

1 Integrating tissue specific mechanisms into GWAS summary 2 results

3 Alvaro N. Barbeira¹, Scott P. Dickinson¹, Jason M. Torres², Rodrigo Bonazzola¹, Jiamao Zheng¹, Eric

4 S. Torstenson³, Heather E. Wheeler⁴, Kaanan P. Shah¹, Todd Edwards³, Tzintzuni Garcia⁵, GTEx

5 Consortium, Dan L. Nicolae¹, Nancy J. Cox³, Hae Kyung Im^{1,*}

6 1 Section of Genetic Medicine, The University of Chicago, Chicago, IL, USA

7 2 Committee on Molecular Metabolism and Nutrition, The University of Chicago,

8 Chicago, IL, USA

9 3 Vanderbilt Genetic Institute, Vanderbilt University Medical Center, Nashville, TN, USA

10 4 Departments of Biology and Computer Science, Loyola University Chicago, Chicago, IL,

11 USA

12 5 Center for Research Informatics, The University of Chicago, IL, USA

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

14 Abstract

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

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

28 Introduction

29 Over the last decade, GWAS have been successful in identifying genetic loci that robustly associate with human
30 complex traits. However, the mechanistic understanding of these discoveries is still limited, hampering the
31 translation of the associations into actionable targets. Studies of enrichment of expression quantitative trait
32 loci (eQTLs) among trait-associated variants [1–3] show the importance of gene expression regulation.
33 Functional class quantification showed that 80% of the common variant contribution to phenotype variability in
34 12 diseases can be attributed to DNAase I hypersensitivity sites, further highlighting the importance of
35 transcript regulation in determining phenotypes [4].

36 Many transcriptome studies have been conducted where genotypes and expression levels are assayed for a
37 large number of individuals [5–8]. The most comprehensive transcriptome dataset, in terms of examined
38 tissues, is the Genotype-Tissue Expression Project (GTEx); a large-scale effort where DNA and RNA were
39 collected from multiple tissue samples from nearly 1000 individuals and sequenced to high coverage [9,10].
40 This remarkable resource provides a comprehensive cross-tissue survey of the functional consequences of
41 genetic variation at the transcript level.

42 To integrate knowledge generated from these large-scale transcriptome studies and shed light on disease
43 biology, we developed PrediXcan [11], a gene-level association approach that tests the mediating effects of
44 gene expression levels on phenotypes. PrediXcan is implemented on GWAS or sequencing studies (i.e. studies
45 with genome-wide interrogation of DNA variation and phenotypes) where transcriptome levels are imputed
46 with models trained in measured transcriptome datasets (e.g. GTEx). These predicted expression levels are
47 then correlated with the phenotype in a gene association test that addresses some of the key limitations of
48 GWAS [11].

49 Meta-analysis efforts that aggregate results from multiple GWAS have been able to identify an increasing
50 number of associations that were not detected with smaller sample sizes [12–14]. We will refer to these results
51 as GWAMA (Genome-wide association meta-analysis) results. In order to harness the power of these increased
52 sample sizes while keeping the computational burden manageable, methods that use summary level data
53 rather than individual level data are needed.

54 A method based on similar ideas to PrediXcan was proposed by Gusev et al. [15] called Transcriptome-wide
55 Association Study (TWAS). For the individual level data based version, the main difference between PrediXcan
56 and TWAS resides in the models used for the prediction of gene expression levels in each implementation. An
57 important extension of this approach was implemented by Gusev et al. [15] that allows the computation of

58 gene-level association results using only summary statistics. We will refer to this method as Summary-TWAS (or
59 S-TWAS for short).

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

66 To examine whether eQTL and GWAS hits in close proximity share the same underlying causal signal,
67 several methods have been developed such as RTC [1], Sherlock [18], COLOC [19], and more recently eCAVIAR
68 [20] and ENLOC [21]. Thorough comparison between RTC, COLOC, and eCAVIAR can be found in [20]. HEIDI,
69 part of SMR, is another approach that computes the degree of non-colocalization or heterogeneity of signals.

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

74 We apply the MetaXcan framework by first training over 1 million elastic net prediction models of gene
75 expression traits, covering protein coding genes across 44 human tissues from GTEx, and then performing gene-
76 level association tests for 101 phenotypes from 37 large meta-analysis consortia.

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

83 **Results**

84 **Inferring PrediXcan results with summary statistics**

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

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

98 **MetaXcan framework**

99 Building on S-PrediXcan and existing approaches, we define a general framework (MetaXcan) to integrate QTL
100 information with GWAS results to map disease-associated genes as illustrated on Figure 2. This evolving
101 framework will incorporate state of the art models and methods to increase the power to detect causal genes
102 and filter out false positives. Existing methods fit within this general framework as instances or components as
103 outlined in Figure 2-A.

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

110 PrediXcan implementations work mostly with elastic net models motivated by our observation that gene
111 expression variation is mostly driven by sparse components [22]. TWAS implementations have used Bayesian
112 Sparse Linear Mixed Models [23] (BSLMM), which allows both polygenic and sparse components. SMR fits into
113 this scheme with prediction models consisting solely of the top eQTL for each gene (weights are not necessary
114 here since only one SNP is used at a time).

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

119 For the last step, we chose COLOC to estimate the probability of colocalization of GWAS and eQTL signals.
120 SMR uses its own estimate of “heterogeneity” of signals calculated by HEIDI. We chose to use COLOC
121 probabilities because COLOC clusters more distinctly into different classes and, unlike other methods, does not
122 require an arbitrary cut off threshold. Another advantage of COLOC is that for genes with low probability of
123 colocalization, it further distinguishes distinct GWAS and eQTL signals from low power. This is a useful feature
124 that future development of colocalization methods should also offer.

125 **Gene expression variation in humans is associated to diverse phenotypes**

126 Next, we downloaded summary statistics from meta analyses of 101 phenotypes from 37 consortia. The full list
127 of consortia and phenotypes is shown in Supplementary Table 3. We tested association between these
128 phenotypes and the predicted expression levels using elastic net models in 44 human tissues from GTEx as
129 described in the Methods section, and a whole blood model from the DGN cohort presented in [11].

130 We used a Bonferroni threshold accounting for all the gene-tissue pairs that were tested ($0.05/\text{total}$
131 number of gene-tissue pairs $\approx 2.5e-7$). This approach is conservative because the correlation between tissues
132 would make the total number of independent tests smaller than the total number of gene-tissue pairs. Height
133 had the largest number of genes significantly associated with 1,690 unique genes (based on a GWAMA of 250K
134 individuals). Other polygenic diseases with a large number of associations include schizophrenia with 307
135 unique significant genes ($n = 150\text{K}$ individuals), low-density lipoprotein cholesterol (LDL-C) levels with 297
136 unique significant genes ($n = 188\text{K}$), other lipid levels, glycemic traits, and immune/inflammatory disorders such
137 as rheumatoid arthritis and inflammatory bowel disease. For other psychiatric phenotypes, a much smaller
138 number of significant associations was found, with 8 significant genes for bipolar disorder ($n = 16,731$) and
139 none for major depressive disorder ($n = 18,759$), probably due to smaller sample sizes, but also smaller effect
140 sizes.

141 When excluding genes with evidence of independent GWAS-eQTL signals ($P_3 > 0.5$), these numbers dropped
142 by about 10-20% to 1377 for height, 231 for schizophrenia, and 244 for LDL-C levels. If we further exclude
143 genes with low power to determine either shared or non-shared GWAS-eQTL signals, we find 642 genes for
144 height, 157 for schizophrenia, and 78 for LDL-C. The quantities for the full set of phenotypes can be found in
145 Supplementary Table 3.

146 Mostly, genome-wide significant genes tend to cluster around known SNP-level genome-wide significant
147 loci or sub-genome-wide significant loci. Regions with sub-genome-wide significant SNPs can yield genome-
148 wide significant results in S-PrediXcan because of the reduction in multiple testing and the increase in power

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

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

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

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

167 **Moderate changes in ClinVar gene expression is associated with milder phenotypes**

168 We reasoned that if complete knock out of monogenic disease genes cause severe forms of the disease, more
169 moderate alterations of gene expression levels (as effected by regulatory variation in the population) could
170 cause more moderate forms of the disease. Thus moderate alterations in expression levels of monogenic
171 disease genes (such as those driven by eQTLs) may have an effect on related complex traits, and this effect
172 could be captured by S-PrediXcan association statistics. To test this hypothesis, we obtained genes listed in the
173 ClinVar database [29] for obesity, rheumatoid arthritis, diabetes, Alzheimer's, Crohn's disease, ulcerative colitis,
174 age-related macular degeneration, schizophrenia, and autism. As postulated, we found enrichment of
175 significant S-PrediXcan associations for ClinVar genes for all tested phenotypes except for autism and
176 schizophrenia. The lack of significance for autism is probably due to insufficient power: the distribution of p-
177 values is close to the null distribution. In contrast, for schizophrenia, many significant genes were found in the
178 S-PrediXcan analysis. There are several reasons that may explain this lack of enrichment: genes identified with
179 GWAS and subsequently with S-PrediXcan have rather small effect sizes, so that it would not be surprising that
180 they were missed until very large sample sizes were aggregated; ClinVar genes may originate from rare

181 mutations that are not well covered by our prediction models, which are based on common variation (due to
182 limited sample sizes of eQTL studies and the minor allele frequency -MAF- filter used in GWAS studies); or the
183 mechanism of action of the schizophrenia linked ClinVar genes may be different than the alteration of
184 expression levels. Also, the pathogenicity of some of the ClinVar entries has been questioned [30]. The list of
185 diseases in ClinVar used to generate the enrichment figures can be found in Supplementary Table 1, along with
186 the corresponding association results.

187 **Agnostic scanning of a broad set of tissues enabled by GTEx improves discovery**

188 The broad coverage of tissues in our prediction models enabled us to examine the tissue specificity of
189 phenotypic associations of GWAS signals. We started by computing average enrichment of significance by
190 tissue. We used several measures of enrichment such as the mean Z-scores squared across all genes, or across
191 significant genes for different thresholds, as well as the proportion of significant genes for different thresholds.
192 We also compared the full distribution of the p-values of a given tissue relative to the remaining tissues.
193 Supplementary Figure 6 shows the average Z-score² as a measure of enrichment of each tissue by phenotype.

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

201 Next we focused on three genes whose functional role has been well established: *C4A* for schizophrenia
202 [31] and *SORT1* [32] and *PCSK9* both for LDL-C and cardiovascular disease. The S-PrediXcan results for these
203 genes and traits and regulatory activity by tissue (as measured by the proportion of expression explained by the
204 genetic component) are shown in Figure 3 with additional details in Supplementary Tables 5, 6, and 7.

205 *SORT1* is a gene with strong evidence for a causal role in LDL-C levels, and as a consequence, is likely to
206 affect risk for cardiovascular disease [32]. This gene is most actively regulated in liver (close to 50% of the
207 expression level of this gene is determined by the genetic component) with the most significant S-PrediXcan
208 association in liver (p-value ≈ 0 , $Z = -28.8$), consistent with our prior knowledge of lipid metabolism. In this
209 example, tissue specific results suggest a causal role of *SORT1* in liver.

210 However, in the following example, association results across multiple tissues do not allow us to
211 discriminate the tissue of action. *C4A* is a gene with strong evidence of causal effect on schizophrenia risk via
212 excessive synaptic pruning in the brain during development [31]. Our results show that *C4A* is associated with

213 schizophrenia risk in all tissues ($p < 2.5 \times 10^{-7}$ in 36 tissue models and $p < 0.05$ for the remaining 4 tissue
214 models).

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

221 *PCSK9* is a target of several LDL-C lowering drugs currently under trial to reduce cardiovascular events [33].
222 The STARNET study [34] profiled gene expression levels in cardiometabolic disease patients and showed tag
223 SNP rs12740374 to be a strong eQTL for *PCSK9* in visceral fat but not in liver. Consistent with this, our S-
224 PrediXcan results also show a highly significant association between *PCSK9* and LDL-C ($p \approx 10^{-13}$) in visceral fat
225 and not in liver (our training algorithm did not yield a prediction model for *PCSK9*, i.e. there was no evidence of
226 regulatory activity). In our results, however, the statistical evidence is much stronger in tibial nerve ($p \approx 10^{-27}$).
227 The association between *PCSK9* and coronary artery disease is also significant in tibial nerve ($p \approx 10^{-8}$) but only
228 nominally significant in visceral fat ($p \approx 0.02$). Accordingly, in our training set (GTEx), there is much stronger
229 evidence of regulation of this gene in tibial nerve compared to visceral fat. Moreover, visceral fat association
230 shows evidence of independent rather than shared GWAS and eQTL signals in the *PCSK9* locus (probability of
231 independent signals $P_3=0.69$ in LDL-C). It is likely that the relevant regulatory activity in visceral adipose tissue
232 was not detected in the GTEx samples for various reasons but it was detected in tibial nerve. Thus by looking
233 into all tissue results we increase the window of opportunities where we can detect the association.

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

239 **Replication in an independent cohort**

240 We used data from the Resource for Genetic Epidemiology Research on Adult Health and Aging study (GERA,
241 phs000674.v1.p1) [35,36]. This is a study led by the Kaiser Permanente Research Program on Genes,
242 Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics with over 100,000 participants.
243 We downloaded the data from dbGaP and performed GWAS followed by S-PrediXcan analysis of 22 conditions

244 available in the dataset in the European subset of the cohort. Genotypes were imputed using the University of
 245 Michigan server and principal components provided by the GERA study were used to adjust for population
 246 stratification. More details can be found in the Methods section.

247 For replication, we chose Coronary Artery Disease (CAD), LDL cholesterol levels, Triglyceride levels, and
 248 schizophrenia, which had closely related phenotypes in the GERA study and had a sufficiently large number of
 249 significant associations (FDR < 0.05) in the discovery set. Analysis and replication of the type 2 diabetes
 250 phenotype can be found in [37]. Coronary artery disease hits were compared with “Any

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

251 **Table 1. Replication of results in GERA.** Significant genes/tissue pairs were replicated using a closely matched
 252 phenotype in an independent dataset from the GERA cohort [35]. The significance threshold for replication was
 253 $p < 0.05$, concordant directions of effect, and meta-analysis p-value less than the Bonferroni threshold in the
 254 discovery set. π_1 is an estimate of proportion of true positives in the replication set. π_1 (all) uses all gene/tissue
 255 pairs whereas π_1 (sig) is computed using only gene/tissue pairs that were significant in the discovery set. The
 256 column “# replicated genes coloc or undeterm” is the number of replicated genes excluding the ones for which
 257 there was strong evidence of independent GWAS and eQTL signals.

258 cardiac event”, LDL cholesterol and triglyceride level signals were compared with “Dyslipidemia”, and
 259 schizophrenia was compared to “Any psychiatric event” in GERA.

260 First, we estimated the proportion of true associations in the replication set (these include LD-induced
 261 ones) using the π_1 statistics from the q-value approach [38]. This approach does not indicate which genes are
 262 true positives but provides an estimate of the proportion. If we take all genes in the replication set, the
 263 estimated proportions of true associations are 0.4% for “Any cardiac event”, 5.8% for “Dyslipidemia”, and 1.2%
 264 for schizophrenia (see third column in Table 1). When we compute π_1 for the subset of genes that were found
 265 to be Bonferroni significant in the discovery analysis we find that π_1 goes up ten to one hundred fold as shown
 266 in Table 1. Following standard practice in meta-analysis, we consider a gene to be replicated if the p-value in
 267 the replication set is <0.05, the direction of discovery and replication effects are the same, and the meta
 268 analyzed p-value is Bonferroni significant with the discovery threshold.

269 Among the 56 genes significantly associated with CAD in the discovery set, 6 (11%) were significantly
 270 associated with “Any cardiac event” in GERA. Using “Dyslipidemia” as the closest matching phenotype, 78.5%

271 and 43.5% of LDL and triglyceride genes replicated, respectively. Among the 285 genes associated with
272 schizophrenia in the discovery set, 51 (21%) replicated. The low replication rate for CAD and
273 Schizophrenia is likely due to the broad phenotype definitions in the replication.

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

275

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

277 Zhu et al. have proposed Summary Mendelian Randomization (SMR) [16], a summary data based Mendelian
278 randomization that integrates eQTL results to determine target genes of complex trait-associated GWAS loci.
279 They derive an approximate χ^2 -statistic (Eq 5 in [16]) for the mediating effect of the target gene expression on
280 the phenotype. This approximation is only valid in two extreme cases: when the eQTL association is much
281 stronger than the GWAS association or vice versa, when the GWAS association is much stronger than the eQTL
282 association. Without this assumption, the mean of the distribution is off by a factor of 4. See Methods section
283 for further details.

284 When the eQTL association is much stronger than the GWAS association, we show that the SMR statistic is
285 approximately equal to the GWAS χ^2 -statistics of the top eQTL for the gene, which is equal to the Summary-
286 PrediXcan Z -score² if top eQTL is used as predictor. See derivation in Methods section.

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

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

293 Gusev et al. have proposed Transcriptome-Wide Association Study based on summary statistics (S-TWAS),
294 which imputes the SNP level Z-scores into gene level Z-scores. This is a natural extension of ImpG [39] or DIST
295 [40], which are SNP-based methods that impute summary statistics of unmeasured SNPs using Gaussian
296 imputation [41]. If restricted to Gaussian imputation, we show that this approach is equivalent to predicting
297 expression levels using BLUP/Ridge Regression, which has been shown to be suboptimal for gene expression
298 traits [22]. However, the mathematical expression used by S-TWAS can be extended to any set of weights such
299 as Bayesian Sparse Linear Mixed Models (BSLMM) as used by Gusev et al. [15]. S-TWAS imputes the Z-score of
300 the gene-level result assuming that under the null hypothesis the Z-scores are normally distributed with the
301 same correlation structure as the SNPs whereas in S-PrediXcan we compute the result of PrediXcan using
302 summary statistics. In the Methods Section we establish the approximate equivalence of the two approaches

303 when the same prediction weights are applied. Figure 4-A illustrates the components of SMR, S-TWAS, and S-
304 PrediXcan methods. All three seek to identify target genes by computing the strength of association between
305 the unobserved predicted expression levels (T_g) of a gene with the complex trait (Y) quantified with $Z_{T_g, Y}$ or its
306 square. SMR also incorporates uncertainty of the predicted expression in the statistics and adds a test for (non-
307) colocalization of GWAS and eQTL hits (HEIDI).

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

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

320

321 **Colocalization estimates complement PrediXcan results**

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

328 COLOC provides the probability of 5 hypotheses: H0 corresponds to no eQTL and no GWAS association, H1
329 and H2 correspond to association with eQTL but no GWAS or vice-versa, H3 corresponds to eQTL and GWAS
330 association but independent signals, and finally H4 corresponds to shared eQTL and GWAS association. P0, P1,
331 P2, P3, and P4 are the corresponding probabilities for each configuration. The sum of the five probabilities is 1.

332 Figure 5 shows ternary plots [42] with P3, P4, and 1-P3-P4 as vertices (for convenience we aggregate H0,
333 H1, and H2 into one event with probability 1-P3-P4). This representation restricts the sum to be 1. The top
334 vertex corresponds to probability of colocalized eQTL and GWAS signals (P4) to be high. The bottom left vertex

335 corresponds to distinct eQTL and GWAS signals (P3 high). The bottom right vertex corresponds to low
336 probability of both colocalization and independent signals, which the authors [19] recommend to interpret as
337 limited power.

338 Figure 5-B shows association results for all gene/tissue pairs to the height phenotype. We find that most
339 gene-tissue pairs' association falls in the bottom right, "undetermined" region. When we restrict the plot to S-
340 PrediXcan significant genes ($p\text{-value} < 1E-6$) (Figure 5-C), three distinct peaks emerge in the high P4 region
341 ("colocalized signals"), high P3 region ("independent signals"), and "undetermined" region. Moreover, when
342 genes with low prediction performance are excluded (Supplementary Figure 7-D) the "undetermined" peak
343 significantly diminishes.

344 These clusters provide a natural way to classify significant genes and complement S-PrediXcan results.
345 Depending on false positive/false negative trade-off choices, genes in the "independent signals" or both
346 "independent signals" and "undetermined" can be filtered out.

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

355 Discussion

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

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

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

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

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

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

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

393 **Software and Resources**

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

399 **Methods**

400 **Summary-PrediXcan formula**

401 Figure 6 shows the main analytic expression used by Summary-PrediXcan for the Z-score (Wald statistic) of the
402 association between predicted gene expression and a phenotype. The input variables are the weights used to
403 predict the expression of a given gene, the variance and covariances of the markers included in the prediction,
404 and the GWAS coefficient for each marker. The last factor in the formula can be computed exactly in principle,
405 but we would need additional information that is unavailable in typical GWAS summary statistics output such
406 as phenotype variance and sample size. Dropping this factor from the formula does not affect the accuracy of
407 the results as demonstrated in the close to perfect concordance between PrediXcan and Summary-PrediXcan
408 results on the diagonal of Figure 1A.

409 The approximate formula we use is:

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

410 where

- 411 • w_{lg} is the weight of SNP l in the prediction of the expression of gene g ,
- 412 • $\hat{\beta}_l$ is the GWAS regression coefficients for SNP l ,
- 413 • $se(\hat{\beta}_l)$ is standard error of $\hat{\beta}_l$,
- 414 • $\hat{\sigma}_l$ is the estimated variance of SNP l ,
- 415 • $\hat{\sigma}_g$ is the estimated variance of the predicted expression of gene g , and

416 • dosage and alternate allele are assumed to be the same.

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

418 • study set (e.g. GWAS study set),

419 • training set (e.g. GTEx, DGN),

420 • population reference set (e.g. the training set or 1000 Genomes).

421 The study set is the main dataset of interest from which the genotype and phenotypes of interest are
422 gathered. The regression coefficients and standard errors are computed based on individual-level data from the
423 study set or a SNP-level meta-analysis of multiple GWAS. Training sets are the reference transcriptome datasets
424 used for the training of the prediction models (GTEx, DGN, Framingham, etc.) thus the weights w_{lg} are
425 computed from this set. Training sets are also used to generate variance and covariances of genetic markers,
426 which will usually be different from the study sets. When individual level data are not available from the
427 training set we use population reference sets such as 1000 Genomes data.

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

431 **Performance in simulated data**

432 We first compared PrediXcan and Summary-PrediXcan using simulated phenotypes and a single transcriptome
433 model trained on Depression Genes and Network's (DGN) Whole Blood data set [5,22] downloaded from
434 PredictDB (<http://predictdb.org>). The phenotype was sampled from a normal distribution without any link to
435 genotype. For genotypes we used three ancestral subsets of the 1000 Genomes project: Africans (n=661), East
436 Asians (n=504), and Europeans (n=503). Each set was taken in turn as reference and study set yielding a total of
437 9 combinations as shown in Figure 1B. For each population combination, we computed PrediXcan association
438 results for the simulated phenotype and compared them with results generated using S-PrediXcan in a scatter
439 plot. In this manner we assess the effect of ancestral differences between study and reference sets.

440 As expected, when the study and reference sets are the same, the concordance between PrediXcan and S-
441 PrediXcan is 100%, whereas for sets of different ancestral origin the R^2 drops a few percentage points, with the
442 biggest loss (down to 85%) when the study set is African and the reference set is Asian. This confirms that our
443 formula works as expected and that the approach is robust to ethnic differences between study and reference
444 sets.

445 **Performance in cellular growth phenotype from 1000 genomes cell lines**

446 Next we tested with an actual cellular phenotype - intrinsic growth. This phenotype was computed based on
447 multiple growth assays for over 500 cell lines from the 1000 Genomes project [43]. We used a subset of values
448 for European (EUR), African (AFR), and Asian (EAS) individuals.

449 We compared Z-scores for intrinsic growth generated by PrediXcan and S-PrediXcan for different
450 combinations of reference and study sets, using whole blood prediction models trained in the DGN cohort. The
451 results are shown in Supplementary Figure 1B. Consistent with our simulation study, the S-PrediXcan results
452 closely match the PrediXcan results. Again, the best concordance occurs when reference and study sets share
453 similar continental ancestry while differences in population slightly reduce concordance. Compared to the plots
454 for the simulated phenotypes, the diagonal concordance is slightly lower than 1. This is due to the fact that
455 more individuals were included in the reference set than in the study set, thus the study and reference sets
456 were not identical for S-PrediXcan.

457 **Performance on disease phenotypes from WTCCC**

458 We show the comparison of PrediXcan and summary-PrediXcan results for two diseases: Bipolar Disorder (BD)
459 and Type 1 Diabetes (T1D) from the WTCCC in Supplementary Figure 1C. Other diseases exhibited similar
460 performance (data not shown). Concordance between PrediXcan and Summary-PrediXcan is over 99% for both
461 diseases (BD $R^2 = 0.996$ and T1D $R^2 = 0.995$). The very small discrepancies are explained by differences in allele
462 frequencies and LD between the reference set (1000 Genomes) and the study set (WTCCC).

463 It is worth noting that the PrediXcan results for diseases were obtained using logistic regression whereas
464 Summary-PrediXcan formula is based on linear regression. As observed before [23], when the number of cases
465 and controls are relatively well balanced (roughly, at least 25% of a cohort are cases or controls), linear
466 regression approximation yields very similar results to logistic regression. This high concordance also shows
467 that the approximation of dropping the factor $\sqrt{\frac{1-R_l^2}{1-R_g^2}}$ does not significantly affect the results.

468 **Derivation of Summary-PrediXcan Formula**

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

472 **Notation and Preliminaries**

473 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
 474 predicted expression (or estimated GREx, genetically regulated expression). w_{lg} are weights to predict
 475 expression $T_g = \sum_{l \in \text{Model } g} w_{lg} X_l$, derived from an independent training set.

476 We model the phenotype as linear functions of X_l and T_g

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

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

477 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
 478 estimated regression coefficients of Y regressed on T_g and X_l , respectively. $\hat{\gamma}_g$ is the result (effect size for gene
 479 g) we get from PrediXcan whereas $\hat{\beta}_l$ is the result from a GWAS for SNP l .

480 We will denote as $\widehat{\text{Var}}$ and $\widehat{\text{Cov}}$ the operators that compute the sample variance and covariance, i.e.:

481 $\widehat{\text{Var}}(Y) = \hat{\sigma}_Y^2 = \sum_{i=1, n} (Y_i - \bar{Y})^2 / (n - 1)$ with $\bar{Y} = \sum_{i=1, n} Y_i / n$. Let $\hat{\sigma}_l^2 = \widehat{\text{Var}}(X_l)$, $\hat{\sigma}_g^2 = \hat{\sigma}_l^2 = \widehat{\text{Var}}(T_g)$

482 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
 483 l has the column mean of \mathbf{X}_l (p being the number of SNPs in the model for gene g , typically $p \ll n$).

484 With this notation, our goal is to infer PrediXcan results ($\hat{\gamma}_g$ and its standard error) using only GWAS results
 485 ($\hat{\beta}_l$ and their standard error), estimated variances of SNPs ($\hat{\sigma}_l^2$), estimated covariances between SNPs in each
 486 gene model (Γ_g), and prediction model weights w_{lg} .

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

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

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

489 and

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

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

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

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

491 By definition

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

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

$$\begin{aligned} \hat{\sigma}_g^2 &= \widehat{\text{Var}} \left(\sum_{l \in \text{Model}_g} w_{lg} X_l \right) \\ &= \widehat{\text{Var}}(\mathbf{W}_g \mathbf{X}_g) \\ &= \mathbf{W}_g' \widehat{\text{Var}}(\mathbf{X}_g) \mathbf{W}_g \end{aligned}$$

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

$$495 \quad \hat{\sigma}_g^2 = \mathbf{W}_g' \Gamma_g \mathbf{W}_g \quad (4)$$

496

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

498 $\hat{\gamma}_g$ can be expressed as

$$\begin{aligned} \hat{\gamma}_g &= \frac{\widehat{\text{Cov}}(T_g, Y)}{\hat{\sigma}_g^2} \\ &= \frac{\widehat{\text{Cov}} \left(\sum_{l \in \text{Model}_g} w_{lg} X_l, Y \right)}{\hat{\sigma}_g^2} \\ &= \sum_{l \in \text{Model}_g} \frac{w_{lg} \widehat{\text{Cov}}(X_l, Y)}{\hat{\sigma}_g^2} \end{aligned}$$

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

500

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

501

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

503 Also from the properties of linear regression we know that

504

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

505 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 :

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

506 Thus:

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

507

508 Notice that the right hand side of (8) is dependent on the SNP l while the left hand side is not. This equality
509 will hold only approximately in our implementation since we will be using approximate values for $\hat{\sigma}_l^2$, i.e. from
510 reference population, not the actual study population.

511 Calculation of Z-score

512 To assess the significance of the association, we need to compute the ratio of the estimated effect size $\hat{\gamma}_g$ and
513 standard error $\text{se}(\hat{\gamma}_g)$, or Z-score,

$$Z_g = \frac{\hat{\gamma}_g}{\text{se}(\hat{\gamma}_g)} \quad (9)$$

514

515

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

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

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

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

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

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

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

522 Expression model training

523 To train our prediction models, we obtained genotype data and normalized gene expression data collected by
 524 the GTEx Project. We used 44 different tissues sampled by GTEx and thus generated 44 different tissue-wide
 525 models (dbGaP Accession phs000424.v6.p1). Sample sizes for different tissues range from 70 (Uterus) to 361
 526 (Muscle - Skeletal). The models referenced in this paper make use of the GTEx Project's V6p data, a patch to
 527 the version 6 data and makes use of improved gene-level annotation. We removed ambiguously stranded SNPs
 528 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
 529 MAF > 0.01. For each tissue, normalized gene expression data was adjusted for covariates such as gender,
 530 sequencing platform, the top 3 principal components from genotype data and top PEER Factors. The number of
 531 PEER Factors used was determined by sample size: 15 for n < 150, 30 for n between 150 and 250, and 35 for n >
 532 250. Covariate data was provided by GTEx. For our analysis, we used protein-coding genes only.

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

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

541 **Online Catalog and SMR, COLOC, TWAS**

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

544 Supplementary Table 3 shows the list of GWA/GWAMA studies we considered in this analysis. We applied S-
545 PrediXcan to these studies using the transcriptome models trained on GTEx studies for patched version 6. For
546 simplicity, S-PrediXcan only considers those SNPs that have a matching set of alleles in the prediction model,
547 and adjusts the dosages (2 – dosage) if the alleles are swapped.

548 To make the results of this study broadly accessible, we built a Postgre SQL relational database to store S-
549 PrediXcan results, and serve them via a web application.

550 We also applied SMR [16] to the same set of GWAMA studies, using the GTEx eQTL associations. We
551 downloaded version 0.66 of the software from the SMR website, and ran it using the default parameters. We
552 converted the GWAMA and GTEx eQTL studies to SMR input formats. In order to have SMR
553 compute the colocalization test, for those few GWAMA studies where allele frequency was not reported, we
554 filled in with frequencies from the 1000 Genomes Project [44] as an approximation. We also used the
555 1000 Genomes genotype data as reference panel for SMR.

556 Next we ran COLOC [19] over the same set of GWAMA and eQTL studies. We used the R package available
557 from CRAN. We used the Approximate Bayes Factor colocalization analysis, with the option that estimates the
558 phenotype variance from the variances and frequencies in each association study. When the frequency
559 information was missing from the GWAS, we filled in with data from the 1000 Genomes
560 Project.

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

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

567 **Comparison with TWAS**

568 Formal similarity with TWAS can be made more explicit by rewriting S-PrediXcan formula in matrix form. With
569 the following notation and definitions:

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

570 and correlation matrix of SNPs in the model for gene g

$$\Sigma_g = \text{diag}\left(\frac{1}{\hat{\sigma}_1}, \dots, \frac{1}{\hat{\sigma}_p}\right) \cdot \Gamma_g \cdot \text{diag}\left(\frac{1}{\hat{\sigma}_1}, \dots, \frac{1}{\hat{\sigma}_p}\right)$$

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

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

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

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

573 , thus

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

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

575 Summary-TWAS imputes the Z-score for the gene-level result assuming that under the null hypothesis, the Z-

576 scores are normally distributed with the same correlation structure as the SNPs; whereas in S-PrediXcan we

577 compute the results of PrediXcan using summary statistics. Thus, S-TWAS and S-PrediXcan yield equivalent

578 mathematical expressions (after setting the factor $\sqrt{\frac{(1-R_g^2)}{(1-R_g^2)}} \approx 1$

579

580 **Summary-PrediXcan with only top eQTL as predictor**

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

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

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

$$Z_{\text{top eQTL S-PrediXcan}}^2 = Z_{\text{GWAS}}^2 \quad (12)$$

584 **Comparison with SMR**

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

$$T_{\text{SMR}} = \frac{Z_{\text{eQTL}}^2 Z_{\text{GWAS}}^2}{Z_{\text{eQTL}}^2 + Z_{\text{GWAS}}^2} \quad (13)$$

587

588 Here Z_{eQTL} is the Z-score (= effect size/standard error) of the association between SNP and gene expression,
589 and Z_{GWAS} is the Z-score of the association between SNP and trait.

590 This SMR statistic is quite different from a χ_1 -square random variable as assumed in [16]. A quick simulation
591 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
592 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
593 Taylor expansions to find an interpretable form of the SMR statistic.

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

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

595

596 , so that for large enough Z_{eQTL}^2 relative to Z_{GWAS}^2 ,

$$T_{SMR} \approx Z_{GWAS}^2 = Z_{\text{top eQTL S-PrediXcan}}^2 \quad (15)$$

597

598 using equation 12. Thus, in this case, the SMR statistic is slightly smaller than the (top eQTL based) S-PrediXcan
599 χ_1 -square. This reduced significance is accounting for the uncertainty in the eQTL association. As the evidence
600 for eQTL association grows, the denominator Z_{eQTL}^2 increases and the difference tends to 0.

601 On the other extreme when the GWAS association is much stronger than the eQTLs, $Z_{eQTL} \ll Z_{GWAS}$,

$$T_{SMR} = \frac{Z_{eQTL}^2}{1 + \frac{Z_{eQTL}^2}{Z_{GWAS}^2}} \approx Z_{eQTL}^2 \left(1 - \frac{Z_{eQTL}^2}{Z_{GWAS}^2} \right) \quad (16)$$

602

603 , so that analogously:

$$T_{SMR} \approx Z_{eQTL}^2 \quad (17)$$

604

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

607 In between the two extremes, the right distribution must be computed using numerical methods. When we
608 look at the empirical distribution of the SMR statistic's p-value against the GWAS and eQTL (top eQTL for the
609 gene) p-values, we find the ceiling of the SMR statistic is maintained as shown in Figure 4-D/E. Given the rate of
610 growth of sample sizes of GWAS studies compared to eQTL studies, the power of eQTL studies will cap the
611 significance attainable by SMR. This approach seems unnecessarily conservative. In our framework, we use a
612 minimum prediction performance threshold and estimates of colocalization to filter out unreliable associations.

613 **GERA imputation**

614 Genotype files were obtained from dbGaP, and updated to release 35 of the probe annotations published by
615 Affymetrix via PLINK [45]. Probes were filtered out that had a minor allele frequency of <0.01, were missing in
616 >10% of subjects, or did not fit Hardy-Weinberg equilibrium. Subjects were dropped that had an unexpected
617 level of heterozygosity ($F > 0.05$). Finally the HRC-1000G-check-bim.pl script (from <http://www.well.ox.ac.uk/wrayner/tools/>)
618 was used to perform some final filtering and split data by chromosome. Phasing (via eagle v2.3
619 [46]) and imputation against the HRC r1.1 2016 panel [47] (via minimac3) were carried out by the Michigan
620 Imputation Server [48].

621 **GERA GWAS and MetaXcan Application**

622 European samples had been split into ten groups during imputation to ease the computational burden on the
623 Michigan server, so after obtaining the imputed .vcf files, we used the software PLINK [45] to convert the
624 genotype files into the PLINK binary file format and merge the ten groups of samples together, while dropping
625 any variants not found in all sample groups. For the association analysis, we performed a logistic regression
626 using PLINK, and following QC practices from [14] we filtered out individuals with genotype missingness > 0.03
627 and filtered out variants with minor allele frequency < 0.01, missingness > 0.05, out of Hardy-Weinberg
628 equilibrium significant at 1E-6, or had imputation quality < 0.8. We used gender and the first ten genetic
629 principal components as obtained from dbGaP as covariates. Following all filtering, our analysis included 61,444
630 European samples with 7,120,064 variants. MetaXcan was then applied to these GWAS results, using the 45
631 prediction models (GTEx and DGN).

632 **Acknowledgments**

633 **Grants**

634 We acknowledge the following US National Institutes of Health grants: R01MH107666 (H.K.I.), T32 MH020065
635 (K.P.S.), R01 MH101820 (GTEx), P30 DK20595 (Diabetes Research and Training Center), F31 DK101202 (J.M.T.),
636 P50 DA037844 (Rat Genomics), P50 MH094267 (Conte). H.E.W. was supported in part by start-up funds from
637 Loyola University Chicago.

638 The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the
639 Director of the National Institutes of Health. Additional funds were provided by the NCI, NHGRI, NHLBI, NIDA,
640 NIMH, and NINDS. Donors were enrolled at Biospecimen Source Sites funded by NCI
641 SAIC-Frederick, Inc. (SAIC-F) subcontracts to the National Disease Research Interchange (10XS170), Roswell Park
642 Cancer Institute (10XS171), and Science Care, Inc. (X10S172). The Laboratory, Data Analysis, and Coordinating
643 Center (LDACC) was funded through a contract (HHSN268201000029C) to The Broad Institute, Inc.
644 Biorepository operations were funded through an SAIC-F subcontract to Van Andel Institute (10ST1035).
645 Additional data repository and project management were provided by SAIC-F (HHSN261200800001E). The
646 Brain Bank was supported by a supplements to University of Miami grants DA006227 & DA033684 and to
647 contract N01MH000028. Statistical Methods development grants were made to the University of Geneva
648 (MH090941 & MH101814), the University of Chicago (MH090951, MH090937, MH101820, MH101825), the
649 University of North Carolina - Chapel Hill (MH090936 & MH101819), Harvard University (MH090948), Stanford
650 University (MH101782), Washington University St Louis (MH101810), and the University of Pennsylvania

651 (MH101822). The data used for the analyses described in this manuscript were obtained from dbGaP accession
652 number phs000424.v6.p1 on 06/17/2016.

653 This work was completed in part with resources provided by the University of Chicago Research Computing
654 Center, Bionimbus [49], and the Center for Research Informatics. The Center for Research Informatics is funded
655 by the Biological Sciences Division at the University of Chicago with additional funding provided by the Institute
656 for Translational Medicine, CTSA grant number UL1 TR000430 from the National Institutes of Health.

657 **References**

- 658 1. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, et al. Candidate causal regulatory
659 effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genetics*.
660 2010;6(4).
- 661 2. Nicolae DL, Gamazon E, Zhang W, Duan S, Eileen Dolan M, Cox NJ. Trait-associated SNPs are more likely
662 to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genetics*. 2010;6(4).
- 663 3. Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary
664 link between genetic variation and disease. *Science*. 2016;352(6285):600–604. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27126046>.
665
- 666 4. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning heritability of regulatory
667 and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics*.
668 2014;95(5):535–552.
- 669 5. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing the genetic
670 basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*.
671 2014;24(1):14–24.
- 672 6. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PaC, Monlong J, Rivas Ma, et al.
673 Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*.
674 2013;501(7468):506–11. Available from: [http://www.pubmedcentral.nih.gov/articlerender.
675 fcgi?artid=3918453&tool=pmcentrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3918453&tool=pmcentrez&rendertype=abstract).
- 676 7. Zhang X, Joehanes R, Chen BH, Huan T, Ying S, Munson PJ, et al. Identification of common genetic
677 variants controlling transcript isoform variation in human whole blood. *Nature Genetics*.
678 2015;47(4):345–352. Available from: <http://www.nature.com/doifinder/10.1038/ng.3220>.

- 679 8. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of Cis regulatory
680 variation in diverse human populations. *PLoS Genetics*. 2012;8(4).
- 681 9. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nature genetics*. 2013;45(6):580–
682 5. Available from: [http://www.pubmedcentral.nih.gov/articlerender.fcgi?
683 artid=4010069&tool=pmcentrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4010069&tool=pmcentrez&rendertype=abstract).
- 684 10. Aguet F, Brown AA, Castel S, Davis JR, Mohammadi P, Segre AV, et al. Local genetic effects on gene
685 expression across 44 human tissues. *bioRxiv*. 2016; Available from: [http://biorxiv.org/
686 content/early/2016/09/09/074450](http://biorxiv.org/content/early/2016/09/09/074450).
- 687 11. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A genebased
688 association method for mapping traits using reference transcriptome data. *Nature genetics*.
689 2015;47(9):1091–1098. Available from: <http://dx.doi.org/10.1038/ng.3367>.
- 690 12. Smoller JW, Craddock N, Kendler K, Lee PH, Neale BM, Nurnberger JI, et al. Identification of risk loci with
691 shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*.
692 2013;381(9875):1371–9. Available from: [http://discovery.ucl.ac.uk/1395494/\\$\
693 delimiter"026E30F\\$nhttp://www.ncbi.nlm.nih.gov/pubmed/23453885](http://discovery.ucl.ac.uk/1395494/$\delimitter).
- 694 13. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, et al. Largescale association
695 analysis identifies new risk loci for coronary artery disease. *Nature genetics*. 2013;45(1):25–33. Available
696 from: [http://www.pubmedcentral.nih.gov/articlerender.fcgi?
697 artid=3679547&tool=pmcentrez&rendertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3679547&tool=pmcentrez&rendertype=abstract).
- 698 14. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, et al. Large-scale association
699 analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature*
700 *Genetics*. 2012;44(9):981–990. Available from:
701 [http://www.ncbi.nlm.nih.gov/pubmed/22885922\\$\
702 nature.com/doi/10.1038/ng.2383](http://www.ncbi.nlm.nih.gov/pubmed/22885922/$\delimitter).
- 703 15. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Integrative approaches for large-
704 scale transcriptome-wide association studies. *Nature Genetics*. 2016;48:245–252.
- 705 16. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS
706 and eQTL studies predicts complex trait gene targets. *Nature genetics*. 2016;48(5):481–7. Available

- 707 from: <http://www.nature.com/doi/10.1038/ng.3538>
708 delimiter"026E30F\$nh<http://www.ncbi.nlm.nih.gov/pubmed/27019110>.
- 709 17. Casella G, Berger R. Statistical Inference. 2nd ed. Imprint Australia; Pacific Grove, CA : Thomson
710 Learning, c2002.; 2002.
- 711 18. He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, et al. Sherlock: Detecting Gene-Disease Associations by
712 Matching Patterns of Expression QTL and GWAS. The American Journal of Human Genetics. 2013
713 May;92(5):667–680. Available from: <http://dx.doi.org/10.1016/j.ajhg.2013.03.022>.
- 714 19. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for
715 colocalisation between pairs of genetic association studies using summary statistics. PLoS Genetics. 2014
716 May;10(5):e1004383. Available from: [http://eutils.ncbi.nlm.nih.gov/entrez/
717 eutils/elink.fcgi?dbfrom=pubmed&id=24830394&retmode=ref&cmd=prlinks](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=24830394&retmode=ref&cmd=prlinks).
- 718 20. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al.; Los Angeles Los Angeles CA 90095
719 USA. Department of Computer Science. Colocalization of GWAS and eQTL
720 Signals Detects Target Genes. Am J Hum Genet. 2016; Available from: [http://dx.doi.org/10.
721 1016/j.ajhg.2016.10.003](http://dx.doi.org/10.1016/j.ajhg.2016.10.003).
- 722 21. Wen X, Pique-Regi R, Luca F. Integrating molecular QTL data into genome-wide genetic association
723 analysis: Probabilistic assessment of enrichment and colocalization. PLoS Genetics. 2017
724 Mar;13(3):e1006646. Available from: <http://dx.plos.org/10.1371/journal.pgen.1006646>.
- 725 22. Wheeler HE, Shah KP, Brenner J, Garcia T, Aquino-Michaels K, Cox NJ, et al. Survey of the Heritability and
726 Sparse Architecture of Gene Expression Traits across Human Tissues. PLoS Genetics. 2016;12(11).
- 727 23. Zhou X, Carbonetto P, Stephens M. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS
728 Genetics. 2013;9(2).
- 729 24. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, et al. Modeling sample
730 variables with an Experimental Factor Ontology. Bioinformatics. 2010;26(8):1112–1118.
- 731 25. Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, et al. The Human Phenotype
732 Ontology project: Linking molecular biology and disease through phenotype data. Nucleic Acids
733 Research. 2014;42(D1).
- 734 26. Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A, Pasaniuc B. Integrating Gene Expression with
735 Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American

- 764 Aug;353(6301):827–830. Available from: <http://www.sciencemag.org/cgi/doi/10.1126/>
765 [science.aad6970](http://www.sciencemag.org/cgi/doi/10.1126/science.aad6970).
- 766 35. Hoffmann TJ, Ehret GB, Nandakumar P, Ranatunga D, Schaefer C, Kwok PY, et al. Genome-wide
767 association analyses using electronic health records identify new loci influencing blood pressure
768 variation. *Nature Genetics*. 2016 Nov;49(1):54–64. Available from: <http://www.nature.com/>
769 [doi/10.1038/ng.3715](http://www.nature.com/doi/10.1038/ng.3715).
- 770 36. Cook JP, Morris AP. Multi-ethnic genome-wide association study identifies novel locus for type 2
771 diabetes susceptibility. *European Journal of Human Genetics*. 2016 Aug;24(8):1175–1180. Available
772 from: <http://www.nature.com/doi/10.1038/ejhg.2016.17>.
- 773 37. Torres JM, Barbeira AN, Bonazzola R, Morris AP, Shah KP, Wheeler HE, et al. Integrative cross tissue
774 analysis of gene expression identifies novel type 2 diabetes genes. *bioRxiv*. 2017; Available from:
775 <http://biorxiv.org/content/early/2017/02/27/108134>.
- 776 38. Storey JD. Statistical significance for genomewide studies. *Proceedings of the National Academy of*
777 *Sciences*. 2003 Jul;100(16):9440–9445. Available from: [http://scholar.google.com.proxy.uchicago.edu/scholar?hl=en&lr=&q=info:eSXwkHMI-nQJ:](http://scholar.google.com.proxy.uchicago.edu/scholar?hl=en&lr=&q=info:eSXwkHMI-nQJ:scholar.google.com/&output=search)
778 scholar.google.com/&output=search.
- 779
- 780 39. Pasaniuc B, Zaitlen N, Shi H, Bhatia G, Gusev A, Pickrell J, et al. Fast and accurate imputation of summary
781 statistics enhances evidence of functional enrichment. *Bioinformatics (Oxford, England)*.
782 2014;30(20):2906–2914.
- 783 40. Lee D, Bigdeli TB, Riley BP, Fanous AH, Bacanu SA. DIST: Direct imputation of summary
784 statistics for unmeasured SNPs. *Bioinformatics*. 2013;29(22):2925–2927.
- 785 41. Wen X, Stephens M. Using linear predictors to impute allele frequencies from summary or pooled
786 genotype data. *The Annals of Applied Statistics*. 2010;4(3):1158–1182. Available from: <http://projecteuclid.org/euclid.aas/1287409368>.
- 787
- 788 42. Hamilton N. ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams; 2016. R package
789 version 2.2.0. Available from: <https://CRAN.R-project.org/package=ggtern>.
- 790 43. Im HK, Gamazon ER, Stark AL, Huang RS, Cox NJ, Dolan ME. Mixed effects modeling of proliferation rates
791 in cell-based models: Consequence for pharmacogenomics and Cancer. *PLoS Genetics*. 2012;8(2).

- 792 44. Auton A, Altshuler DM, Durbin RMJA, Wang J, Yang H, Auton A, et al. A global
793 reference for human genetic variation. *Nature*. 2015;526(7571):68–74. Available from:
794 [http://www.nature.com/nature/journal/v526/n7571/fig_{ }tab/nature15393_{ }SF1.](http://www.nature.com/nature/journal/v526/n7571/fig_{ }tab/nature15393_{ }SF1.html)
795 <http://dx.doi.org/10.1038/nature15393>{ }[http://www.ncbi.nlm.nih.gov/](http://www.ncbi.nlm.nih.gov/pubmed/26432245)
796 <http://www.nature.com/doi/10.1038/nature15393>.
- 797 45. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the
798 challenge of larger and richer datasets. *GigaScience*. 2015 dec;4(1):7. Available from:
799 <http://gigascience.biomedcentral.com/articles/10.1186/s13742-015-0047-8>.
- 800 46. Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Referencebased phasing
801 using the Haplotype Reference Consortium panel. *Nature Genetics*. 2016 oct;48(11):1443–1448.
802 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27694958><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5096458>[http://www.nature.](http://www.nature.com/doi/10.1038/ng.3679)
803 [com/doi/10.1038/ng.3679](http://www.nature.com/doi/10.1038/ng.3679).
- 804
- 805 47. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976
806 haplotypes for genotype imputation. *Nature Genetics*. 2016 aug;48(10):1279–1283. Available from:
807 <http://www.nature.com/doi/10.1038/ng.3643>.
- 808 48. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation
809 service and methods. *Nature Genetics*. 2016 aug;48(10):1284–1287. Available from:
810 <http://www.nature.com/doi/10.1038/ng.3656>.
- 811 49. Heath AP, Greenway M, Powell R, Spring J, Suarez R, Hanley D, et al. Bionimbus: a cloud for managing,
812 analyzing and sharing large genomics datasets. *Journal of the American Medical Informatics Association : JAMIA*.
813 2014 Nov;21(6):969–975. Available from: [https://academic.oup.com/jamia/article-](https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-002155)
814 [lookup/doi/10.1136/amiajnl-2013-002155](https://academic.oup.com/jamia/article-lookup/doi/10.1136/amiajnl-2013-002155).
- 815

816 Figure Captions

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

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

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

831

832 **Figure 2. MetaXcan Framework description and application.**

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

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

839

840

841 **Figure 3. A) ClinVar genes show significant S-PrediXcan associations.** Genes implicated in ClinVar tended to be
842 more significant in S-PrediXcan for most diseases tested, except for schizophrenia and autism. Blue circles
843 correspond to the qq-plot of genes in ClinVar that were annotated with the phenotype and black circles
844 correspond to all genes. **B) S-PrediXcan association for PCSK9, SORT1, and C4A.** R^2_{pred} is a performance
845 measure computed as the correlation squared between observed and predicted expression, cross validated in
846 the training set. Darker points indicate larger genetic component and consequently more active regulation in
847 the tissue. The size of the points represent the significance of the association between predicted expression
848 and the traits indicated on the top labels. C4A associations with schizophrenia (SCZ) are found across all tissues.
849 SORT1 associations with LDL-C, coronary artery disease (CAD), and myocardial infarction (MI) are most
850 significant in liver. PCSK9 associations with LDL-C, coronary artery disease (CAD), and myocardial infarction (MI)
851 are most significant in tibial nerve.

852 Tissue abbreviation: Adipose - Subcutaneous (ADPSBQ), Adipose - Visceral (Omentum) (ADPVSC), Adrenal Gland (ADRNLG), Artery - Aorta
853 (ARTAORT), Artery - Coronary (ARTCRN), Artery - Tibial (ARTTBL), Bladder (BLDDER), Brain - Amygdala (BRNAMY), Brain - Anterior cingulate
854 cortex (BA24) (BRNACC), Brain - Caudate (basal ganglia) (BRNCDT), Brain - Cerebellar Hemisphere (BRNCHB), Brain - Cerebellum (BRNCHA),
855 Brain - Cortex (BRNCTXA), Brain - Frontal Cortex (BA9) (BRNCTXB), Brain - Hippocampus (BRNHPP), Brain - Hypothalamus (BRNHPT), Brain -
856 Nucleus accumbens (basal ganglia) (BRNNCC), Brain - Putamen (basal ganglia) (BRNPTM), Brain - Spinal cord (cervical c-1) (BRNSPC), Brain -
857 Substantia nigra (BRNSNG), Breast - Mammary Tissue (BREAST), Cells - EBV-transformed lymphocytes (LCL), Cells - Transformed fibroblasts
858 (FIBRBLS), Cervix - Ectocervix (CVXECT), Cervix - Endocervix (CVSEND), Colon - Sigmoid (CLNSGM), Colon - Transverse (CLNTRN), Esophagus
859 - Gastroesophageal Junction (ESPGEJ), Esophagus - Mucosa (ESPMCS), Esophagus - Muscularis (ESPMSL), Fallopian Tube (FLLPNT), Heart -
860 Atrial Appendage (HRTAA), Heart - Left Ventricle (HRTLTV), Kidney - Cortex (KDNCTX), Liver (LIVER), Lung (LUNG), Minor Salivary Gland
861 (SLVRYG), Muscle - Skeletal (MSCLSK), Nerve - Tibial (NERVET), Ovary (OVARY), Pancreas (PNCREAS), Pituitary (PTTARY), Prostate (PRSTTE),
862 Skin - Not Sun Exposed (Suprapubic) (SKINNS), Skin - Sun Exposed (Lower leg) (SKINS), Small Intestine - Terminal Ileum (SNTRM), Spleen
863 (SPLEEN), Stomach (STMACH), Testis (TESTIS), Thyroid (THYROID), Uterus (UTERUS), Vagina (VAGINA), Whole Blood (WHLBLD).

864

865 **Figure 4. Comparison Summary-PrediXcan with Summary-TWAS, and SMR.**

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

868 Multiple SNPs are linked to the expression level of a gene via weights W_{X,T_g} .

869 **Panel B)** shows the significance of Summary-TWAS (BSLMM) vs. Summary-PrediXcan (elastic net). There is a
870 small bias caused by using S-TWAS results available from [26], which only lists significant hits.

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

873 **Panels D)** and **E)** show that the SMR statistics significance is bounded by GWAS and eQTL p-values. The p-values
874 (-log10) of the SMR statistics are plotted against the GWAS p-value of the top eQTL SNP (panel **D)**, and the
875 gene's top eQTL p-value (panel **E)**.

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

878

879 **Figure 5. Colocalization status of S-PrediXcan results.**

880 **Panel A)** shows a triangle that contains the probabilities of all five COLOC configurations. This ternary plot
881 constrains the values such that the sum of the probabilities is 1. All points in a horizontal line have the same
882 probability of "colocalized" GWAS and eQTL signals (P4), points on a line parallel to the right side of the triangle
883 (NW to SE) have the same probability of "Independent signals" (P3), and lines parallel to the left side of the
884 triangle (NE to SW) correspond to constant $P1+P2+P3$. Top sub-triangle corresponds to high probability of
885 colocalization ($P4>0.5$), lower left sub-triangle corresponds to probability of independent signals ($P3>0.5$), and
886 lower right parallelogram corresponds to genes without enough power to determine or reject colocalization.

887 The following panels present scatter plots of COLOC probabilities with a density overlay for S-PrediXcan results
888 of the Height phenotype.

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

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

894 **Panel D)** shows that HEIDI significant genes (to be interpreted as high heterogeneity between GWAS and eQTL
895 signals) mostly cluster in the "independent signal" region, in concordance with COLOC. A few genes fall in the
896 "colocalized" region, in disagreement with COLOC classification. Unlike COLOC results, HEIDI does not partition
897 the genes into distinct clusters and an arbitrary cutoff p-value has to be chosen.

898 **Panel E)** shows genes with large HEIDI p-value (no evidence of heterogeneity) which fall in large part in the
899 "colocalized" region but also substantial number fall in "independent signal" region, contradicting COLOC's
900 classification.

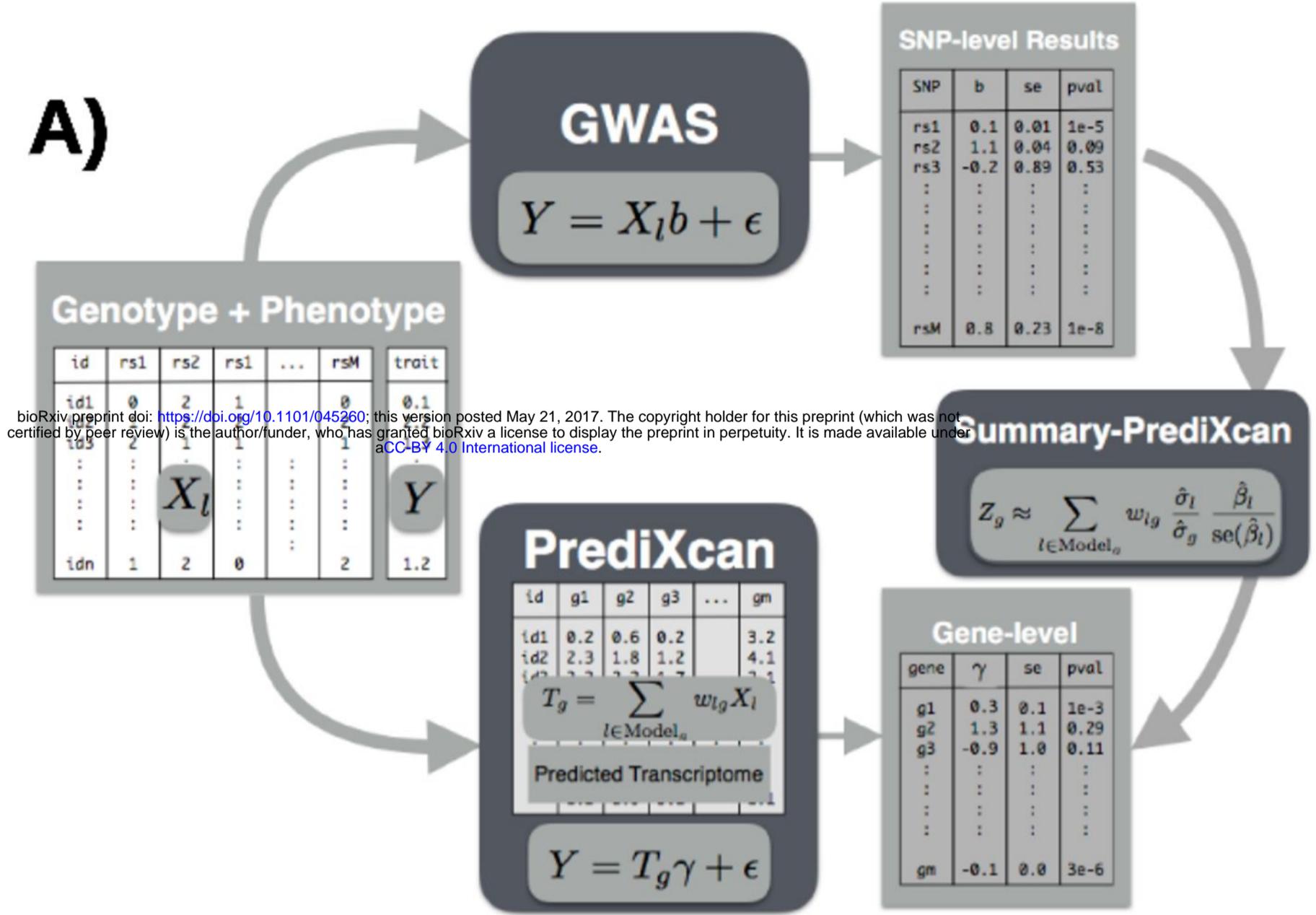
901

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

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

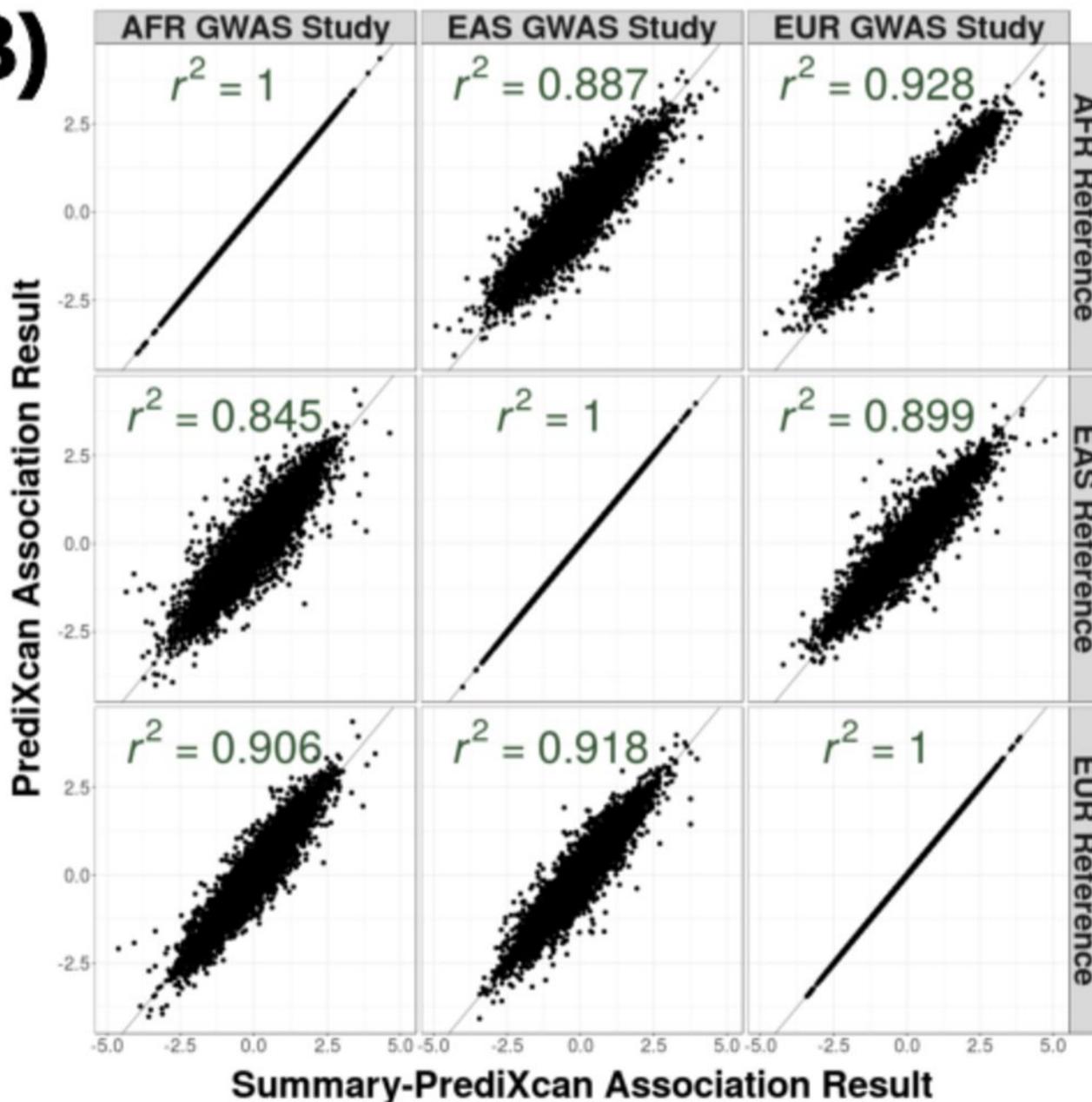
911

A)

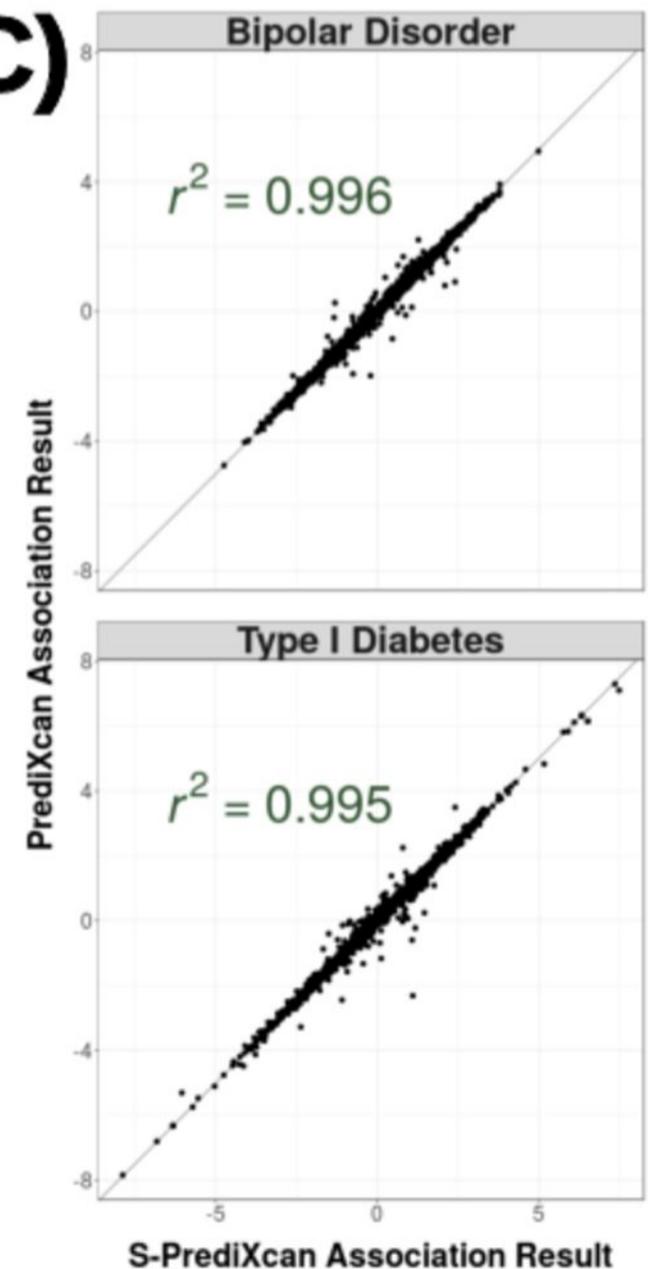


bioRxiv preprint doi: <https://doi.org/10.1101/045260>; this version posted May 21, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

B)



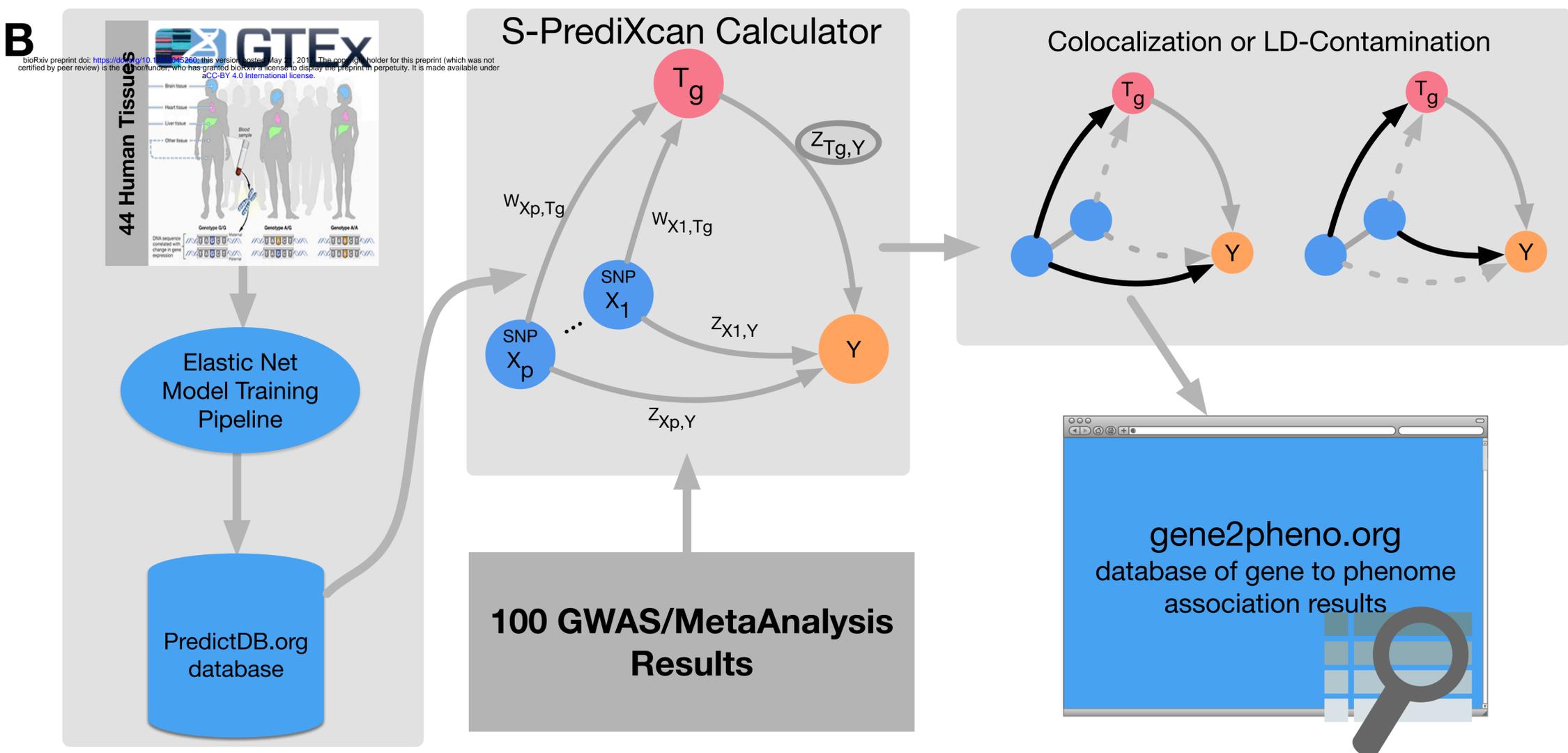
C)



A

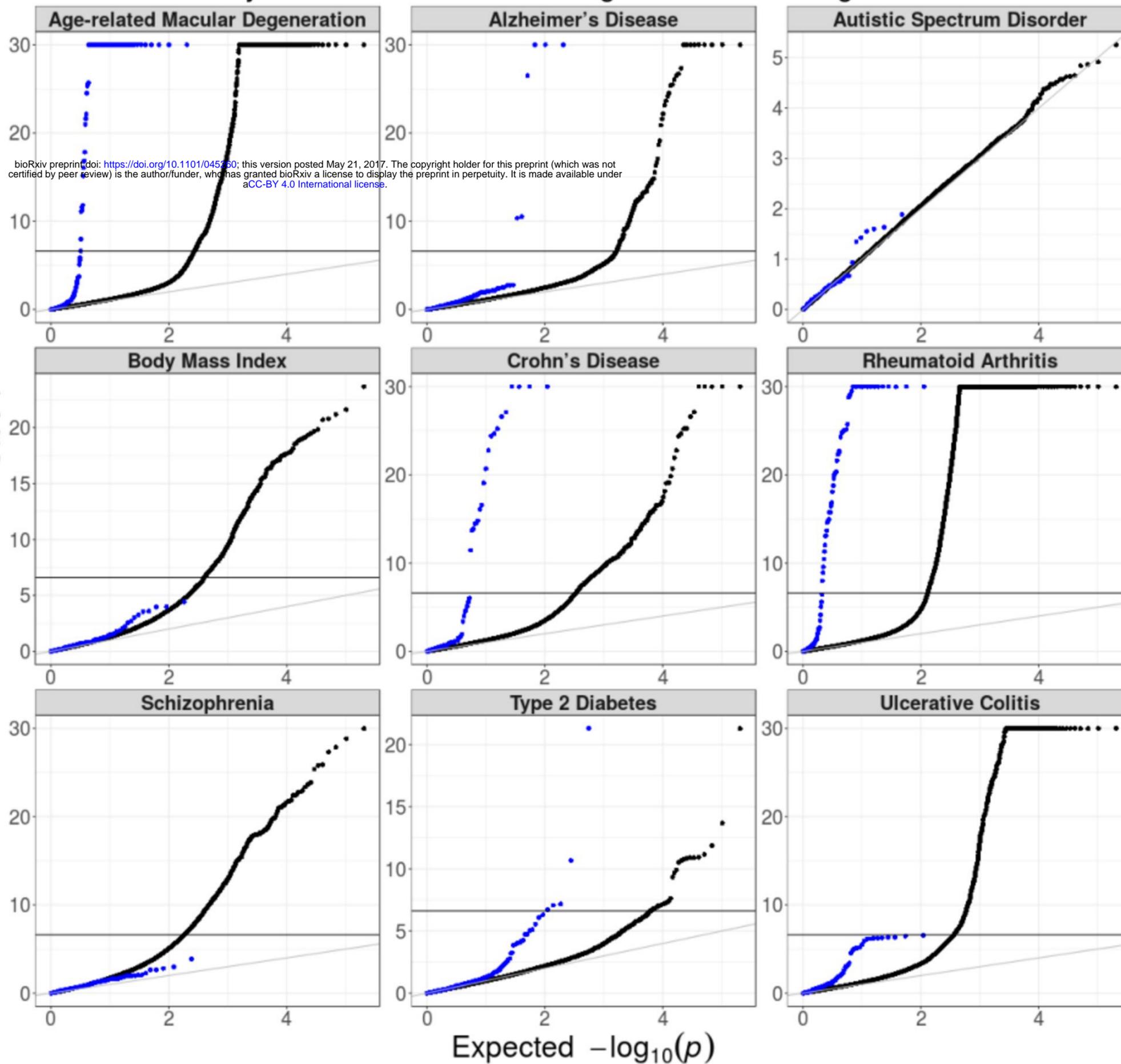
	Train Prediction Models	Filter-out Unreliable Models	Perform Gene-level Association	Adjust for Model Uncertainty	Filter-out Non-Colocalized Signals
PrediXcan	Elastic-Net	Correlation Observed vs Predicted FDR<0.05	Individual Level Data $Y \sim \text{Predicted } X$		Colocalization status (COLOC)
S-PrediXcan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

B



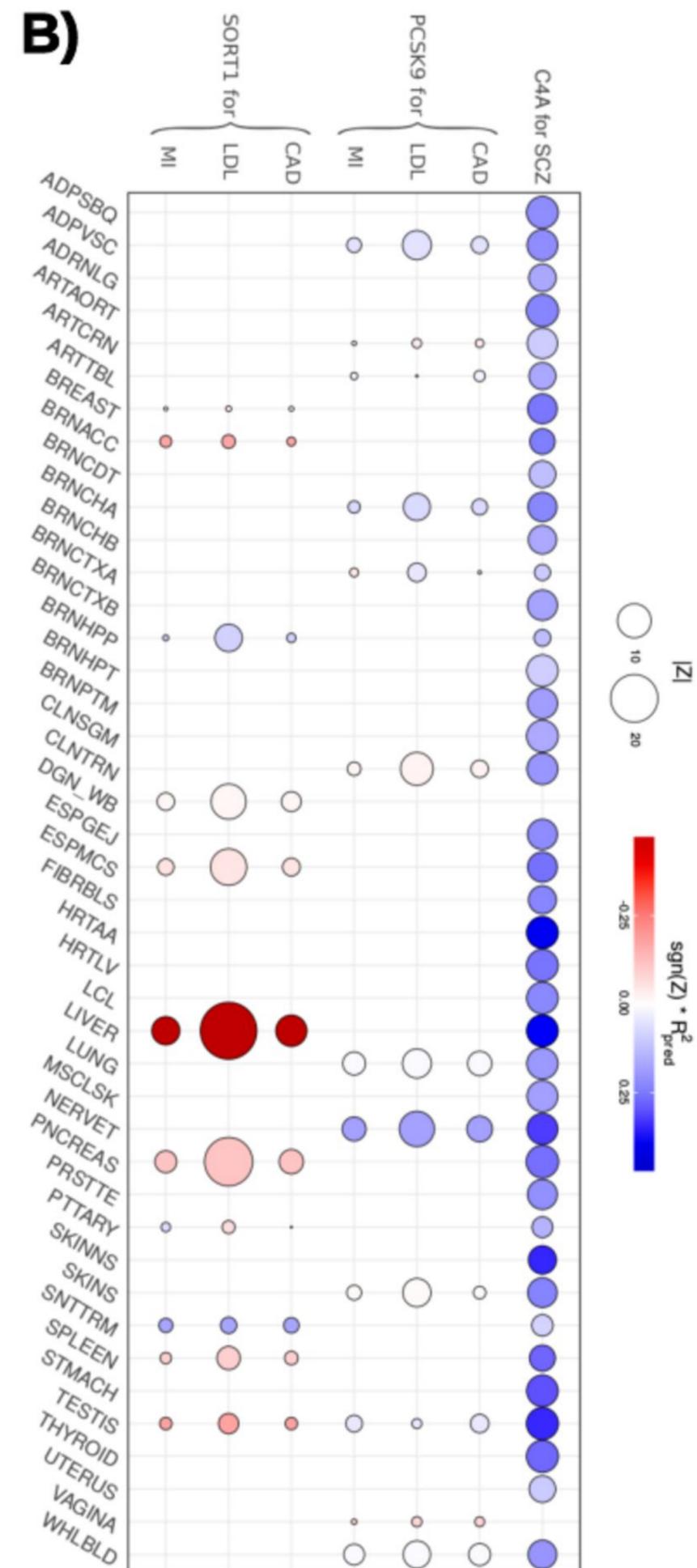
A)

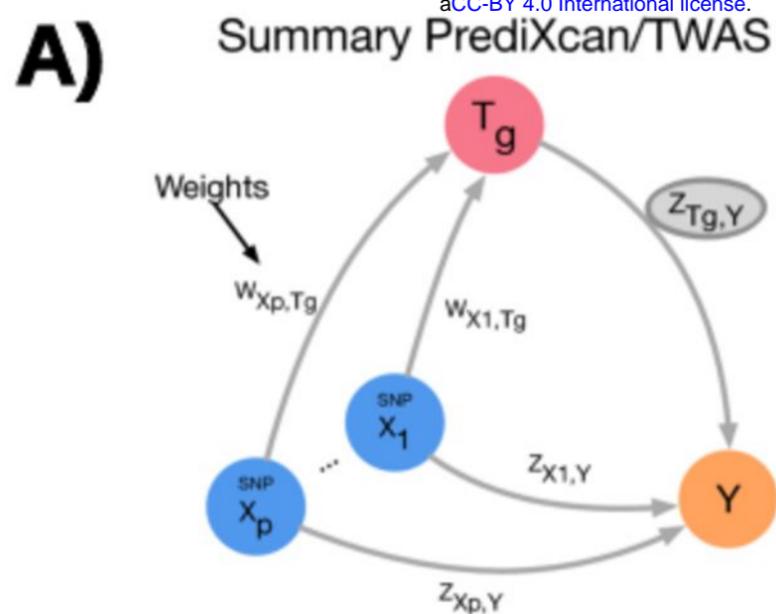
QQ Uniform plot: Summary PrediXcan results across all genes vs results for genes in ClinVar



Colors : ● All Genes, ● Clinvar,

B)

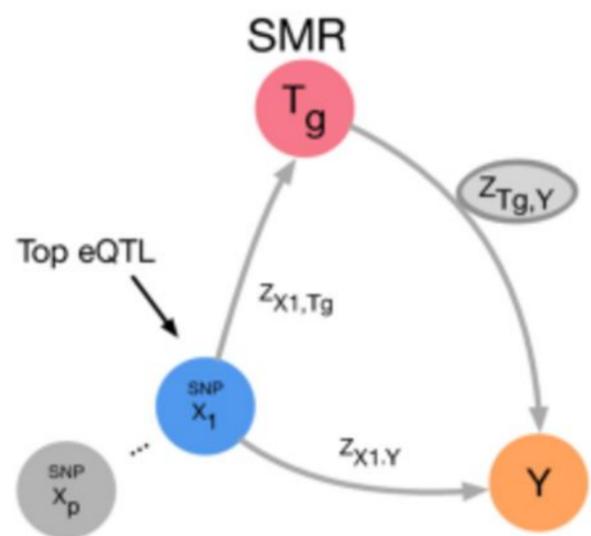




$$\hat{Z}_{T_g, Y}^{\text{s-predixcan}} = \sum_{l \in \text{Model}_{T_g}} w_{X_l, T_g} \frac{\hat{\sigma}_{X_l}}{\hat{\sigma}_{T_g}} Z_{X_l, Y} \sqrt{\frac{1 - R_{X_l}^2}{1 - R_{T_g}^2}}$$

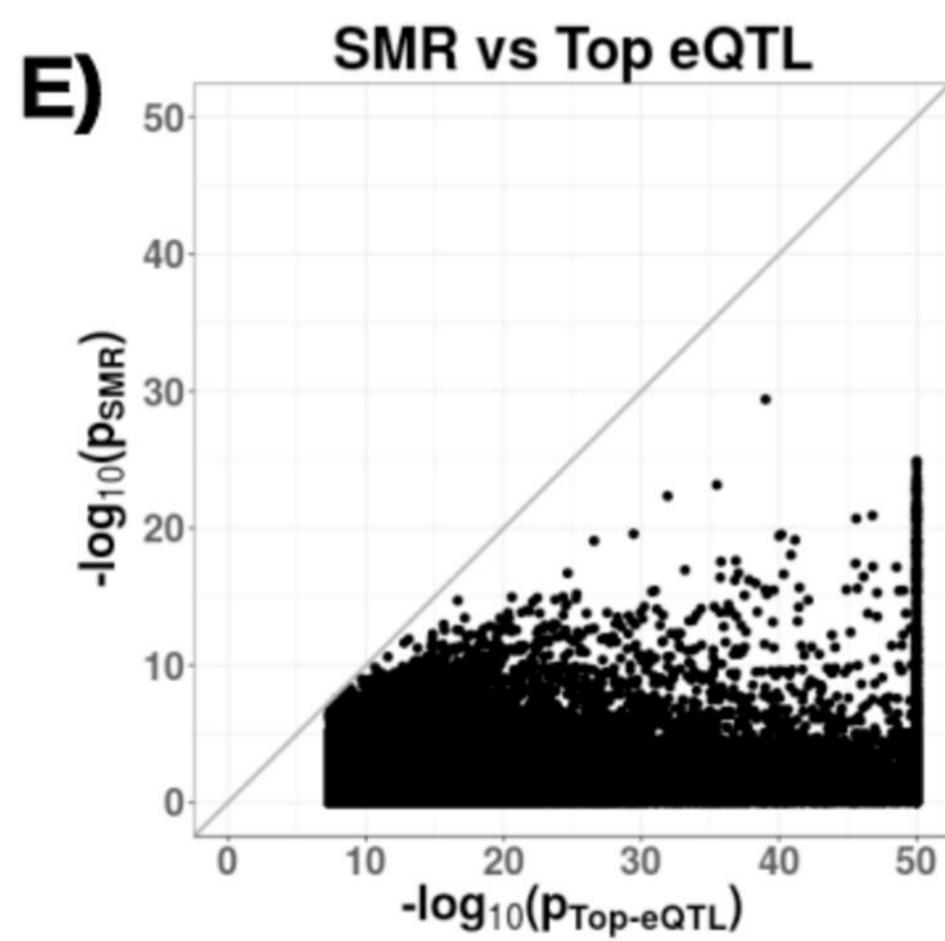
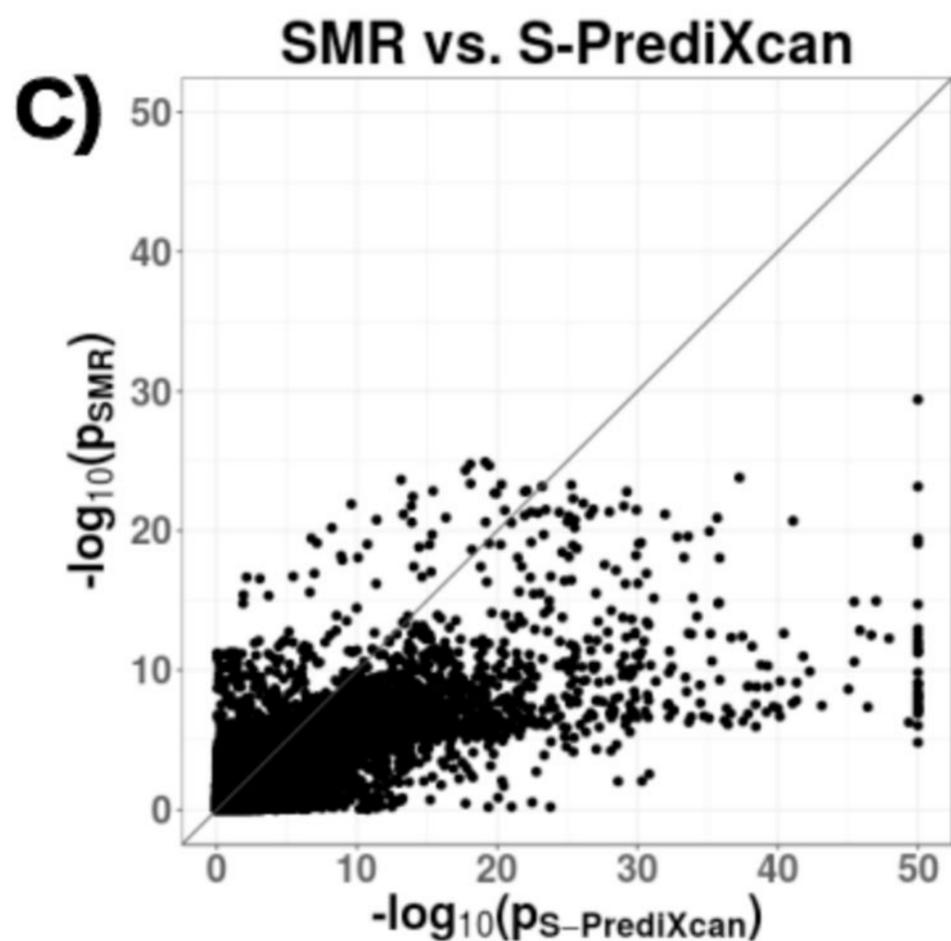
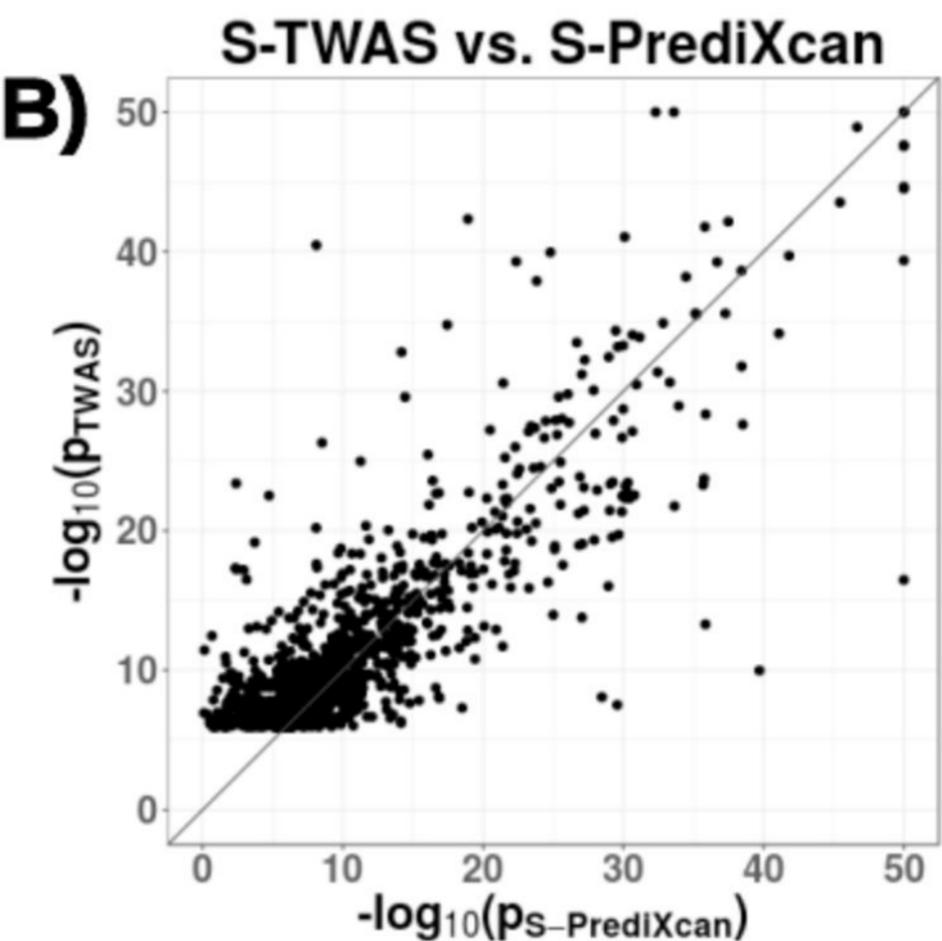
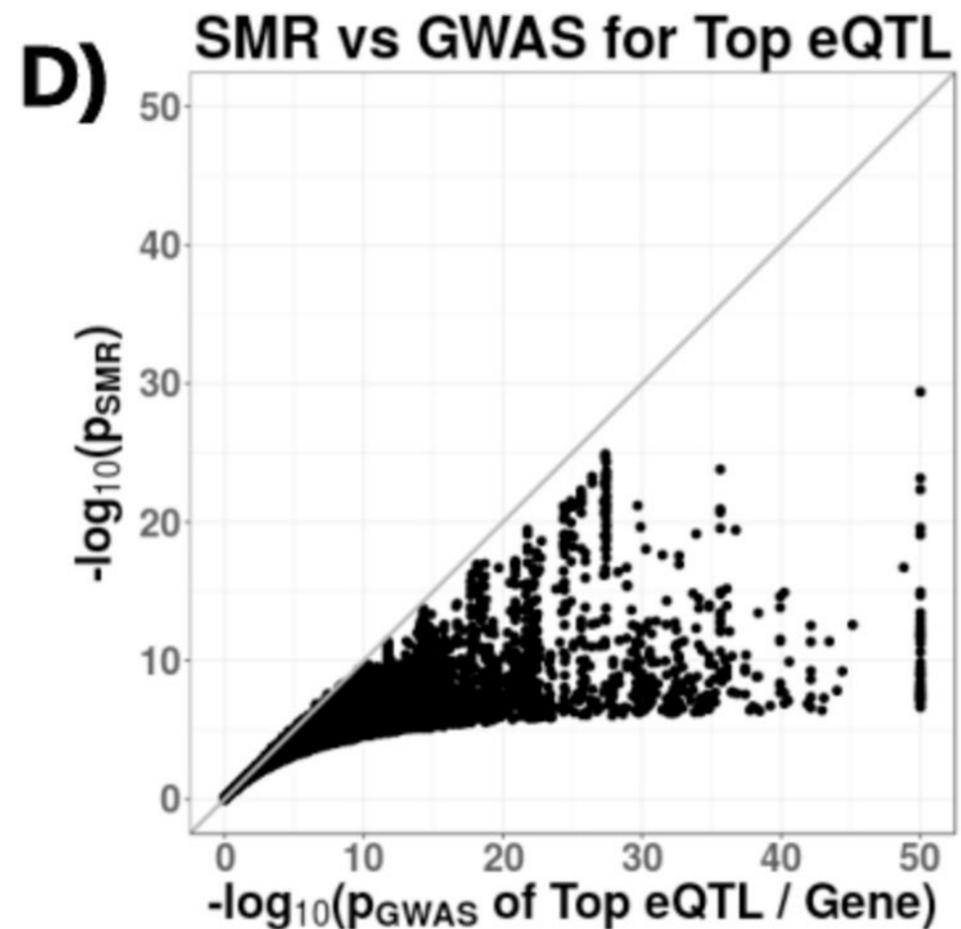
$$\approx \sum_{l \in \text{Model}_{T_g}} w_{X_l, T_g} \frac{\hat{\sigma}_{X_l}}{\hat{\sigma}_{T_g}} Z_{X_l, Y}$$

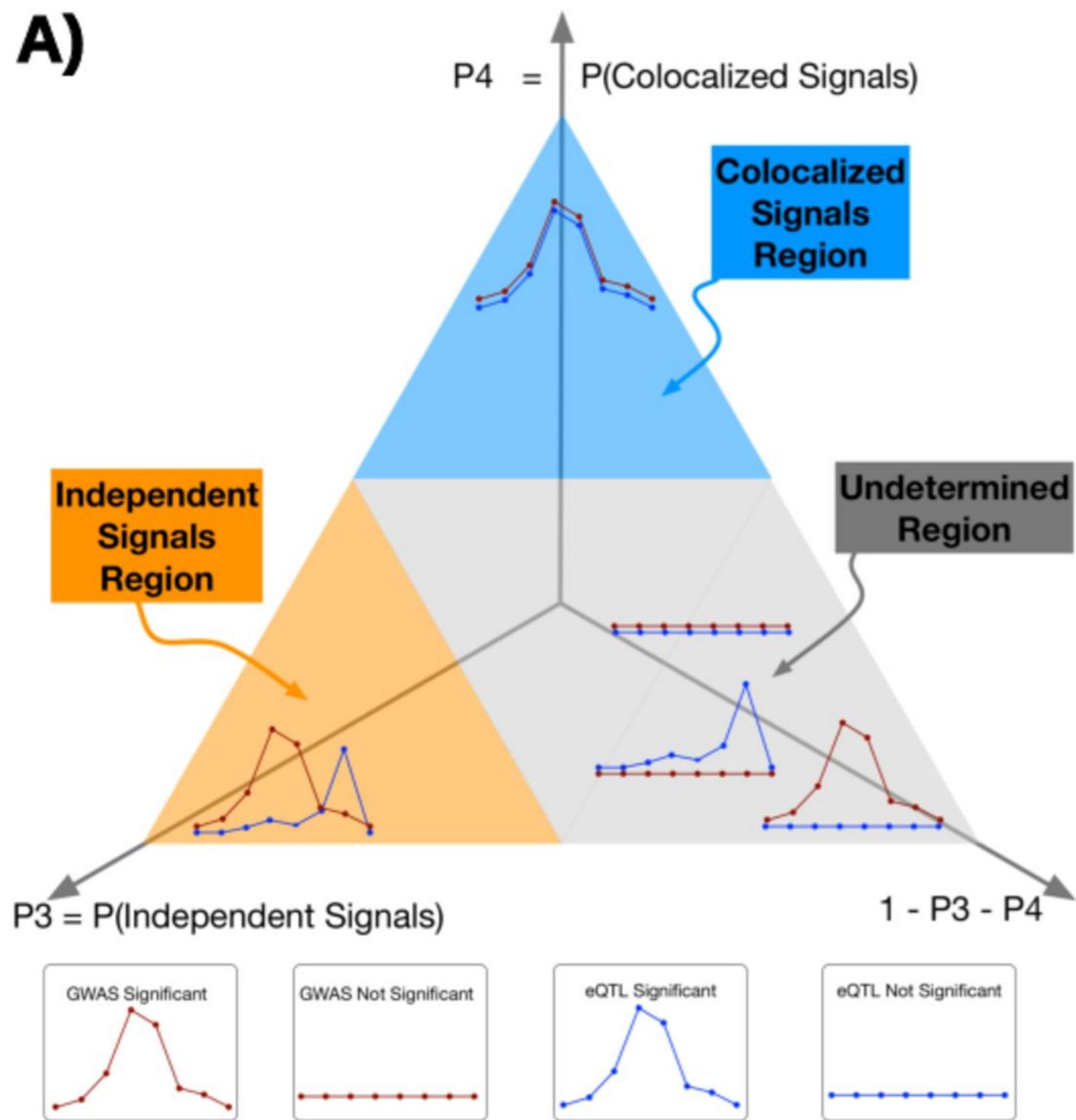
$$\approx \frac{W' Z_{X, Y}}{W' \Sigma_{X, X} W} = \hat{Z}_{X, Y}^{\text{s-twaspredixcan}}$$



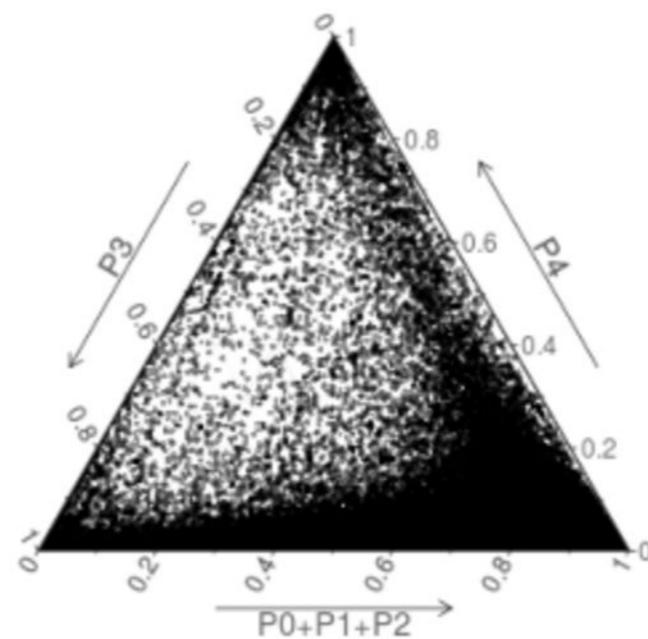
$$\frac{1}{\hat{Z}_{T_g, Y}^{2, \text{smr}}} = \frac{1}{Z_{X_{\text{top.eqtl}}, T_g}^2} + \frac{1}{Z_{X_{\text{top.eqtl}}, Y}^2}$$

$$= \frac{1}{Z_{\text{gwas}}^2} + \frac{1}{Z_{\text{eqtl}}^2}$$

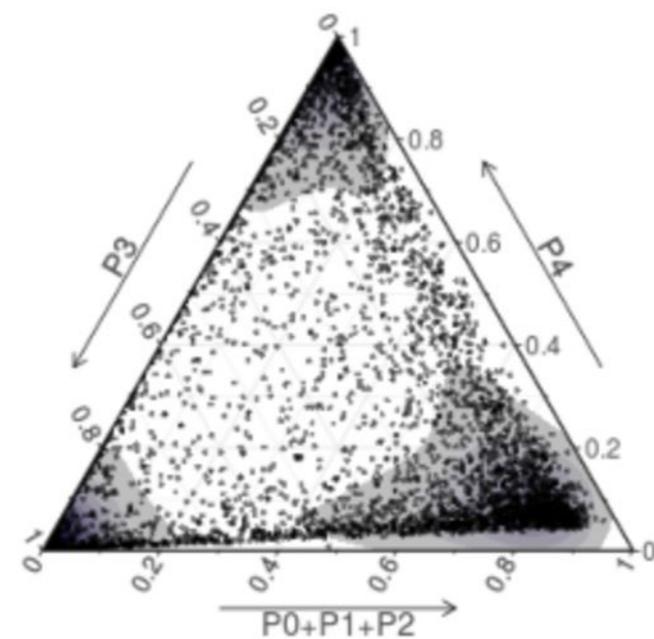




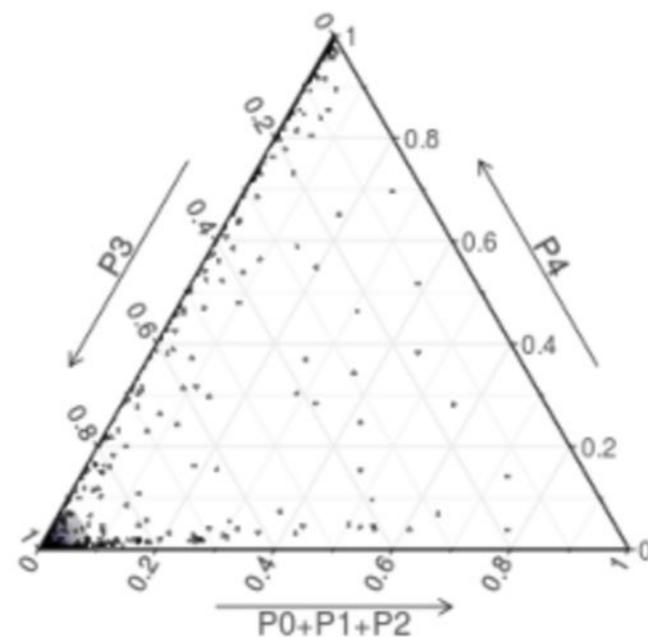
B) Summary-PrediXcan, All Results



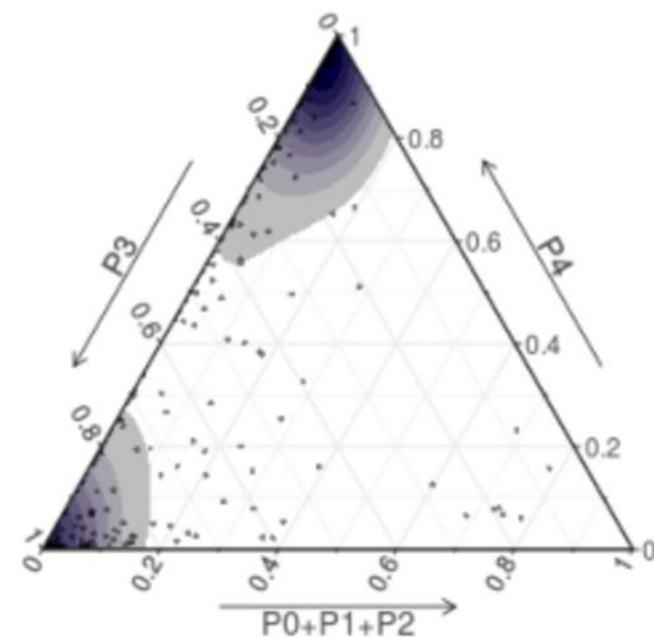
C) Summary-PrediXcan, $p_{\text{val}} < 1e-6$



D) Summary-PrediXcan, $p_{\text{val}} < 1e-6$, $p_{\text{HEIDI}} < 0.05$



E) Summary-PrediXcan, $p_{\text{val}} < 1e-6$, $p_{\text{HEIDI}} > 0.37$



GWAS Summary
Results
Study Set



$$Z_g =$$

$$\sum_{l \in \text{Model}_g}$$

$$w_{lg}$$

$$\frac{\sigma_l}{\hat{\sigma}_g}$$

$$\frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)}$$

~~$$\sqrt{\frac{1 - R_l^2}{1 - R_g^2}}$$~~



Weights from
PredictDB
Training Set



Reference Set:
1000G or
Training set