Integrating tissue specific mechanisms into GWAS summary results

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

To understand the biological mechanisms underlying the thousands of genetic variants robustly associated with complex traits, scalable methods that integrate GWAS and functional data generated by large-scale efforts are needed. We derived a mathematical expression to compute PrediXcan results using summary data (S-PrediXcan) and showed its accuracy and robustness to misspecified reference populations. We compared S-PrediXcan with existing methods and combined them into a best practice framework (MetaXcan) that integrates GWAS with QTL studies and reduces LD-confounded associations. We applied this framework to 44 GTEx tissues and 101 phenotypes from GWAS and meta-analysis studies, creating a growing catalog of associations that captures the effects of gene expression variation on human phenotypes. Most of the associations were tissue specific, indicating context specificity of the trait etiology. Colocalized significant associations in unexpected tissues underscore the advantages of an agnostic scanning of multiple contexts to increase the probability of detecting causal regulatory mechanisms.

Prediction models, efficient software implementation, and association results are shared as a resource for the research community.

Introduction

Over the last decade, GWAS have been successful in identifying genetic loci that robustly associate with human complex traits. However, the mechanistic understanding of these discoveries is still limited, hampering the translation of 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) where transcriptome levels are imputed with models trained in measured transcriptome datasets (e.g. GTEx). These predicted expression levels are then correlated with the phenotype in a gene 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.

A method based on similar ideas to PrediXcan was proposed by Gusev et al. [15] called Transcriptome-wide Association Study (TWAS). For the individual level data based version, the main difference between PrediXcan and TWAS resides in the models used for the prediction of gene expression levels in each implementation. An important extension of this approach was implemented by Gusev et al. [15] that allows the computation of gene-level association results using only summary statistics. We will refer to this method as Summary-TWAS (or S-TWAS for short).

Zhu et al [16] proposed another method that integrates eQTL data with GWAS results based on summary data. The method, Summary Mendelian Randomization (SMR), uses Wald statistics (effect size/standard error) from GWAS and eQTL studies to estimate the effect of the genetic component of gene expression on a phenotype using the delta approximation [17]. By design, this approach uses one eQTL per gene so that in practice only the top eQTL is used per gene. SMR incorporates uncertainty in the eQTL association and a post-filtering step, HEIDI, that tests the heterogeneity of the GWAS and eQTL hits.

To examine whether eQTL and GWAS hits in close proximity share the same underlying causal signal, several methods have been developed such as RTC [1], Sherlock [18], COLOC [19], and more recently eCAVIAR [20] and ENLOC [21]. Thorough comparison between RTC, COLOC, and eCAVIAR can be found in [20]. HEIDI, part of SMR, is another approach that computes the degree of non-colocalization or heterogeneity of signals.

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 the applicability of PrediXcan. We compare with existing methods and outline a best practice 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 for 101 phenotypes from 37 large meta-analysis consortia.

A limitation of this approach is linkage disequilibrium (LD) confounding: when different causal SNPs are affecting expression levels and the phenotypic trait in a GWAS, PrediXcan may yield significant results if the SNPs are in LD. To reduce false positive links caused by this confounding, we filter out associations based on the colocalization status of the eQTL and GWAS signals. Using these results, we build a growing catalog of downstream phenotypic associations with molecular traits across multiple tissues and contexts, and make it publicly available at gene2pheno.org.

Results

Inferring PrediXcan results with summary statistics

We have derived an analytic expression that allows us to compute the outcome of PrediXcan using only summary statistics from genetic association studies. Details of the derivation are shown in the Methods section. In Figure 1-A, we illustrate the mechanics of Summary-PrediXcan (S-PrediXcan) in relation to traditional GWAS and the individual-level PrediXcan method [11].

For both GWAS and PrediXcan, the input is a genotype matrix and phenotype vector. GWAS computes the regression coefficient of the phenotype on each marker in the genotype matrix and generates SNP-level results. PrediXcan starts by estimating the genetically regulated component of the transcriptome (using weights from the publicly available PredictDB database) and then computes regression coefficients of the phenotype on each predicted gene expression level generating gene-level results. S-PrediXcan, on the other hand, can be viewed as a shortcut that uses the output from a GWAS to infer the output from PrediXcan, using the LD structure (covariances) from a reference population. Since S-PrediXcan only uses summary statistics, it can effectively take advantage of the considerably larger sample sizes available from GWAMA studies, avoiding the computational and regulatory burden of handling large amounts of protected individual-level data.

MetaXcan framework

Building on S-PrediXcan and existing approaches, we define a general framework (MetaXcan) to integrate QTL information with GWAS results to map disease-associated genes as illustrated on Figure 2. This evolving framework will incorporate state of the art 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 as outlined in Figure 2-A.

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-confounded associations, which can occur when expression predictor SNPs and phenotype causal SNPs are different but in LD, we use state of the art methods that estimate the probability of shared or independent signals.

PrediXcan implementations work mostly with elastic net models motivated by our observation that gene expression variation is mostly driven by sparse components [22]. TWAS implementations have used Bayesian Sparse Linear Mixed Models [23] (BSLMM), which allows both polygenic and sparse components. 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).

SMR has implemented an adjustment for model uncertainty by using half of the harmonic average of GWAS and eQTL χ^2-statistics. It is in principle possible to extend this idea to S-PrediXcan, but this would bound the significance of the association to the smaller of the prediction model or GWAS significance, which is an overly stringent penalization of the uncertainty in the prediction model (see the comparison subsection for details).

For the last step, we chose COLOC to estimate the probability of colocalization of GWAS and eQTL signals. SMR uses its own estimate of "heterogeneity" of signals calculated by HEIDI. We chose to use COLOC probabilities because COLOC clusters more distinctly into different classes and, unlike other methods, does not require an arbitrary cut off threshold. 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.

Gene expression variation in humans is associated to diverse phenotypes

Next, we downloaded summary statistics from meta analyses of 101 phenotypes from 37 consortia. The full list of consortia and phenotypes is shown in Supplementary Table 3. We tested association between these phenotypes and the predicted expression levels using elastic net models in 44 human tissues from GTEx as described in the 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 ≈ 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 genes significantly associated with 1,690 unique genes (based on a GWAMA of 250K individuals). Other polygenic diseases with a large number of associations include schizophrenia with 307 unique significant genes (n = 150K individuals), low-density lipoprotein cholesterol (LDL-C) levels with 297 unique significant genes (n = 188K), 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 none for major depressive disorder (n = 18,759), probably due to smaller sample sizes, but also smaller effect sizes.

When excluding genes with evidence of independent GWAS-eQTL signals (P3>0.5), these numbers dropped by about 10-20% to 1377 for height, 231 for schizophrenia, and 244 for LDL-C levels. If we further exclude genes with low power to determine either shared or non-shared GWAS-eQTL signals, we find 642 genes for height, 157 for schizophrenia, and 78 for LDL-C. The quantities for the full set of phenotypes can be found in Supplementary Table 3.

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 from taking into account the combined effects of multiple variants. Supplementary Table 2 lists a few examples where this occurs.

As expected, results of S-PrediXcan tend to be more significant as the genetic component of gene expression increases (larger cross-validated prediction performance R^2). Similarly, S-PrediXcan associations tend to be more significant when prediction performance p-values are more significant. The trend is seen both when results are averaged across all tissues for a given phenotype or across all phenotypes for a given tissue. All tissues and representative phenotypes are shown in Supplementary Figures 2-5. This trend was also robust across different monotone functions of the Z-scores.

The full set of results can be queried in our online catalog gene2pheno.org, and we provide the significant association results in Supplementary Table 4. Our web application allows filtering the results by gene, phenotype, tissue, p-value, prediction performance, and colocalization status. For each trait we assigned ontology terms from the Experimental Factor Ontology (EFO) [24] and Human Phenotype Ontology (HPO) [25], if applicable. As the catalog grows, the ontology annotation will facilitate analysis by hierarchy of phenotypes. Supplementary Table 3 shows the list of consortia and phenotypes for which gene-level associations are available.

To facilitate comparison, the catalog contains all SMR results we generated and the S-TWAS results reported by [26] for 30 GWAS traits and GTEx BSLMM models. SMR application to 28 phenotypes was reported by [27] using whole blood eQTL results from [28].

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 [29] for obesity, rheumatoid arthritis, diabetes, Alzheimer’s, Crohn’s disease, ulcerative colitis, age-related macular degeneration, schizophrenia, and autism. 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 significant genes were found 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 [30]. 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

The broad coverage of tissues in our prediction models enabled us to examine the tissue specificity of phenotypic associations of GWAS signals. We started by computing average enrichment of significance by tissue. We used several measures of enrichment such as the mean Z-scores squared across all genes, or across significant genes for different thresholds, as well as the proportion of significant genes for different thresholds. We also compared the full distribution of the p-values of a given tissue relative to the remaining tissues. Supplementary Figure 6 shows the average Z-score^2 as a measure of enrichment of each tissue by phenotype.

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

Next we focused on three genes whose functional role has been well established: *C4A* for schizophrenia [31] and *SORT1* [32] and *PCSK9* both for LDL-C and cardiovascular disease. The 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 3 with additional details in Supplementary Tables 5, 6, and 7.

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 [32]. 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 [31]. 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).

Note that p-values of 0.02 and 0.03 for the Brain Hippocampus and Cortex results should not be interpreted as not being associated. Brain tissues have limited sample size which could be one of the reasons why this association is less significant than in other tissues. There is no significant eQTL for this gene in Brain Hippocampus and Cortex so that SMR runs, performed using significant eQTL dataset from GTEx as recommended, did not return any result. By using a multi snp model we obtain significant models even when single eQTL analysis does not produce significant results.

PCSK9 is a target of several LDL-C lowering drugs currently under trial to reduce cardiovascular events [33]. The STARNET study [34] 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 ($p \approx 10^{-27}$). The association between *PCSK9* and coronary artery disease is also significant in tibial nerve ($p \approx 10^{-8}$) but only nominally significant in visceral fat ($p \approx 0.02$). Accordingly, in our training set (GTEx), there is much stronger evidence of regulation of this gene in tibial nerve compared to visceral fat. Moreover, visceral fat association shows evidence of independent rather than shared GWAS and eQTL signals in the PCSK9 locus (probability of independent signals P3=0.69 in LDL-C). It is likely 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 tissue results we increase the window of opportunities where we can detect the association.

These examples demonstrate the power of studying the regulation in a broad set of tissues and contexts and emphasize the challenges of determining causal tissues of complex traits based on in-silico analysis alone. Based on these results, we would recommend to scan all tissue models to increase the chances to detect the relevant regulatory mechanism that mediates the phenotypic association. False positives will be controlled by accounting for the multiple testing with a more stringent significance cutoff.

Replication in an independent cohort

We used data from the Resource for Genetic Epidemiology Research on Adult Health and Aging study (GERA, phs000674.v1.p1) [35,36]. 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 dataset in the European subset of the cohort. Genotypes were imputed using the University of Michigan server and principal components provided by the GERA study were used to adjust for population stratification. More details can be found in the Methods section.

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 significant associations (FDR< 0.05) in the discovery set. Analysis and replication of the type 2 diabetes phenotype can be found in [37]. 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.

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 [35]. The significance threshold for replication was $p< 0.05$, 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.

First, we estimated the proportion of true associations in the replication set (these include LD-induced ones) using the π_1 statistics from the q-value approach [38]. This approach does not indicate which genes are true positives but provides an estimate of the proportion. If we take all genes in the replication set, the estimated proportions of true associations are 0.4% for "Any cardiac event", 5.8% for "Dyslipidemia", and 1.2% for schizophrenia (see third column in Table 1). When we compute π_1 for the subset of genes that were found to be Bonferroni significant in the discovery analysis we find that π_1 goes up ten to one hundred fold as shown in Table 1. Following standard practice in meta-analysis, we consider a gene to be replicated if 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 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.

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

Comparison of S-PrediXcan to other integrative methods based on summary results

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 traitassociated 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. This 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. Without this assumption, the mean of the distribution is off by a factor of 4. See Methods section for further details.

When the eQTL association is much stronger than the GWAS association, we show that the SMR statistic is approximately equal to the GWAS χ^2-statistics of the top eQTL for the gene, which is equal to the Summary-PrediXcan Z-score^2 if top eQTL is used as predictor. See derivation in Methods section.

On the other extreme, when the GWAS association is much stronger, the SMR statistic is approximately equal the top eQTL χ^2-statistic (slightly smaller). In general, the SMR statistic is bounded by the eQTL and GWAS significance in practically all cases as shown in Figure 4-D and E.

Given the cost differences, the current trend of much larger GWAS studies compared to eQTL studies will continue. This means that the SMR significance will be bounded by the significance of the eQTL association, which seems too conservative.

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 a natural extension of ImpG [39] or DIST [40], which are SNP-based methods that impute summary statistics of unmeasured SNPs using Gaussian imputation [41]. If restricted to Gaussian imputation, we show that this approach is equivalent to predicting expression levels using BLUP/Ridge Regression, which has been shown to be suboptimal for gene expression traits [22]. However, the mathematical expression used by S-TWAS can be extended to any set of weights such as Bayesian Sparse Linear Mixed Models (BSLMM) as used by Gusev et al. [15]. S-TWAS imputes the Z-score of 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 result of PrediXcan using summary statistics. In the Methods Section we establish the approximate equivalence of the two approaches when the same prediction weights are applied. Figure 4-A illustrates the components of SMR, S-TWAS, and S-PrediXcan methods. All three seek to identify target genes by computing the strength of association between the unobserved predicted expression levels (T_g) of a gene with the complex trait (Y) quantified with $Z_{T_g,Y}$or its square. SMR also incorporates uncertainty of the predicted expression in the statistics and adds a test for (non-) colocalization of GWAS and eQTL hits (HEIDI).

Next we show the comparison of S-PrediXcan associations to SMR, S-TWAS in practice. We computed SMR and COLOC results using the software provided by the authors and the GTEx eQTL data [9,10]. For S-TWAS we use the results made available by [26], which only include significant associations. We show results for the height phenotype and all GTEx tissues. Other phenotypes exhibit qualitatively similar patterns.

SMR, S-TWAS, and S-PrediXcan are directly comparable since all three provide the significance of the association between the mediating gene and the phenotype. Figure (4-B and -C) compare the significance of S-PrediXcan (elastic net) associations with S-TWAS and SMR results. As expected, SMR p-values tend to be less significant than Summary-PrediXcan's in large part due to the additional adjustment for the uncertainty in the eQTL association. S-TWAS (≈S-PrediXcan BSLMM) results are similar to S-PrediXcan with elastic net models (there is a small bias favoring the results of S-TWAS because only significant results were available). Overall all three methods rank results similarly with some differences that in part are a consequence of the effect size distributions of the eQTL and GWAS variants in each locus.

Colocalization estimates complement PrediXcan results

Here we compare to another class of methods that attempts to determine whether eQTL and GWAS signals are colocalized or are distinct although linked by LD. Among this class of methods are COLOC [19], Sherlock [18], and RTC [1], and more recently eCAVIAR [20], and ENLOC [21]. Thorough comparison between these methods can be found in [19,20]. HEIDI, the post filtering step in SMR that estimates heterogeneity of GWAS and eQTL signals, is another method in this class. We focus here on COLOC, whose quantification of the probability of five configurations complements well with the 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.

Figure 5 shows ternary plots [42] with P3, P4, and 1-P3-P4 as vertices (for convenience we aggregate H0, H1, and H2 into one event with probability 1-P3-P4). This representation restricts the sum to be 1. The top vertex corresponds to probability of colocalized eQTL and GWAS signals (P4) to be high. The bottom left vertex corresponds to distinct eQTL and GWAS signals (P3 high). The bottom right vertex corresponds to low probability of both colocalization and independent signals, which the authors [19] recommend to interpret as limited power.

Figure 5-B shows association results for all gene/tissue pairs to the height phenotype. We find that most gene-tissue pairs' association falls in the bottom right, “undetermined” region. When we restrict the plot to S-PrediXcan signficant genes (p-value<1E-6) (Figure 5-C), three distinct peaks emerge in the high P4 region ("colocalized signals"), high P3 region ("independent signals"), and "undetermined" region. Moreover, when genes with low prediction performance are excluded (Supplementary Figure 7-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.

This post-filtering idea was first implemented in the SMR approach using HEIDI. Comparison of COLOC results with HEIDI is shown in Figure 5-D and E. Panel D shows the colocalization probabilities of genes with small HEIDI p-values, which indicates heterogeneity of GWAS and eQTL signals. As expected, most genes fall in the lower left region, "independent signals" although there is a small cluster of genes that fall in the colocalized region, showing the disagreement between the two methods. When HEIDI p-values are large, i.e. the majority of genes cluster in the "colocalized" region, but there is a substantial number of genes that fall on the opposite end. COLOC and HEIDI tend to agree but in a number of cases they provide opposite conclusions. HEIDI does not provide a natural cutoff point for classification as COLOC does.

Discussion

Here we derive a mathematical expression to compute the results of PrediXcan (an integrative method that combines eQTL and GWAS data to map genes associated with complex traits) using only summary results, avoiding the need to use individual-level data. We show that our approach is accurate and robust to population mismatches. This allows us to greatly expand the applicability of PrediXcan given the widespread availability of summary results for massive sample size GWAS.

It also allows us to infer the downstream phenotypic association of any molecular trait as long as it can be approximately represented as linear functions of SNPs. These traits include expression levels of genes, intron usage, methylation status, telomere length, within different spatial, temporal, and developmental contexts.

Building on this derivation and existing methods to integrate GWAS and QTL data, we outline a general framework in which our method and others can be placed. We term this MetaXcan and view it as an evolving framework that computes the downstream phenotypic associations of genetic regulation of molecular (intermediate) traits. So far it is built on transcriptome prediction models based on elastic net, the calculator itself (implementation of the formula), adjustment for model uncertainty (hard threshold on minimum prediction performance), and an LD-confounding filter (colocalization of GWAS and eQTL status). SMR and S-TWAS can be considered different implementations of this framework. SMR uses top eQTL as predictor whereas Summary-TWAS has been implemented with BSLMM for prediction. SMR incorporates uncertainty of the prediction model into the association Z-score but the distribution of the combined statistics should be computed numerically instead of using the χ^2-square approximation, which will be valid only in extreme cases where the eQTL significance is much larger than the GWAS or vice-versa.

Methods to estimate colocalization is an active area of research. For example, COLOC assumes that there is a single causal variant for each gene. As they evolve, we will include the improved assessments to the MetaXcan framework.

We applied the MetaXcan framework by training transcriptome models in 44 human tissues from GTEx and estimating their effect on phenotypes from over 101 available GWAMA studies. We find known disease and trait associated genes active in relevant tissues but we also discover patterns of regulatory activity in tissues that are not traditionally associated with the trait. Further investigation of context and tissue specificity of these processes is needed but our results emphasize the importance of methods that integrate functional data across a broad set of tissues and contexts to augment our ability to identify novel target genes and provide mechanistic insight.

We also replicate some of our phenotypes using an independent cohort from the GERA study. Using the most related phenotypes available to us in GERA, we found that the proportion of true associations (estimated using the replication results) for the set of genes (BF significant in discovery) was between 48% and 91%. For LDL cholesterol, we find that 79% of discovery genes replicate in GERA.

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

Software and Resources

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

Methods

Summary-PrediXcan formula

Figure 6 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 1A.

The approximate formula we use is:

$$Z_g \approx \sum_{l \in Model_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} \tag{1}$$

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,
- se($\hat{\beta}_l$) is standard error of $\hat{\beta}_l$,
- $\hat{\sigma}_l$ is the estimated variance of SNP l,
- $\hat{\sigma}_g$ is the estimated variance of the predicted expression of gene g, and

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

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

Performance in simulated data

We first compared PrediXcan and Summary-PrediXcan using simulated phenotypes and a single transcriptome model trained on Depression Genes and Network’s (DGN) Whole Blood data set [5,22] downloaded from PredictDB (http://predictdb.org). The phenotype was sampled from a normal distribution without any link to genotype. 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 1B. 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 100%, whereas for sets of different ancestral origin the R^2 drops a few percentage points, with the biggest loss (down to 85%) when the study set is African and the reference set is Asian. This confirms that our formula works as expected and that the approach is robust to ethnic differences between study and reference sets.

Performance in cellular growth phenotype from 1000 genomes cell lines

Next we tested with an actual cellular phenotype - intrinsic growth. This phenotype was computed based on multiple growth assays for over 500 cell lines from the 1000 Genomes project [43]. 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 Supplementary Figure 1B. Consistent 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.

Performance on 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 Supplementary Figure 1C. 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 [23], 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.

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 $\widehat{\mathrm{Var}}$ and $\widehat{\mathrm{Cov}}$ the operators that compute the sample variance and covariance, i.e.: $\widehat{\mathrm{Var}}(Y) = \hat{\sigma}_Y^2 = \sum_{i=1,n}(Y_i - \bar{Y})^2/(n-1)$ with $\bar{Y} = \sum_{i=1,n} Y_i/n$. Let $\hat{\sigma}_l^2 = \widehat{\mathrm{Var}}(X_l)$, $\hat{\sigma}_g^2 = \hat{\sigma}_l^2 = \widehat{\mathrm{Var}}(T_g)$ and $\Gamma_g = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})/n$, where $\mathbf{X}'$ is the p × n matrix of SNP data and $\bar{\mathbf{X}}$ is a n × 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 $p \ll n$).

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

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

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

and

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

The proportion of variance explained by the covariate (T_g or X_l) can be expressed as:

$$R_g^2 = \hat{\gamma}_g^2 \frac{\hat{\sigma}_g^2}{\hat{\sigma}_Y^2}$$
$$R_l^2 = \hat{\gamma}_l^2 \frac{\hat{\sigma}_l^2}{\hat{\sigma}_Y^2}$$

By definition

$$T_g = \sum_{l \in Model_g} w_{lg} X_l$$

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

$$\hat{\sigma}_g^2 = \widehat{\mathrm{Var}}\left(\sum_{l \in Model_g} w_{lg}\, X_l \right)$$
$$= \widehat{\mathrm{Var}}(\mathbf{W}_g \mathbf{X}_g)$$
$$= \mathbf{W}_g' \;\; \widehat{\mathrm{Var}}(\mathbf{X}_g)\, \mathbf{W}_g$$

, where $\mathbf{W}_g$ is the vector of w_{lg} for SNPs in the model of g. By definition, Γ_g is $\widehat{\mathrm{Var}}(\mathbf{X}_g)$, the sample covariance of $\mathbf{X}_g$, so that we arrive to:

$$\boldsymbol{\hat{\sigma}_g^2} = \mathbf{W}_g' \boldsymbol{\Gamma}\, \mathbf{W_g} \tag{4}$$

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

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

$$\hat{\gamma}_g = \frac{\widehat{\mathrm{Cov}}(T_g, Y)}{\hat{\sigma}_g^2}$$
$$= \frac{\widehat{\mathrm{Cov}}\left(\sum_{l \in Model_g} w_{lg} X_l\, , Y\right)}{\hat{\sigma}_g^2}$$
$$= \sum_{l \in Model_g} \frac{w_{lg} \widehat{\mathrm{Cov}}(X_l, Y)}{\hat{\sigma}_g^2}$$

, where we used the linearity of $\widehat{Cov}$ in the last step. Using equation (3), we arrive to:

$$\hat{\gamma}_g = \sum_{l \in Model_g} \frac{w_{lg}\hat{\beta}_l\sigma_l^2}{\hat{\sigma}_g^2} \qquad (5)$$

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

Also from the properties of linear regression we know that

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

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

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

Thus:

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

Notice that the right hand side of (8) is dependent on the SNP l while the left hand side is not. This equality will hold only approximately in our implementation since we will be using approximate values for $\sigma\hat{}_l{}^2$, 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 se(γ_g), or Z-score,

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

with which we can compute the p-value as $p = 2\Phi(-|Z_g|)$ where $\Phi(.)$ is the normal CDF function. Thus:

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

$$= \sum_{l \in Model_g} \frac{w_{lg}\hat{\beta}_l \hat{\sigma}_l^2}{\hat{\sigma}_g^2} \sqrt{\frac{n}{\hat{\sigma}_Y^2} \frac{\hat{\sigma}_g^2}{(1-R_g^2)}}$$

$$= \sum_{l \in Model_g} \frac{w_{lg}\hat{\beta}_l \hat{\sigma}_l^2}{\hat{\sigma}_g^2} \sqrt{\frac{(1-R_l^2)}{\mathrm{se}^2(\hat{\beta}_l)\hat{\sigma}_l^2} \frac{\hat{\sigma}_g^2}{(1-R_g^2)}}$$

, where we used equations (5) and (6) in the second line and equation (8) in the last step. So:

$$Z_g = \sum_{l \in Model_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} \sqrt{\frac{(1-R_l^2)}{(1-R_g^2)}} \tag{10}$$

$$\approx \sum_{l \in Model_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)} \tag{11}$$

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

Expression model training

To train our prediction models, we obtained genotype data and normalized gene expression data collected by the GTEx Project. We used 44 different tissues sampled by GTEx and thus generated 44 different tissue-wide models (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 which reached significance at a False Discovery Rate of less than 0.05. For each tissue examined, we created a sqlite database to store the weights of the prediction models, as well as other statistics regarding model training. These databases have been made available for download at PredictDB.org.

Online Catalog and SMR, COLOC, TWAS

We have executed all methods and programs in the High-Performance Cluster of the Center for Research Informatics.

Supplementary Table 3 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 − 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.

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 [44] as an approximation. We also used the 1000 Genomes genotype data as reference panel for SMR.

Next we ran COLOC [19] 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 the option that estimates the phenotype variance from the variances and frequencies in each association study. When the frequency information was missing from the GWAS, we filled in with data from the 1000 Genomes Project.

For both the cases of SMR and COLOC, we discarded those SNPS where the allele sets in the GWAMA and the eQTL studies differed. After obtaining these results, we uploaded the results to the relational databases and linked to the appropriate S-PrediXcan result.

For comparison purposes, we have also included the results of the application of Summary-TWAS to 30 traits [26]. We linked each TWAS result to a matching S-PrediXcan result with the same GWAS Study, gene and transcriptome data source (i.e. GTEx tissue study).

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:

$$\begin{aligned}\widetilde{\mathbf{W}}_g &= \left(\sigma_1 w_{1g}, \dots, \sigma_p w_{pg}\right)' \\ \mathbf{Z}_{SNPs} &= \left(Z_1, \dots, Z_p\right)' \\ &= \left(\frac{\hat\beta_1}{se(\hat\beta_1)}, \dots, \frac{\hat\beta_p}{se(\hat\beta_p)}\right)'\end{aligned}$$

and correlation matrix of SNPs in the model for gene g

$$\Sigma_g = \text{diag}\left(\frac{1}{\hat\sigma_1}, \dots, \frac{1}{\hat\sigma_p}\right) \cdot \Gamma_g \cdot \text{diag}\left(\frac{1}{\hat\sigma_1}, \dots, \frac{1}{\hat\sigma_p}\right)$$

it is quite straightforward to write the numerator in (1) and (11) as

$$\widetilde{\mathbf{W}}_g \cdot \mathbf{Z}_g$$

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

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

, thus

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

This equation has the same form as the TWAS expression if we use the scaled weight vector $\widetilde{\mathbf{W}}_g$ instead of $\mathbf{W}_g$. 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

$$
\begin{aligned}
Z_{S-PrediXcan} &= \sum_{l \in Model_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\mathrm{se}(\beta_l)} \\
&= w_{1g} \frac{\hat{\sigma}_1}{\sqrt{w_{1g}^2 \hat{\sigma}_1^2}} Z_1 \\
&= Z_1
\end{aligned}
$$

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

$$Z^2_{\text{top eqtl S−PrediXcan}} = Z^2_{\text{GWAS}} \qquad (12)$$

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:

$$T_{SMR} = \frac{Z^2_{\text{eQTL}} Z^2_{\text{GWAS}}}{Z^2_{\text{eQTL}} + Z^2_{\text{GWAS}}} \qquad (13)$$

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. Only in two extreme cases, the chi-square approximation holds: when $Z_{\text{eQTL}} \gg Z_{\text{GWAS}}$ or $Z_{\text{eQTL}} \ll Z_{\text{GWAS}}$. In these extremes, we can apply Taylor expansions to find an interpretable form of the SMR statistic.

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

$$T_{SMR} = \frac{Z^2_{\text{GWAS}}}{1 + \frac{Z^2_{\text{GWAS}}}{Z^2_{\text{eQTL}}}} \approx Z^2_{\text{GWAS}} \left(1 - \frac{Z^2_{\text{GWAS}}}{Z^2_{\text{eQTL}}}\right) \qquad (14)$$

, so that for large enough Z^2_{eQTL} relative to Z^2_{GWAS},

$$T_{SMR} \approx Z^2_{\text{GWAS}} = Z^2_{\text{top eQTL S-PrediXcan}} \tag{15}$$

using equation 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 eQTLs, $Z_{\text{eQTL}} \ll Z_{\text{GWAS}}$,

$$T_{SMR} = \frac{Z^2_{\text{eQTL}}}{1 + \frac{Z^2_{\text{eQTL}}}{Z^2_{\text{GWAS}}}} \approx Z^2_{\text{eQTL}} \left(1 - \frac{Z^2_{\text{eQTL}}}{Z^2_{\text{GWAS}}}\right) \tag{16}$$

, so that analogously:

$$T_{SMR} \approx Z^2_{\text{eQTL}} \tag{17}$$

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 4-D/E. Given the rate of growth of sample sizes of GWAS studies compared to eQTL studies, the power of eQTL studies will cap the significance attainable by SMR. This approach seems unnecessarily conservative. In our framework, we use a minimum prediction performance threshold and estimates of colocalization to filter out unreliable associations.

GERA imputation

Genotype files were obtained from dbGaP, and updated to release 35 of the probe annotations published by Affymetrix via PLINK [45]. 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 [46]) and imputation against the HRC r1.1 2016 panel [47] (via minimac3) were carried out by the Michigan Imputation Server [48].

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 [45] 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 [49], 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 genebased 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. Largescale 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. Casella G, Berger R. Statistical Inference. 2nd ed. Imprint Australia; Pacific Grove, CA : Thomson Learning, c2002.; 2002.
18. 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.

19. 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.

20. 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.

21. 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.

22. 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).

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

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

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

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

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

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

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

42. 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.

43. 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).

44. 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.

45. 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.

46. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Referencebased 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.

47. 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.

48. 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.

49. 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.

Figure Captions

Figure 1. Comparison between GWAS, PrediXcan, and Summary-PrediXcan.

A) illustrates the Summary-PrediXcan method in relationship to GWAS and PrediXcan. Both GWAS and PrediXcan take genotype and phenotype data as input. GWAS computes the regression coefficients of Y on X_l using the model $Y = a + X_l b + \epsilon$, where Y is the phenotype and X_l the individual SNP dosage. The output is the table of SNP-level results. PrediXcan, in contrast, starts first by predicting/imputing the transcriptome. Then it calculates the regression coefficients of the phenotype Y on each gene's predicted expression T_g. The output is a table of gene-level results. Summary-PrediXcan directly computes the gene-level association results using the output from GWAS.

Comparison of results for **B)** a simulated phenotype; and **C)** a Bipolar Disorder study and a Type 1 Diabetes study from Wellcome Trust Case Control Consortium (WTCCC). For the simulated phenotype, study sets and S-PrediXcan population reference sets consisted of European, African, and Asian individuals from the 1000 Genomes Project. For the WTCCC phenotypes, the study set consisted of British individuals, and the S-PrediXcan population reference was the European subset of 1000 Genomes Project. Gene expression prediction models were based on the DGN cohort presented in [11].

Figure 2. MetaXcan Framework description and application.

Panel **A)** shows the components of the MetaXcan framework for integrating GWAS and eQTL data.

Panel **B)** displays our application of the MetaXcan framework. Using 44 RNA-seq data from GTEx we trained prediction models using elastic-net and deposited the weights and SNP covariance in the publicly available (PredictDB) resource. The weights and covariances were entered in the Summary-PrediXcan calculator, which when combined with 101 GWAS summary results, computed the gene/tissue pairs' associations. Colocalization status was computed and the full set of results were deposited in gene2pheno.org

Figure 3. A) 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. Blue circles correspond to the qq-plot of genes in ClinVar that were annotated with the phenotype and black circles correspond to all genes. **B) S-PrediXcan association for PCSK9, SORT1, and C4A.** $\mathrm{R}^2_{\mathrm{pred}}$ 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. The size of the points represent the significance of the association between predicted expression and the traits indicated on the top labels. C4A associations with schizophrenia (SCZ) are found across all 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.

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 4. Comparison Summary-PrediXcan with Summary-TWAS, and SMR.

The height phenotype association results across 44 GTEx tissues are analyzed in this figure. **Panel A)** depicts the test of the mediating role of gene expression level T_g in PrediXcan/TWAS summary versions and SMR.

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). There is a small bias caused by using S-TWAS results available from [26], which only lists significant hits.

Panel C) shows the significance of SMR vs Summary-PrediXcan. As expected, SMR associations tend to be smaller than S-PrediXcan's and S-TWAS'.

Panels D) and **E)** 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 **D**), and the gene's top eQTL p-value (panel **E**).

Some of the GWAS and eQTL p-values were more significant than shown since they were thresholded at $1\text{E-}50$ to improve visualization.

Figure 5. Colocalization status of S-PrediXcan results.

Panel **A)** shows a triangle that contains the probabilities of all five COLOC configurations. This ternary plot 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 corresponds to high probability of colocalization (P4>0.5), lower left sub-triangle 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 scatter plots of COLOC probabilities with a density overlay for S-PrediXcan results of the Height phenotype.

Panel **B)** shows the scatter plot of colocalization probabilities for all gene-tissue pairs. Most results fall into the "undetermined" region.

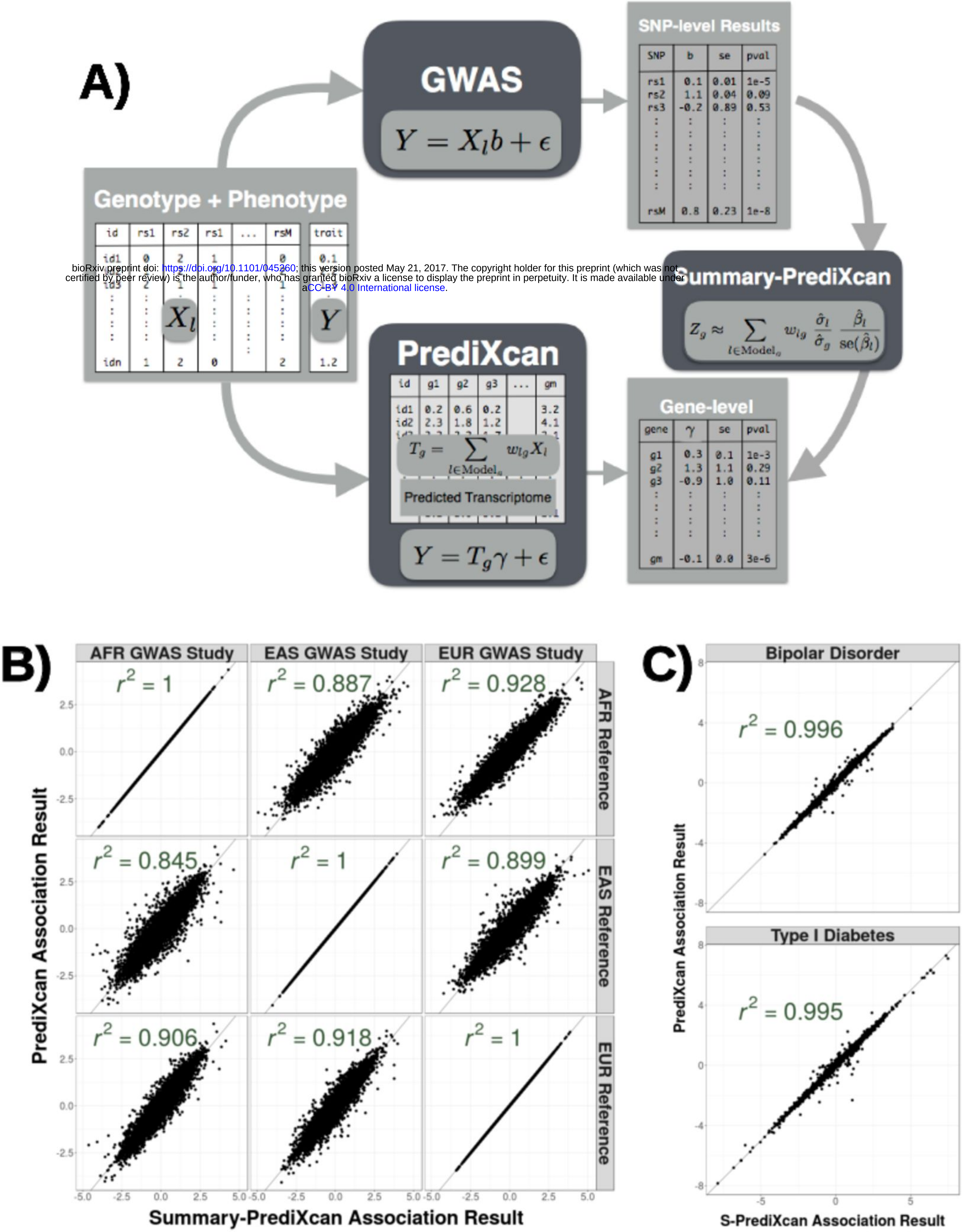
Panel **C)** shows that if we keep only significant results ($p_{\text{s-predixcan}} < 1 \times 10^{-6}$), associations tend to cluster into three distinct regions: "independent signals", "colocalized" and "undertermined", with most results in the "undetermined" region.

Panel **D)** shows that HEIDI significant genes (to be interpreted as high heterogeneity between GWAS and eQTL signals) mostly 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 but also substantial number fall in "independent signal" region, contradicting COLOC's classification.

Figure 6. Components of the S-PrediXcan formula. This plot shows the formula to infer

PrediXcan gene-level association results using summary statistics. The different sets involved in input data are shown. The regression coefficient between the phenotype and the genotype is obtained from the study set. The training set is the reference transcriptome dataset where the prediction models of gene expression levels are trained. The reference set, the training set (preferable) or 1000 Genomes, is used to compute the variances and covariances (LD structure) of the markers used in the predicted expression levels. Both the reference set and training set values are pre-computed and provided to the user so that only the study set results need to be provided to the software. The crossed out term was set to 1 as an approximation, since its calculation depends on generally unavailable data. We found this approximation to have negligible impact on the results.



Genotype + Phenotype

id	rs1	rs2	rs1	...	rsM
id1	0	2	1		0
id3	2	1	1		1
⋮	⋮	⋮	⋮	⋮	⋮
idn	1	2	0		2

trait
0.1
⋮
1.2

SNP-level Results

SNP	b	se	pval
rs1	0.1	0.01	1e-5
rs2	1.1	0.04	0.09
rs3	-0.2	0.89	0.53
⋮	⋮	⋮	⋮
rsM	0.8	0.23	1e-8

PrediXcan

id	g1	g2	g3	...	gm
id1	0.2	0.6	0.2		3.2
id2	2.3	1.8	1.2		4.1

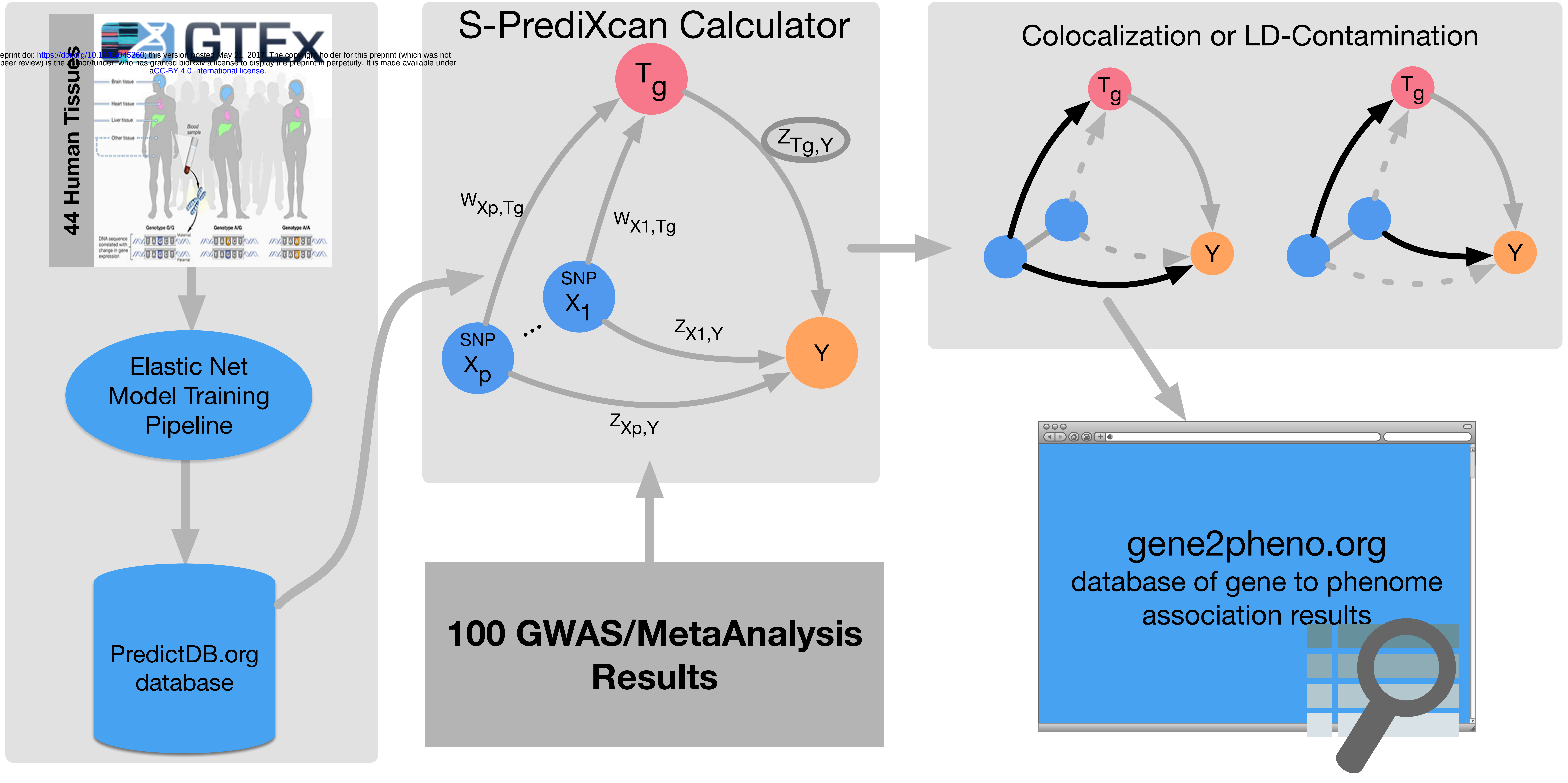
Gene-level

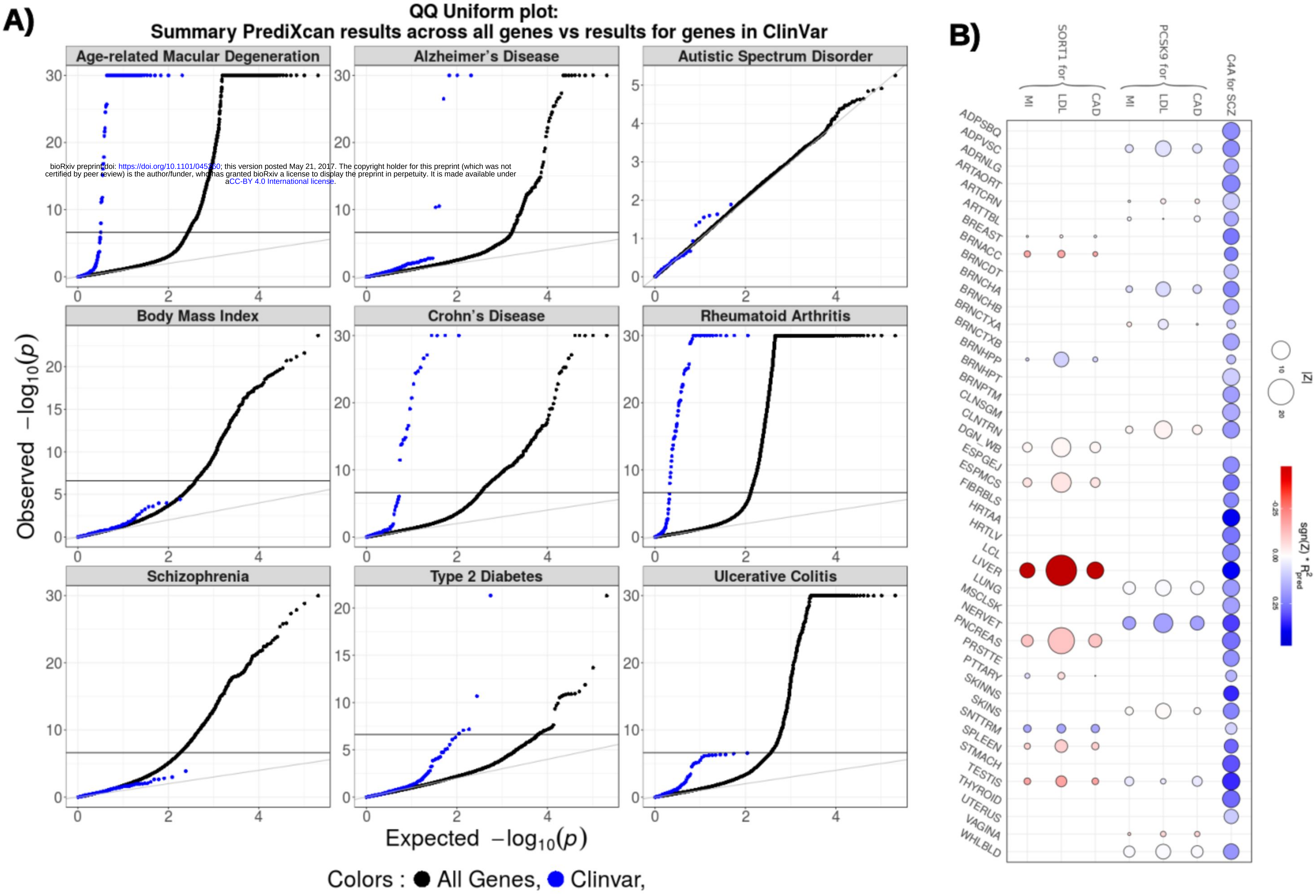
gene	γ	se	pval
g1	0.3	0.1	1e-3
g2	1.3	1.1	0.29
g3	-0.9	1.0	0.11
⋮	⋮	⋮	⋮
gm	-0.1	0.0	3e-6

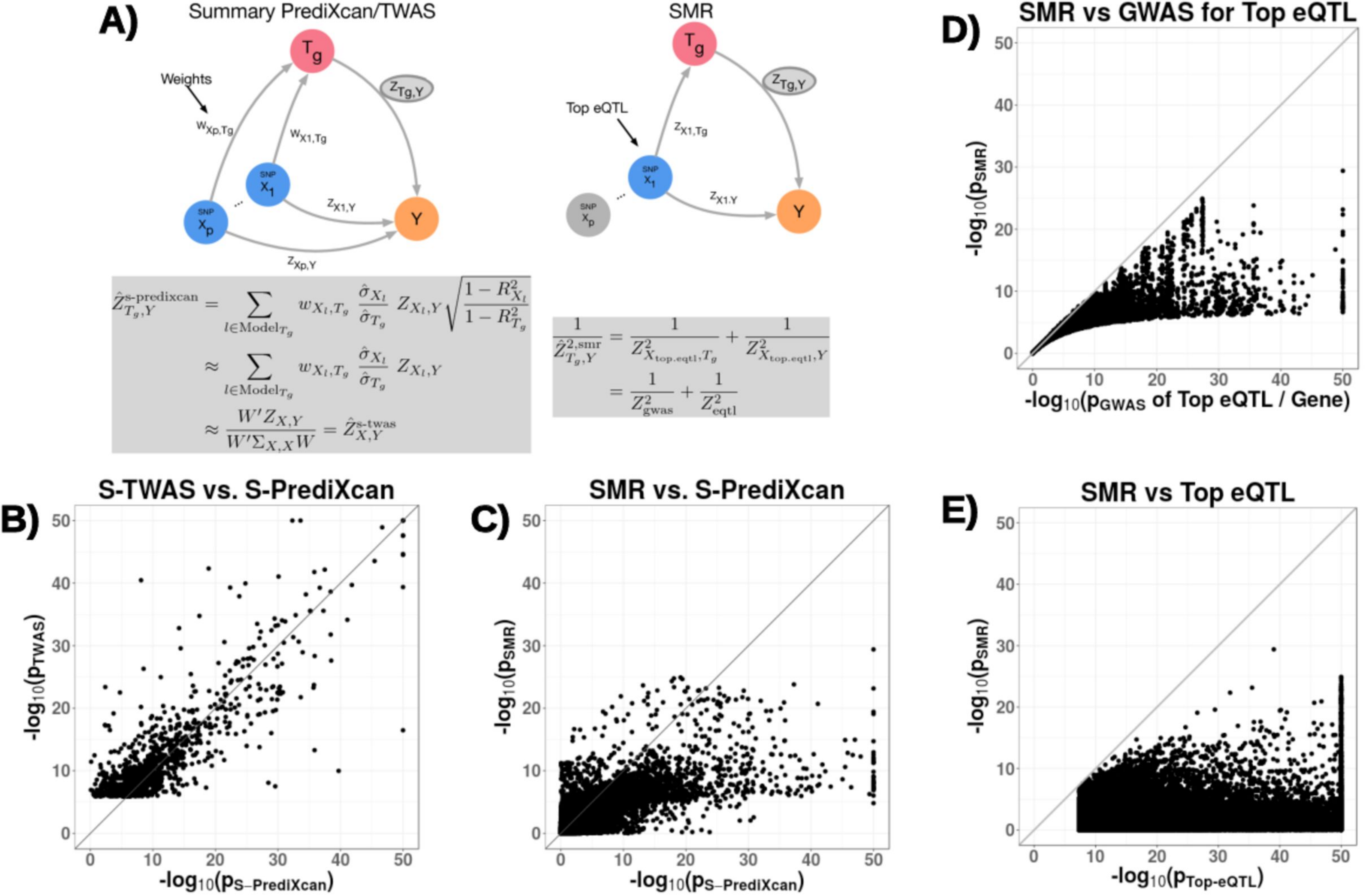
A

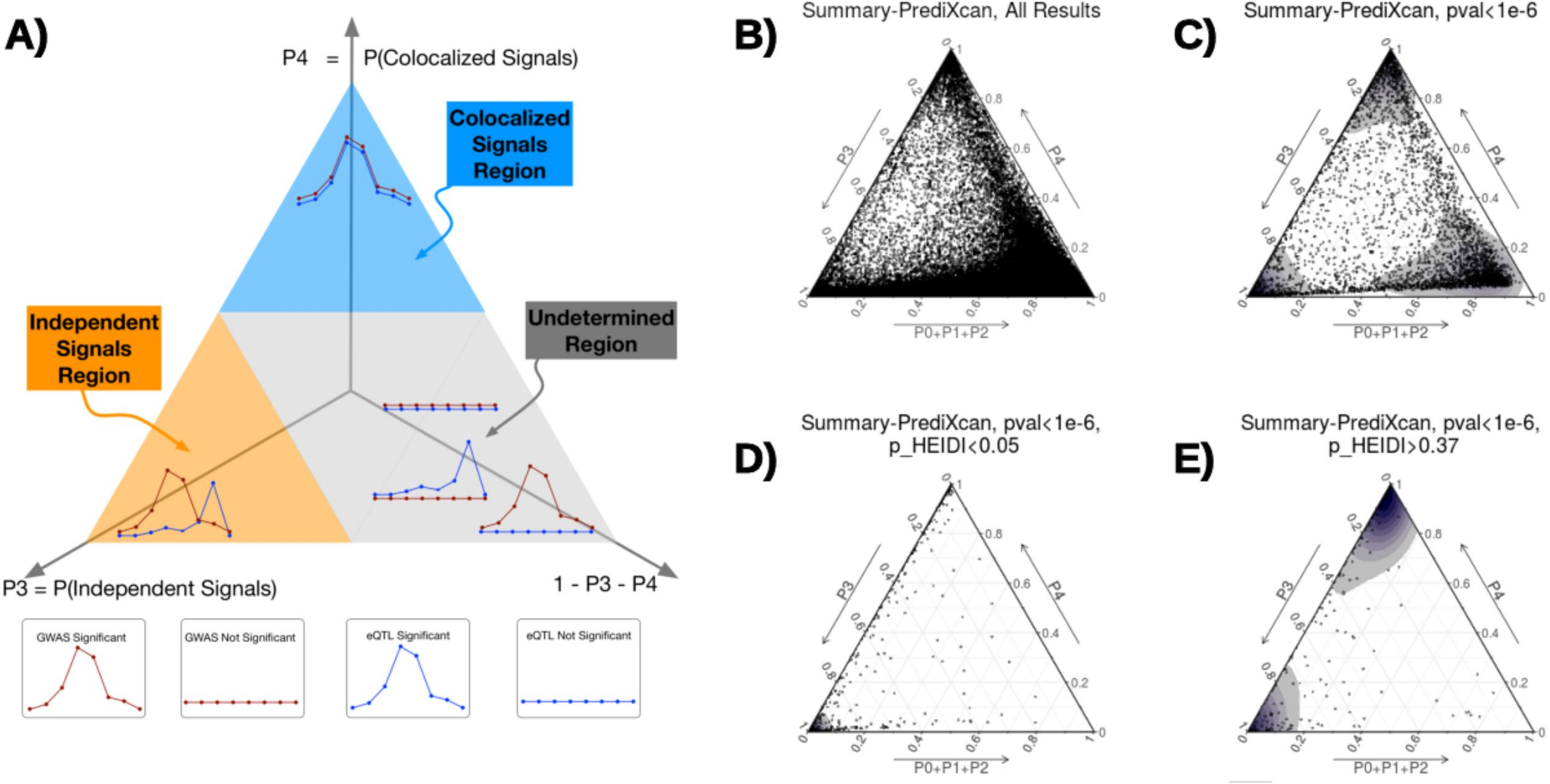
	Train Prediction Models	Filter-out Unreliable Models	Perform Gene-level Association	Adjust for Model Uncertainty	Filter-out Non-Colocalized Signals
PrediXcan	Elastic-Net	Correlation Observed vs Predicted FDR<0.05	Individual Level Data Y ~ Predicted X		Colocalization status (COLOC)
S-PrediXcan	Elastic-Net	Correlation Observed vs Predicted FDR<0.05	Summary Data $\sum_{l \in \text{Model}_{T_g}} w_{X_l,T_g} \frac{\hat\sigma_l}{\hat\sigma_{T_g}} Z_{X_l,Y}$		Colocalization status (COLOC)
SMR	Top eQTL	eQTL Significant	Summary Data $\frac{1}{\hat Z^{2,\text{smr}}_{T_g,Y}} = \frac{1}{Z^2_{\text{gwas}}} + \frac{1}{Z^2_{\text{eqtl}}}$		Heterogeneity (HEIDI)
S-TWAS	BSLMM	Heritability Above Threshold	Summary Data $\frac{W' Z_{X,Y}}{W' \Sigma_g W}$		

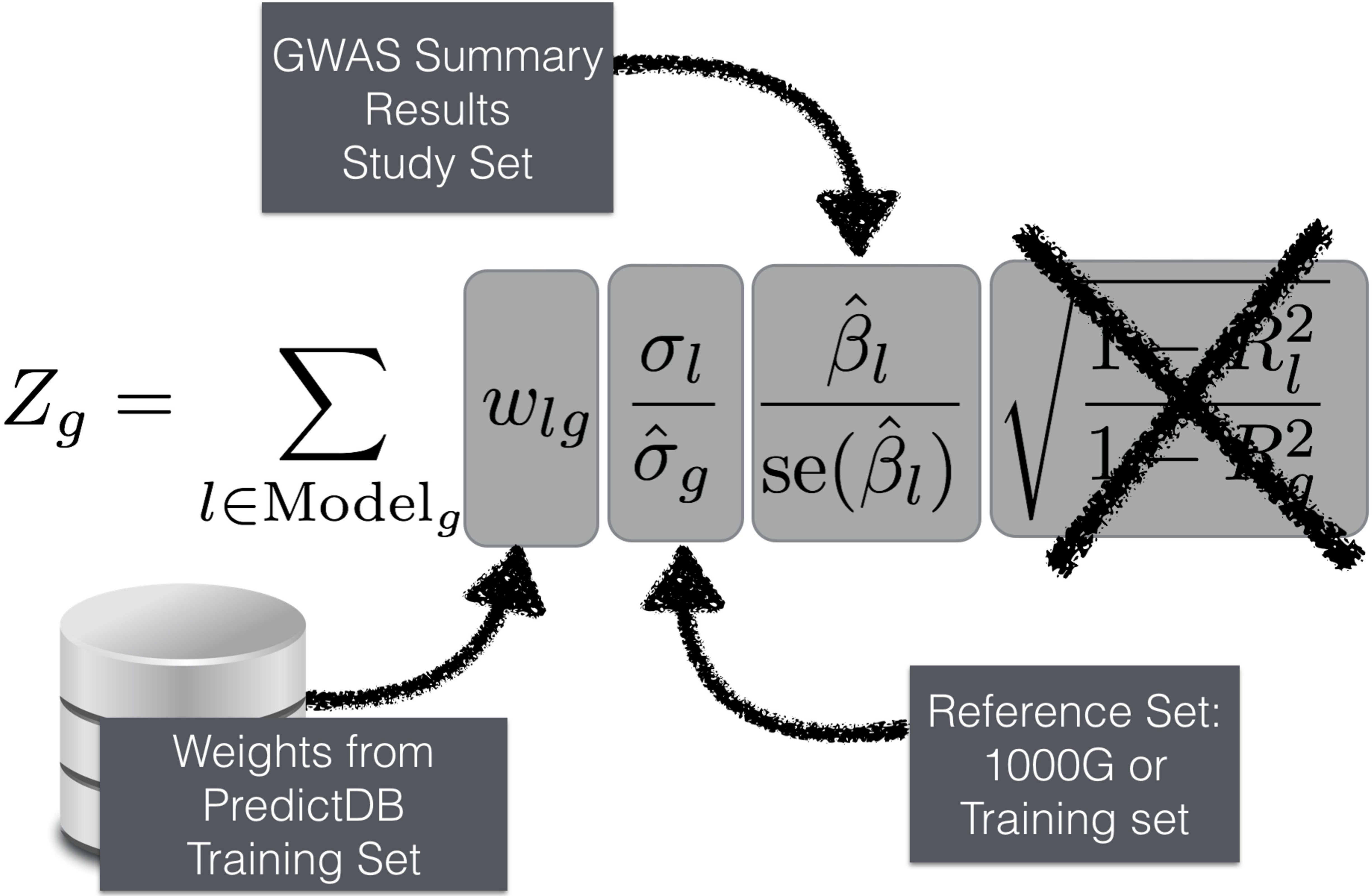
B




A)
QQ Uniform plot:
Summary PrediXcan results across all genes vs results for genes in ClinVar
Age-related Macular Degeneration
Alzheimer's Disease
Autistic Spectrum Disorder
Body Mass Index
Crohn's Disease
Rheumatoid Arthritis
Schizophrenia
Type 2 Diabetes
Ulcerative Colitis
Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
2
4
5
10
15
20
30
Colors : ● All Genes, ● Clinvar,
B)
SORT1 for
MI
LDL
CAD
PCSK9 for
MI
LDL
CAD
C4A for SCZ
ADPSBQ
ADPVSC
ADRNLG
ARTAORT
ARTCRN
ARTTBL
BREAST
BRNACC
BRNCDT
BRNCHA
BRNCHB
BRNCTXA
BRNCTXB
BRNHPP
BRNHPT
BRNPTM
CLNSGM
CLNTRN
DGN_WB
ESPGEJ
ESPMCS
FIBRBLS
HRTAA
HRTLV
LCL
LIVER
LUNG
MSCLSK
NERVET
PNCREAS
PRSTTE
PTTARY
SKINNS
SKINS
SNTTRM
SPLEEN
STMACH
TESTIS
THYROID
UTERUS
VAGINA
WHLBLD
|Z|
10
20
$\mathrm{sgn}(Z) * R^2_{pred}$
-0.25
0.00
0.25


A)
Summary PrediXcan/TWAS
T_g
Weights
$Z_{Tg,Y}$
$w_{Xp,Tg}$
$w_{X1,Tg}$
SNP X_1
SNP X_p
...
$Z_{X1,Y}$
Y
$Z_{Xp,Y}$
$$\hat{Z}^{\text{s-predixcan}}_{T_g,Y} = \sum_{l\in\text{Model}_{T_g}} w_{X_l,T_g}\frac{\hat{\sigma}_{X_l}}{\hat{\sigma}_{T_g}} Z_{X_l,Y}\sqrt{\frac{1-R^2_{X_l}}{1-R^2_{T_g}}}$$
$$\approx \sum_{l\in\text{Model}_{T_g}} w_{X_l,T_g}\frac{\hat{\sigma}_{X_l}}{\hat{\sigma}_{T_g}} Z_{X_l,Y}$$
$$\approx \frac{W'Z_{X,Y}}{W'\Sigma_{X,X}W} = \hat{Z}^{\text{s-twas}}_{X,Y}$$
SMR
T_g
$Z_{Tg,Y}$
Top eQTL
$Z_{X1,Tg}$
SNP X_1
SNP X_p
...
$Z_{X1,Y}$
Y
$$\frac{1}{\hat{Z}^{2,\text{smr}}_{T_g,Y}} = \frac{1}{Z^2_{X_{\text{top.eqtl}},T_g}} + \frac{1}{Z^2_{X_{\text{top.eqtl}},Y}} = \frac{1}{Z^2_{\text{gwas}}} + \frac{1}{Z^2_{\text{eqtl}}}$$
D)
SMR vs GWAS for Top eQTL
$-\log_{10}(p_{\text{SMR}})$
$-\log_{10}(p_{\text{GWAS}}$ of Top eQTL / Gene)
0 10 20 30 40 50
B)
S-TWAS vs. S-PrediXcan
$-\log_{10}(p_{\text{TWAS}})$
$-\log_{10}(p_{\text{S-PrediXcan}})$
0 10 20 30 40 50
C)
SMR vs. S-PrediXcan
$-\log_{10}(p_{\text{SMR}})$
$-\log_{10}(p_{\text{S-PrediXcan}})$
0 10 20 30 40 50
E)
SMR vs Top eQTL
$-\log_{10}(p_{\text{SMR}})$
$-\log_{10}(p_{\text{Top-eQTL}})$
0 10 20 30 40 50


A)
P4 = P(Colocalized Signals)
Colocalized Signals Region
Independent Signals Region
Undetermined Region
P3 = P(Independent Signals)
1 - P3 - P4
GWAS Significant
GWAS Not Significant
eQTL Significant
eQTL Not Significant
B)
Summary-PrediXcan, All Results
C)
Summary-PrediXcan, pval<1e-6
D)
Summary-PrediXcan, pval<1e-6, p_HEIDI<0.05
E)
Summary-PrediXcan, pval<1e-6, p_HEIDI>0.37
P3
P4
P0+P1+P2
0
0.2
0.4
0.6
0.8
1



GWAS Summary Results Study Set

$$Z_g = \sum_{l \in \mathrm{Model}_g} \boxed{w_{lg}} \boxed{\frac{\sigma_l}{\hat{\sigma}_g}} \boxed{\frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)}} \boxed{\xcancel{\sqrt{\frac{1-R_l^2}{1-R_g^2}}}}$$

Weights from PredictDB Training Set

Reference Set: 1000G or Training set