1.2e-01	0.07	0.60
		0.06	2.5e-10	Brain Cerebellum	0.07	6.1e-03	0.14	0.26
		0.03	3.6e-03	Brain Cortex	0.04	3.9e-02		
		-0.02	4.4e-01	Vagina	0.08	1.2e-02		
		0.05	4.7e-01	Testis	0.04	9.1e-03		
		-0.10	5.1e-01	Artery Coronary	0.04	3.8e-02		

Supplementary Table 9. MetaXcan results for *C4A*, *PCSK9*, *SORT1*. Full set of results available in `gene2pheno.org`. Data included in **supplementary_table_results_known_functional_genes.csv**

Supplementary Table 10. 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

Supplementary Note

There is a large amount of evidence indicating that a substantial portion of genetic effect on phenotype is mediated via alteration of gene expression levels. Studies of enrichment of expression quantitative trait loci (eQTLs) among trait-associated variants [1–3] show the importance of this relationship.

Given the success of GWAS approaches, large-scale GWAS/GWAMA efforts with ever increasing sample sizes are underway, which will be able to detect variants of smaller effects sizes. QTL studies, on the other hand, are currently limited to smaller sample sizes.

To take advantage of GWAS and QTL study data, PrediXcan was designed to test the mediating molecular trait's effect on a phenotype [11]. Its purpose is to identify trait-associated genes using genetically predicted molecular traits such as gene expression. PrediXcan uses independent QTL studies to train prediction models of the molecular trait. Using these models, PrediXcan imputes gene expression levels in the GWAS study cohort, and then correlates the association of the predicted gene expression levels to the phenotype of interest.

This means that PrediXcan has the following features [11]:

- **Reduced testing burden.** By performing gene-level tests, the computational burden is significantly reduced when compared to single variant tests. $20,000$ gene-level tests are need at most, whereas $10,000,000$ test may be needed in a single-variant analysis.

- **Direction of effect.** PrediXcan provides the direction of effect of the association, so that potential targets for either down-regulation or up-regulation can be identified.

- **Reduced reverse causality problems.** Drug treatment or disease status may affect a molecular trait without modifying germline genomic variation. Germline is not affected by disease and prediction models are built in independent cohorts such that when there is a causal relationship the direction goes mostly from the predicted gene expression to the disease.

- **Does not claim causality.** PrediXcan may yield significant association in cases where the causal variant for the gene expression and the causal variant for the phenotype are different but are in LD. This can be mitigated using colocalization measures to filter out these LD-contaminated associations. Also, PrediXcan cannot distinguish between causal relationship and pleiotropy. For example, if the same causal SNP affects the expression level of two genes, where one is causal and

the other one is not, PrediXcan will not be able to distinguish between them.

- **Detection of aggregates of small effects**. In cases when multiple causal variants affect gene expression levels, it may happen that the individual SNP association does not reach genome-wide significance but the combined multi-SNP effect does clear the threshold because of the combination of reduced multiple testing burden and increased effect size of the multi-SNP combination.

In summary, one of the main benefits of PrediXcan (and S-PrediXcan) over GWAS is that it returns genes rather than SNPs since much more is known about the function of genes. Another important benefit is the increased power because of the reduced multiple testing burden and the potentially larger effect sizes of the imputed genes.