

1 **A multi-tissue transcriptome analysis of human metabolites guides the**  
2 **interpretability of associations based on multi-SNP models for gene expression**

3 Anne Ndungu,<sup>1,4</sup> Anthony Payne,<sup>1,4</sup> Jason Torres,<sup>1,4</sup> Martijn van de Bunt,<sup>1, 2, 3,5</sup> Mark I. McCarthy,<sup>1, 2, 5,6</sup>

4

5 **Author Affiliations**

6 <sup>1</sup>The Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford,  
7 Oxford, OX3 7BN, UK.

8 <sup>2</sup>Oxford Centre for Diabetes, Endocrinology and Metabolism, Radcliffe Department of Medicine,  
9 University of Oxford, Oxford, OX3 7LE, UK.

10 <sup>3</sup>Department of Bioinformatics and Data Mining, Novo Nordisk A/S, Måløv, 2760, DK.

11

12 **Author Notes**

13 <sup>4</sup>These authors contributed equally to this work.

14 <sup>5</sup>These authors jointly supervised this work.

15 Present addresses: <sup>6</sup>OMNI Human Genetics, Genentech, 1 DNA Way, South San Francisco, CA 94080,  
16 USA

17 \* Corresponding author: [mccarthy.mark@gene.com](mailto:mccarthy.mark@gene.com)

18

19

20

21 **Abstract**

22

23 There is particular interest in transcriptome-wide association studies (TWAS) - gene-level tests based on  
24 multi-SNP predictive models of gene expression - for identifying causal genes at loci associated with  
25 complex traits. However, interpretation of TWAS associations may be complicated by divergent effects  
26 of model SNPs on trait phenotype and gene expression. We developed an iterative modelling scheme for  
27 obtaining multi-SNP models of gene expression and applied this framework to generate expression  
28 models for 43 human tissues from the Genotype-Tissues Expression (GTEx) Project. We characterized  
29 the performance of single- and multi-SNP TWAS models for identifying causal genes in GWAS data for  
30 46 circulating metabolites. We show that: (a) multi-SNP models captured more variation in expression  
31 than the top *cis*-eQTL (median 2 fold improvement); (b) predicted expression based on multi-SNP models  
32 was associated (FDR<0.01) with metabolite levels for 826 unique gene-metabolite pairs, but, after step-  
33 wise conditional analyses, 90% were dominated by a single eQTL SNP; (c) amongst the 35% of  
34 associations where a SNP in the expression model was a significant *cis*-eQTL and metabolomic-QTL  
35 (met-QTL), 92% demonstrated colocalization between these signals, but interpretation was often  
36 complicated by incomplete overlap of QTLs in multi-SNP models; (d) using a “truth” set of causal genes  
37 at 61 met-QTLs, the sensitivity was high (67%), but the positive predictive value was low, as only 8% of  
38 TWAS associations at met-QTLs involved true causal genes. These results guide the interpretation of  
39 TWAS and highlight the need for corroborative data to provide confident assignment of causality.

40

41

## 42 **Introduction**

43

44 Genome wide association studies (GWAS) have been a powerful tool in revealing many loci that  
45 influence complex traits and diseases. However, most SNP associations map to non-coding regions of the  
46 genome, thereby complicating the task of identifying the (causal) genes through which the observed  
47 effects on disease predisposition are mediated<sup>1</sup>. To address this challenge, researchers have implemented  
48 a variety of approaches to link regulatory variants implicated in disease predisposition to their  
49 downstream effectors. One of the most widely adopted approaches leverages expression quantitative trait  
50 loci (eQTLs) to identify regional genes that are under the direct regulatory influence of the disease risk  
51 variant(s), and which thereby represent candidate mediators of disease predisposition. Empirical support  
52 for this approach is provided by the enrichment of *cis*-eQTL regulatory variants among significant  
53 GWAS variants and evidence that such variants explain a disproportionate share of trait heritability<sup>2-6</sup>.

54

55 A range of approaches have been deployed to detect coincident *cis*-eQTL and trait association signals.  
56 The simplest involves limiting the search space to trait variants that also demonstrate significant eQTL  
57 signals in a disease relevant tissue. In such analyses, it is now widely accepted that it is essential to test  
58 for statistical evidence of colocalization between eQTLs and trait-associated SNPs to avoid assigning  
59 relationships between eQTL and trait signals that map to distinct causal variants, and which cannot  
60 therefore be assumed to have any biological connection<sup>7,8</sup>.

61

62 Recently, this approach has been supplemented by a suite of methods (collectively, transcriptome-wide  
63 association studies or TWAS), built around a Mendelian randomization framework, which test for  
64 relationships between the genetic components of both complex traits and gene expression<sup>5,9-13</sup>. For  
65 example, the PrediXcan method generates predictive models of transcript expression from eQTL mapping  
66 data, and then uses these to “impute” estimates of gene expression into case-control or cohort-based

67 GWAS datasets: those imputed estimates can then be subjected to trait association testing<sup>12</sup>. Although  
68 PrediXcan requires individual-level genotype data as input, there are conceptually similar approaches  
69 available that can accept GWAS summary statistics with linkage disequilibrium (LD) estimates from a  
70 reference population (e.g. S-PrediXcan, Fusion)<sup>5,13,14</sup>. Collectively, these methods have been applied to a  
71 broad range of complex traits and diseases and have spotlighted novel and biologically plausible  
72 candidate genes that had evaded detection in conventional GWAS approaches<sup>5,11-13</sup>.

73  
74 The prediction models generated by these approaches range from those that feature only the single best  
75 (i.e. most strongly associated) eQTL for each gene, to those that support a polygenic model which  
76 comprises all SNPs within a locus (e.g. best linear unbiased predictor; BLUP). However, it has been  
77 shown that more sparse multivariate linear models (such as those generated by LASSO regression or a  
78 Bayesian Sparse Linear Mixed Model (BSLMM)) outperform both single variant and polygenic models in  
79 predicting gene expression<sup>5,11-13,15</sup>. Unlike single variant models, these sparse multi-SNP models can  
80 capture the effects of allelic heterogeneity (i.e. genes whose transcription is under the influence of  
81 multiple *cis*-regulatory signals). They also better reflect current understanding of the genetic architecture  
82 of gene expression than do polygenic models<sup>5,12,15</sup>.

83  
84 The fact that multi-SNP models better predict gene expression than single-SNP models might suggest that  
85 trait associations based on these models would themselves involve multiple SNPs with shared effects on  
86 both expression and phenotype. However, the extent to which this is true is unknown. Moreover, if such  
87 models better reflect the number of independent genetic signals acting on a phenotype, are they supported  
88 by evidence of shared identity between the trait-associated and eQTL variants within the model?  
89 Furthermore, the extent to which novel genes implicated by colocalized associations represent genuine  
90 biological relationships, causal for disease, is unclear and inference is further complicated by the shared  
91 regulatory architecture of gene expression and by horizontal pleiotropy<sup>16,17</sup>.

92

93 To address these outstanding questions and guide the interpretability of predicted gene expression studies,  
94 we systematically evaluated sparse multi-SNP models underlying significant gene associations for  
95 evidence of independent effects on both phenotype and expression. We did this by generating multi-SNP  
96 gene expression models for 43 human tissues from the GTEx project and evaluating their utility through a  
97 large-scale analysis of GWAS data for 46 metabolites. We focused on metabolomic phenotypes as they  
98 provide a singular opportunity to assess the biological plausibility of multi-SNP gene associations.  
99 Insights from both genetic and experimental studies have led to well-curated sets of effector genes at loci  
100 with *cis*-associations to the levels of particular metabolites<sup>18-21</sup>. The subsets of genes so implicated encode  
101 enzymes, transporters, and regulators that can be directly tied to the specific metabolite, based on known  
102 functional relationships. These provide a “truth” gene set that can then be used to assess the performance  
103 (i.e. sensitivity and positive predictive value) of alternative analytical approaches for identifying effector  
104 transcripts, and which can inform the utility of applying TWAS approaches to the interpretation of  
105 GWAS data for other complex traits.

106

## 107 **Material and Methods**

108

### 109 **GTEx expression data and *Cis*-eQTL analysis**

110 Genotype data (variant call format), gene expression (quantified gene-level counts), and sample  
111 phenotype data from GTEx version 7 were obtained through dbGaP accession phs000424.v7.p2<sup>22</sup>.  
112 Genotypes were filtered to keep only bi-allelic variants with minor allele frequency of at least 0.05.  
113 Finally, only remaining SNPs that were tested in all metabolite GWAS were used for analyses to ensure  
114 consistent downstream modelling and testing across metabolites.  
115 Only non-sex-specific tissue types with sample size of  $n \geq 50$  were analysed. For each tissue, genes  
116 reaching a threshold of  $> 6$  raw reads and  $> 1$  count per million in at least 10 individuals were carried

117 forward for analysis. Remaining genes were TMM normalised, then log transformed to counts per million  
118 using Voom<sup>23</sup>. Surrogate variables were calculated after explicitly defining sex in the models, and  
119 residual expression values after regressing out all surrogate variables and sex were used for analyses<sup>24</sup>.  
120 Cis-eQTLs analysis was performed using QTLtools (Version 1.1) with a *cis*-distance limit of 1,000,000  
121 base pair (1 Mb) from each gene<sup>25</sup>. The top eQTL SNP per gene was defined as the SNP with the lowest  
122 p-value for that gene.

123

#### 124 **GWAS summary data**

125 GWAS summary data for 46 metabolites were downloaded from the Metabolomics GWAS Server<sup>20,26</sup>.  
126 Metabolites for this analysis were selected based on having GWAS significant loci where at least one  
127 gene was identified as having a plausible or established biochemical link to the associated metabolite.  
128 Unknown metabolites and metabolite ratios were excluded from this analysis.

129

#### 130 **LASSO regression, model filtering and final model selection**

131 LASSO regression was used to select an optimal set of SNPs for explaining the expression of each gene.  
132 Regression was performed using GLMNET in R on each gene, with all SNPs less than 1MB from any  
133 part of each gene as potential covariates<sup>27</sup>. To select the optimal penalty factor for each gene, mean  
134 squared error (MSE) was calculated using 10-fold cross-validation across 100 automatically selected  
135 potential penalty factors. Given that data partitioning is random for cross-validation, this process was  
136 repeated 200 times per gene, and the penalty factor that had the mean lowest MSE across all iterations  
137 was selected as recommended in the reference manual for GLMNET.

138 For genes with multiple SNPs selected by LASSO regression, all selected SNPs were first linearly  
139 modelled against the gene's expression. For any groups of SNPs in perfect LD, one was randomly  
140 selected and retained. Model  $R^2$  was calculated for the full linear model. Iteratively, starting with the SNP  
141 with the lowest p-value in the model, SNPs were added back one-at-a-time, each time calculating the

142 subset model's  $R^2$  (i.e. forward regression). Once 95% of the full model's  $R^2$  value was attained; any  
143 SNPs not in the current subset model were eliminated. The final subset of SNPs was then modelled  
144 against expression and smoothed using ridge regression to minimize overfitting; with penalty factors  
145 selected using 25 iterations of 10-fold cross-validated ridge regression. For genes with only one SNP  
146 selected by LASSO, this SNP alone was modelled against gene expression using 25 iterations of 10-fold  
147 cross-validated ridge regression. The final coefficients from ridge regression models were carried forward  
148 for use in S-PrediXcan. Model fit p-values were determined by modelling pre-validated predicted  
149 expression of each gene against the observed expression. Model fit p-values were study wide FDR  
150 corrected (all genes and all tissues simultaneously), and those with adjusted  $p \geq 0.01$  were excluded from  
151 further analysis due to poor model fit.

152

### 153 **Transcriptome wide association analysis with S-PrediXcan**

154 For each modelled gene, S-PrediXcan (version 0.5.4) was used to calculate a z-score, which is a linear  
155 model of SNP effects for all SNPs in the gene's final ridge regression model described above<sup>14</sup>. Each  
156 SNP's effect is the product of its expression association coefficient, its GWAS z-score, and a SNP  
157 variance term (the SNP's standard error divided by the standard error of the gene's predicted expression).  
158 The SNP expression association coefficients used were those resulting from the final filtered gene  
159 expression ridge regression models. GWAS z-scores were calculated manually from effect size and  
160 standard error, since some SNPs had published p-values of 0 due to rounding.

161

### 162 **Conditional analysis**

163 For significant genes identified by S-PrediXcan, we decomposed the z-scores into per-SNP scores. For  
164 each significant gene, for SNPs from the S-PrediXcan model that had the same individual direction of  
165 effect as the overall S-PrediXcan z-score, the SNP that had the highest absolute S-PrediXcan magnitude  
166 was considered the top contributing SNP for conditional analysis. Conditional analysis was performed on

167 each significant S-PrediXcan gene using GCTA (version 1. 26.0)<sup>28</sup>. Each lead SNP effect was conditioned  
168 out of the GWAS summary data. S-PrediXcan was then performed as previously described, excluding the  
169 SNP/s being conditioned on, and using the GWAS z-scores resulting from the conditional GWAS  
170 analysis.

171

## 172 **Establishing biological plausibility of novel genes**

173 Annotated protein information was downloaded from the Human Metabolome Database (version 3.6) on  
174 December 11, 2017<sup>29</sup>. HUGO gene names, metabolism pathways, and gene ontology classifications listed  
175 in this database were referenced to assess membership of significant S-PrediXcan associated genes.  
176 Metabolic pathways and GO classifications annotated to novel genes were compared with those for  
177 putative causal genes associated to the same metabolites to assess shared metabolic processes.

178

## 179 **Results**

180

### 181 **Multi-SNP models explain more variance in gene expression than single eQTL models**

182 To investigate gene associations based on multi-SNP models, we first evaluated the extent to which these  
183 models improve prediction of gene expression relative to single variant models. We obtained single  
184 variant models by performing standard univariate eQTL analysis to identify the top associated *cis*-SNP  
185 for each gene in each of 43 tissues from the GTEx study (version 7) with a sample size exceeding 50  
186 (Methods)<sup>22</sup>. The number of expressed genes (defined as genes with >6 raw reads and >1 count per  
187 million in at least 10 individuals), ranged from 15,483 in EBV-transformed lymphocytes to 19,846 in  
188 lung.

189

190 To obtain multi-SNP genetic prediction models of gene expression, we employed LASSO regression - a  
191 multivariate penalized regression procedure that simultaneously performs feature selection along with



192 model fitting<sup>27</sup> - to select an optimal and sparse set of *cis*-SNPs to jointly model expression of each gene  
193 in each tissue. We then compared the variation in gene expression explained by these multi-SNP models  
194 to that accounted for by the single eQTL models.

195

196 In **Figure 1**, we show representative results, in this case for skeletal muscle, the tissue with the largest  
197 sample size (n=385). LASSO regression selected multiple SNPs in the models for the majority of genes  
198 (n=11,210), and for these genes, there was a median of 2.4-fold increase (interquartile range or IQR, 1.7  
199 to 3.9 fold) in expression variation explained by LASSO models versus the top eQTL alone (**Figure 1A**,  
200 **B**). There was a 2.0-fold median increase in expression variation explained across all gene models (i.e.  
201 including single-eQTL models) in skeletal muscle. LASSO selected the intercept-only model (i.e. model  
202 without any SNPs) for 2,667 genes out of 15,780 expressed genes, and the top-eQTL-only model (or a  
203 perfect proxy SNP) for 1,903 genes in skeletal muscle. The impact of multi-SNP selection seen for  
204 skeletal muscle was typical of that seen across all tissues and all genes (**Table S1**).

205

206 Despite the sparse nature of LASSO selection, for those genes with at least two modelled variants, 7,406  
207 genes (66%) in skeletal muscle retained at least one pair of SNPs with LD  $r^2 > 0.8$ . Moreover, LASSO  
208 expression models contained up to 159 SNPs and a median of 9 SNPs (IQR, 4 to 18 SNPs) for models  
209 with  $>1$  SNP. Since correlated SNPs can result in invalid inference for summarised Mendelian  
210 randomisation (MR) analyses<sup>30</sup>, we performed additional filtering of SNPs based on LD and proportion  
211 of variation explained ( $R^2$ ), iteratively adding SNPs into the model until 95% of the full model's  $R^2$  was  
212 achieved. For groups of SNPs in perfect LD ( $r^2 = 1$ ), one SNP was randomly selected and retained  
213 (Methods). This reduced the median number of SNPs per gene in the model in skeletal muscle to six  
214 (IQR, 3 to 12 SNPs, **Table S1**). Moreover, 18% of gene models (2,015 out of 11,210 models that  
215 included multiple SNPs in the unfiltered analysis) contained only the top eQTL (or a perfect proxy). This  
216 further round of filtering had little impact on model performance; the mean reduction in model  $R^2$  was

217 only 1.6% (calculated as percentages of the full LASSO models'  $R^2$  values; **Figure 1C,D**). Similar to the  
218 unfiltered LASSO models in skeletal muscle, there was still a 2.0-fold median increase in expression  
219 variation explained across all gene models and across all tissues. Overall, these results demonstrate that  
220 multi-SNP models should generally be optimised and explained more variation in gene expression  
221 relative to the single top eQTL for the majority of genes across tissues.

222

### 223 **Transcriptome-wide association analysis of 46 metabolites across 43 tissues**

224 Given these estimates of the extent to which multi-SNP models enhance the prediction of gene  
225 expression, we next sought to assess their utility in understanding genetic variation associated with  
226 complex diseases and traits. Metabolites offer a singular opportunity for such analyses as recent GWAS  
227 have identified strongly associated loci that regulate metabolite levels (met-QTLs)<sup>18-21</sup>.

228 At some of these loci, extensive genetic and experimental evidence have identified nearby genes for  
229 which the biological evidence for a causal role in mediating the metabolomics association is  
230 overwhelming, providing a “truth” set for causal gene localization not available in most other trait GWAS  
231 settings.

232

233 We focused on 46 metabolites with publicly available GWAS data for which at least one gene mapped  
234 near a significant met-QTL signal with high confidence biochemical links to the associated metabolite  
235 (**Table S2**)<sup>20</sup>. We performed transcriptome-wide association analysis with S-PrediXcan<sup>14</sup> to test for  
236 associations between predicted gene expression across 43 tissues and these 46 metabolite levels. Analysis  
237 was restricted to filtered LASSO prediction models with a strict significant expression model fit (model  
238  $q < 0.01$ ;  $n = 568,185$  total gene models).

239

240 A total of 2,834 associations between predicted gene expression values and metabolite levels reached  
241 significance at study wide FDR  $< 0.01$ , corresponding to 826 unique gene-metabolite pairs (i.e. some pairs

242 were significantly associated in multiple tissues) (**Figure 2A**). The largest number of associations  
243 identified for any tissue was 100 (tibial nerve). There were only 66 associations arising from predictive  
244 models generated from liver expression data (8% of 826 unique associations), even though liver could be  
245 considered the most biologically relevant tissue for most of these metabolites. This is likely due to the  
246 relatively small sample size for liver in GTEx (153 samples compared to 361 in tibial nerve) (**Figure 2B**,  
247 **Table S3**).

248  
249 For these 826 unique gene-metabolite pairs, we next sought to understand the extent to which multiple  
250 independent SNPs selected by the model were contributing to these metabolite associations. To do this,  
251 we performed conditional analyses for each of the 2,593 (from the total of 2,834) significant S-PrediXcan  
252 associations where the gene model had more than one SNP. We conditioned the metabolite GWAS on the  
253 SNP with the greatest effect on each gene's S-PrediXcan score and re-ran the S-PrediXcan association  
254 test using the conditioned GWAS summary statistics. After correcting for the number of genes, tissues,  
255 and metabolites tested after conditional analysis ( $p\text{-value}_{\text{conditional}} \leq 1.93 \times 10^{-5}$ ), 2,320 of the 2,593  
256 associations (89.5%) were no longer significant. This proportion was similar if we instead analysed only  
257 the most significant tissue for each gene; 684 out of 758 gene-metabolite pairs (90.2%) were no longer  
258 significant ( $p\text{-value}_{\text{conditional}} \leq 6.61 \times 10^{-5}$ ). Thus, for over 90% of significant S-PrediXcan associations,  
259 evidence for mediation of metabolite levels was dominated by a single SNP within the multi-SNP  
260 prediction models. Of the 273 still significant associations, over half (148) involved genes within 1 Mb of  
261 the highly complex *ACADS* gene region, which features multiple independent met-QTLs significantly  
262 associated with butyrylcarnitine levels (**Figure 3, Table S4**).

263  
264 **Colocalization analysis of model SNPs reveals the distinct relationships between *cis*-eQTL and met-**  
265 **QTL signals**

266

267 It is possible that overlaps between GWAS met-QTLs and *cis*-eQTL variants in multi-SNP models could  
268 be due to chance, rather than representing true colocalization of causal signals. Consider, for example, a  
269 multi-SNP model with two SNPs where one SNP is a strong eQTL but weakly associated with metabolite  
270 levels, and the other SNP displays the converse arrangement: this configuration could still yield a  
271 significant association between gene expression and metabolite levels. We therefore questioned to what  
272 extent multi-SNP S-PrediXcan associations were driven by *cis*-eQTL and met-QTL signals that shared  
273 the same identity (i.e. the associations were attributable to SNPs that influence metabolite levels through  
274 their effects on gene expression).

275  
276 We addressed this by performing colocalization analysis using eCAVIAR to obtain colocalization  
277 posterior probability (CLPP) values as evidence of shared causal signals, benefiting from the fact that  
278 eCAVIAR allows for multiple causal variants within a locus<sup>8</sup>. To increase our power to detect genuine  
279 colocalisation, we restricted this analysis to those SNPs in the prediction models that were significant *cis*-  
280 eQTLs (per tissue FDR<0.01) and met-QTLs (p-value<=5.0x10<sup>-8</sup>).

281  
282 We found that, among the 2,834 significant S-PrediXcan associations, about 35% of associations (990 of  
283 2,834 total; 214 unique gene-metabolite pairs) contained at least one SNP in the prediction model that  
284 influenced both metabolite levels at genome-wide significance and expression levels at FDR<0.05. Of  
285 these, 907 associations (92% of 990 associations tested; 202 unique gene-metabolite pairs) had at least  
286 one significant *cis*-eQTL with a CLPP > 0.01, evidence of a shared causal signal between met-QTL and  
287 *cis*-eQTL, in at least one tissue<sup>8</sup> (**Table S5**). Therefore, for the SNPs that corresponded to gene models  
288 and that were amenable to colocalization analysis, there was strong evidence of shared eQTL and met-  
289 QTL signals.

290

291 We then analysed the context within which *cis*-eQTL SNPs in the multi-SNP models colocalized with  
292 met-QTLs. For the 907 associations with evidence of colocalization, we observed instances of a one-to-  
293 one overlap whereby the significant *cis*-eQTL in the multi-SNP model colocalized with the corresponding  
294 met-QTL. An example of this arrangement is displayed in **Figure 4A**. However, determining the evidence  
295 for or against colocalization of the met-QTL and *cis*-eQTLs was not always as simple, since many loci  
296 had a more complex topography. For example, expression of *SLC16A9* was significantly associated with  
297 carnitine levels in S-PrediXcan analyses in tibial nerve. Two significant *cis*-eQTLs with low LD  
298 ( $r^2=0.002$ ) were selected in the prediction model, but, as the locus plot shows, only one of these signals  
299 colocalized with the met-QTL (**Figure 4B**).

300  
301 In contrast, we observed significant TWAS associations where model SNPs had divergent effects on  
302 expression and metabolite levels and were thereby excluded from colocalization analysis (i.e. associations  
303 not included in the 907 associations with evidence of colocalized QTL signals). For example, the  
304 expression of *FNDC1* in skeletal muscle was significantly associated with circulating isobutyrylcarnitine  
305 levels. However, the met-QTL and *cis*-eQTL were clearly not colocalized even though the genetically  
306 predicted expression of *FNDC1* was significantly associated with metabolite levels. This is because the  
307 set of SNPs in the *FNDC1* prediction model includes both the SNP driving the strong met-QTL (which  
308 explains a small portion of the variance in *FNDC1* expression) and a strong *cis*-eQTL that is only weakly  
309 associated with metabolite levels (**Figure 4C**).

310

### 311 **Determining the sensitivity and positive predictive value of multi-SNP prediction models**

312

313 Across the 46 metabolite GWAS that we used as the substrate for our analyses, Shin et al. previously  
314 reported 61 SNP-metabolite associations at which the associated met-QTL SNP mapped near a gene that  
315 was highly likely to be causal for the association. This assessment was based on either experimental

316 validation or a strong biological plausibility that the encoded protein was involved in the synthesis or  
317 degradation of the metabolite concerned<sup>20</sup>. These 61 SNP-gene-metabolite groupings provide a “truth” set  
318 of causal genes that can be used to explore the performance of expression QTL based mapping strategies,  
319 information relevant to more common applications (e.g. in a disease GWAS) where the causal gene is  
320 typically not known with equivalent certainty.

321

322 Of these 61 gene-metabolite pairs in the “truth” set, we were able to detect 41 through significant S-  
323 PrediXcan associations in at least one GTEx tissue (**Table 1**), indicating a sensitivity for *cis*-eQTL  
324 validation of the causal gene of 67%. Thirty-three of these gene-metabolite assignments were supported  
325 in more than one tissue with the *GCDH*-glutaryl carnitine association being the most widely represented  
326 (detected in 38 tissues, **Table 1**). Only eight of the 41 were detected in liver, though this may in part  
327 reflect the relatively small sample size of liver in GTEx. We assessed the extent to which overlaps  
328 between eQTLs and GWAS at these truth set genes represented true colocalization of signals. Of these 41  
329 genes, 23 were amenable to colocalization analysis (i.e. at least one of the SNPs in the model was a  
330 significant *cis*-eQTL and a significant met-QTL) and all of these 23 genes showed evidence of  
331 colocalization, where at least one SNP in the multi-SNP model colocalised with the met-QTL in at least  
332 one tissue.

333

334 As described earlier, our genome-wide trawl for associations between metabolite levels and predicted  
335 expression levels across GTEx tissues had implicated 826 unique gene-metabolite pairs. Of these, more  
336 than half (514; 62%) involved genes that mapped within 1Mb of the 61 truth set genes (including the 41  
337 detected truth set genes). At only four of the truth set loci did these analyses identify the true causal gene  
338 only with no nearby bystander gene. This indicates that, at many of these loci, there are multiple  
339 “bystander” genes, other than the truth set genes, that are also being detected through predicted  
340 expression. From this analysis of TWAS associations at metabolite-associated loci, we estimate that the

341 positive predictive value (PPV; i.e. number of true positives divided by the sum of true and false  
342 positives) for detecting true positive associations is only 8% (41/514 gene-metabolite pairs).

343

344 There were 20 of the 61 gene-metabolite pairs in the truth set that did not yield significant S-PrediXcan  
345 associations in any tissue. However, for 15 of these, significant S-PrediXcan associations (from the set of  
346 514 gene-metabolite pairs described above) were seen for nearby bystander genes in at least one tissue,  
347 with eight of these showing significant bystander gene colocalization (**Table S4**). Taken together with the  
348 results for the 41 true positive signals, these analyses indicate substantial pleiotropy at the level of *cis*-  
349 eQTLs, with many met-QTL loci harbouring a substantial excess of “bystander” genes alongside the true  
350 causal gene (or at some loci, only “bystander” genes).

351

352 To illustrate these concepts, consider SNP rs8012, which is a significant met-QTL for glutarylcarnitine  
353 levels ( $p\text{-value}_{\text{GWAS}}=1.24\times 10^{-43}$ ), and maps 8kb from the *GCDH* gene that encodes the enzyme glutaryl-  
354 CoA dehydrogenase. This enzyme catalyses the conversion of glutaryl-CoA to crotonyl-CoA, making  
355 *GCDH* a highly plausible effector gene mediating the effects of rs8012 on glutarylcarnitine levels<sup>31</sup>. In  
356 GTEx, whilst rs8012 is a *cis*-eQTL for *GCDH* in 31 tissues, the same SNP is also associated with the  
357 expression of other nearby genes including *HOOK2*, *SYCE2*, *FARSA*, *AD000092.3* and *CALR*. For all  
358 these genes, the *cis*-eQTL and the met-QTL signal clearly colocalized in at least one tissue (**Figure S1**).  
359 In the absence of the strong biological prior favouring *GCDH* at this locus, there would be at least five  
360 other genes that could be equally plausible candidates.

361 We next asked whether there were any features of the 473 bystander genes that might allow them to be  
362 distinguished from truth set genes. We found that bystander genes did not differ with respect to the  
363 strength of association with the metabolite, distance to transcription start site, the effect sizes of the  
364 individual eQTLs included in the multi-SNP models, or the CLPP values for model SNPs (**Figure 5**).

365 However, we did find that causal genes tended to be significant in more tissues than bystander genes at  
366 the same locus (**Figure S2**).

367

368 In addition to the 61 SNP-metabolite pairs in the truth set, Shin et al. reported 18 SNP-metabolite pairs  
369 that reached genome-wide significance in their analysis, but for which it was not possible to assign a  
370 causal gene with high confidence, as none of the genes could be implicated based on known biology<sup>20</sup>. In  
371 this setting, the authors assigned each associated SNP to the nearest gene at the locus (**Table S2**). The  
372 results of our analyses for these 18 signals recapitulated those we saw for the 61 genes in the “truth set”.  
373 We could recover 10 of these “nearest gene” candidates (a sensitivity of 56%) using S-PrediXcan applied  
374 to GTEx (**Table S6**), of which seven colocalized in at least one tissue. However, a further 92 bystander  
375 associations at these loci were also significant.

376

377 We also used a complementary approach to quantify the performance of the predicted expression analysis  
378 for identifying novel, plausible gene candidates. We focused on the 312 gene-metabolite pairs that  
379 involved genes that did not map to known met-QTL regions and evaluated metabolite and gene  
380 annotations in the Human Metabolome Database (version 4.0)<sup>29</sup>. We found that 96 of these pairs -  
381 corresponding to 83 genes - involved genes annotated to metabolic pathways. These included two genes  
382 involved in uridine metabolism - *CDA* and *UPPI*. Notably, SNPs in at these two loci were sub-genome-  
383 wide significant in the GWAS but were implicated from our S-PrediXcan analysis and subsequent  
384 studies<sup>17</sup> (**Table S7**). We expanded the search further by querying a recently curated dataset<sup>17</sup> and found  
385 an additional 18 genes annotated to at least one metabolic pathway. Thus, as many as 37% (114/312 gene-  
386 metabolite pairs) of novel TWAS gene associations can be considered biologically plausible albeit based  
387 on the rather permissive overlap between “metabolic pathway” and met-QTL.

388



389 We then performed a more stringent evaluation by determining the number of novel gene-metabolite  
390 associations (again excluding “bystander” genes) where the novel gene either shared at least one  
391 metabolic pathway with a reported truth set gene for the associated metabolite or has been curated as a  
392 high confidence causal gene with the associated metabolite in recent publications<sup>17</sup>. We found that 16  
393 (5%) of the 312 novel gene-metabolite pairs met this criterion (**Table S8**). Taking this as a lower limit  
394 and the previous less stringent estimate as an upper limit, we estimate that 5-37% of novel gene-  
395 metabolite relationships are biologically plausible. Notably, this range encompasses our PPV estimate of  
396 8% obtained from evaluating the true positive rate at met-QTLs with known causal genes. Therefore,  
397 most novel gene associations based on multi-SNP models represented either false positives or “bystander”  
398 genes that are not biologically relevant *per se* but rather driven by variants with pleiotropic effects on  
399 gene expression. Overall, these findings emphasize that, whilst the multi-SNP *cis*-eQTL approach has  
400 respectable sensitivity in detecting the causal gene in these data, performance in terms of PPV is poor and  
401 additional lines of evidence will be needed at most loci to establish causality.

402

## 403 **Discussion**

404

405 In this study, we have assessed the utility of multi-SNP prediction models for explaining variation in gene  
406 expression and their application in transcriptome-wide association analysis (TWAS). We quantified the  
407 extent to which these models outperform expression models based on a single eQTL, demonstrating,  
408 across all evaluated tissues, a median 2-fold improvement in variance explained. When applied in a  
409 TWAS of genome-wide data for 46 metabolites across 43 human tissues, these multi-SNP models  
410 identified 826 significant gene-metabolite associations. By leveraging knowledge of genes highly likely  
411 to be causally involved in the regulation of metabolite levels, we were able to quantify the accuracy with  
412 which multi-SNP TWAS detects such high-confidence effectors. The results from these analyses offer  
413 several key insights relevant to the interpretation of TWAS results.

414

415 We found that, notwithstanding the use of LASSO regression as a sparse form of variable selection, it is  
416 still prone to select sets of SNPs that are highly correlated, introducing multicollinearity into resulting  
417 regression models. This notion has been described before in real and simulated GWAS data<sup>32</sup>. We showed  
418 that a simple iterative approach to LASSO modelling that involved LD-based filtering resulted in  
419 increased model sparsity and decreased multicollinearity, leading to more confident genetic instruments  
420 for gene expression.

421

422 Despite the improved performance in predicting gene expression attributable to models with multiple,  
423 independent SNPs, we found that, using available GTEx data, TWAS associations based on these models  
424 were, in most instances, driven by a single SNP within each trait-associated locus: 90% of associations  
425 were no longer significant after stepwise conditional analysis. Although this proportion is likely to fall as  
426 eQTL sample sizes increase (increasing the power to detect the additional impact of conditioned variants),  
427 these results indicate that, for many genes, the increment in power gained by moving from single to multi-  
428 SNP analyses is modest.

429

430 The genetic architecture underlying metabolite traits provides a unique opportunity to quantify the  
431 performance of gene associations based on multi-SNP models. By leveraging a “truth” set of  
432 experimentally validated genes linked to metabolites, we have shown, using GTEx, that TWAS has  
433 reasonable sensitivity (67%) at identifying causal genes. However, the PPV is low (8%), as a great  
434 majority of associations in the vicinity of a known causal gene involved nearby “bystander” genes.  
435 Furthermore, the process of resolving true causal from false-positive associations is complicated by the  
436 fact that these types of associations were indistinguishable in their model SNP effect sizes (GWAS and  
437 eQTL), colocalization probabilities, and distance to transcription start sites. In the case of the metabolite  
438 glutarylcarntine, for example, the met-QTL rs8012 not only regulates the expression of the causal *GCDH*

439 gene but also five other genes at the same locus, all of which are associated with glutaryl carnitine levels  
440 in TWAS. These insights temper the extent to which it can be assumed that genes implicated by  
441 significant TWAS associations are causal.

442  
443 These “bystander” effects reflect their shared regulatory architecture with known causal genes, and our  
444 observations around met-QTLs mirror recent findings at the *SORT1* and *NOD2* loci (associated with LDL  
445 cholesterol and Crohn’s disease, respectively)<sup>33</sup>. By anchoring our analysis on a wide range of  
446 metabolomic phenotypes, we have been able to extend those observations, and to develop more  
447 generalizable estimates of the sensitivity and PPV of TWAS. Recent analyses from Stacey and colleagues  
448 using an alternative gene prioritisation method (ProGeM) are also instructive<sup>17</sup>. Using ProGeM to address  
449 a similar problem (the detection of causal effector genes at met-QTL loci), the performance was  
450 appreciably better than that we observed with a sensitivity of 98%, and a specificity ranging from 38.4%  
451 to 84.6% (PPV was not measured, and the true negatives needed for estimation of specificity were  
452 derived using different criteria for delineating sets of candidate causal genes). However, in contrast to  
453 TWAS, ProGeM explicitly integrates SNP-level annotations (i.e. eQTLs) with functional gene and  
454 pathway annotations across five databases to prioritize causal genes. That is to say, ProGeM directly  
455 leverages molecular pathway annotations whereas TWAS is agnostic to this information. Accordingly,  
456 ProGeM is intended for a specific trait class - molecular QTLs (e.g. metabolites, lipids, proteins) - and the  
457 incorporation of additional information relevant to metabolites is likely to have contributed to the better  
458 performance in this specific task. In addition, the sensitivity of ProGeM may be inflated by the fact that  
459 shared database features were used to both prioritise genes and benchmark performance. For these  
460 reasons, ProGeM might be expected not to achieve comparable performance when used to prioritise  
461 effectors at disease GWAS loci, with performance more resembling that of the more agnostic approach  
462 we achieve with TWAS.

463

464 Our analyses were focused on the use of expression QTLs to map causal genes at metabolomic-QTL  
465 signals: the extent to which similar observations apply to other molecular QTLs remains to be  
466 determined. Previous studies have shown that the genetic architecture of protein-QTLs (pQTLs) is  
467 distinct from that of eQTLs: only half of pQTLs identified in lymphoblastoid cell lines (LCLs) were also  
468 eQTLs, and pQTL effect sizes were typically lower than those for eQTLs<sup>34</sup>. However, these apparently  
469 distinct architectures are likely in part the consequence of disparities in sample sizes and differences in  
470 the technologies used to profile these features. Further work is required to assess if the confounding effect  
471 of co-regulation observed in TWAS based on predicted gene expression will be present to the same extent  
472 for other molecular features.

473  
474 TWAS approaches provide an attractive option for prioritizing candidate genes at trait-associated loci.  
475 Here, we have demonstrated the potential for these approaches to identify associations that are not causal,  
476 through a combination of incomplete colocalization and pleiotropy in gene expression regulation.  
477 Ultimately, the process of identifying causal genes at GWAS signals represents an integrative enterprise  
478 that is dependent on combining results from multiple complementary approaches, including, in addition to  
479 QTL-mapping, epigenome profiling (e.g. chromatin co-accessibility or conformation capture methods),  
480 functional screens (e.g. high-throughput gene knock-out CRISPR screens) and the detection of coding  
481 variant associations. All of these prioritization approaches – including TWAS – will become more  
482 accurate, as the data sets available encompass a wider range of tissues and cell types captured in  
483 circumstances (e.g. developmental stages, physiological states, environmental exposures) that better  
484 reflect the underlying pathophysiology of the particular traits and diseases under investigation.

485

#### 486 **Supplemental Data**

487 Supplemental Figures. Figures S1 and S2

488 Supplemental Tables. Tables S1-S8

489

490 **Declaration of Interests**

491 MMcC has served on advisory panels for Pfizer, NovoNordisk, Zoe Global; has received honoraria from  
492 Merck, Pfizer, NovoNordisk and Eli Lilly; has stock options in Zoe Global and has received research  
493 funding from Abbvie, AstraZeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk,  
494 Pfizer, Roche, Sanofi Aventis, Servier & Takeda. As of June 2019, MMcC is an employee of Genentech,  
495 and holds stock in Roche. MvdB has been a full time employee of Novo Nordisk A/S since May 2017,  
496 and holds stock in Novo Nordisk.

497

498 **Acknowledgments**

499 MMcC is a Wellcome Investigator and an NIHR Senior Investigator. Relevant funding support for this  
500 work comes from Wellcome (090532, 106130, 098381, 203141, and 212259), NIDDK (U01-DK105535;  
501 U01-DK085545) and NIHR (NF-SI-0617-10090). AP was supported by the Rhodes Trust, the Natural  
502 Sciences and Engineering Research Council of Canada, and the Canadian Centennial Scholarship Fund.  
503 While employed at the University of Oxford, MvdB was supported by a Novo Nordisk postdoctoral  
504 fellowship run in partnership with the University of Oxford. This work was also supported by Oxford  
505 Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for  
506 Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford  
507 Biomedical Research Centre. The views expressed are those of the author and not necessarily those of the  
508 NHS, the NIHR or the Department of Health.

509

510 **Web Resources**

511 The pre-trained multi-SNP models across 43 GTEx (version 7) tissues are available at  
512 <http://mccarthy.well.ox.ac.uk/pub/>

513

514 **References**

- 515 1. Visscher, P.M., Brown, M.A., McCarthy, M.I., and Yang, J. (2012). Five years of GWAS discovery.  
516 *Am J Hum Genet* 90, 7-24.
- 517 2. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-Associated  
518 SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. *PLOS Genetics* 6,  
519 e1000888.
- 520 3. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S.,  
521 Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major  
522 determinant of human expression variation. *Nature* 482, 390-394.
- 523 4. Davis, L.K., Yu, D., Keenan, C.L., Gamazon, E.R., Konkashbaev, A.I., Derks, E.M., Neale, B.M.,  
524 Yang, J., Lee, S.H., Evans, P., et al. (2013). Partitioning the heritability of Tourette syndrome and  
525 obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet* 9, e1003864.
- 526 5. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W., Jansen, R., de Geus, E.J., Boomsma,  
527 D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association  
528 studies. *Nat Genet* 48, 245-252.
- 529 6. Torres, J.M., Gamazon, E.R., Parra, E.J., Below, J.E., Valladares-Salgado, A., Wachter, N., Cruz, M.,  
530 Hanis, C.L., and Cox, N.J. (2014). Cross-tissue and tissue-specific eQTLs: partitioning the heritability of  
531 a complex trait. *Am J Hum Genet* 95, 521-534.
- 532 7. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and  
533 Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using  
534 summary statistics. *PLoS Genet* 10, e1004383.
- 535 8. Hormozdiari, F., van de Bunt, M., Segre, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman,  
536 S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes.  
537 *Am J Hum Genet* 99, 1245-1260.

- 538 9. Smith, G.D., and Ebrahim, S. (2003). 'Mendelian randomization': can genetic epidemiology contribute  
539 to understanding environmental determinants of disease? *International journal of epidemiology* 32, 1-22.
- 540 10. Freeman, G., Cowling, B.J., and Schooling, C.M. (2013). Power and sample size calculations for  
541 Mendelian randomization studies using one genetic instrument. *International journal of epidemiology* 42,  
542 1157-1163.
- 543 11. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard,  
544 M.E., Wray, N.R., Visscher, P.M., et al. (2016). Integration of summary data from GWAS and eQTL  
545 studies predicts complex trait gene targets. *Nat Genet* 48, 481-487.
- 546 12. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J.,  
547 Eyster, A.E., Denny, J.C., Consortium, G.T., Nicolae, D.L., et al. (2015). A gene-based association method  
548 for mapping traits using reference transcriptome data. *Nat Genet* 47, 1091-1098.
- 549 13. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene  
550 Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits.  
551 *Am J Hum Genet* 100, 473-487.
- 552 14. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson,  
553 E.S., Shah, K.P., Garcia, T., Edwards, T.L., et al. (2018). Exploring the phenotypic consequences of  
554 tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun* 9, 1825.
- 555 15. Wheeler, H.E., Shah, K.P., Brenner, J., Garcia, T., Aquino-Michaels, K., Consortium, G.T., Cox, N.J.,  
556 Nicolae, D.L., and Im, H.K. (2016). Survey of the Heritability and Sparse Architecture of Gene  
557 Expression Traits across Human Tissues. *PLoS Genet* 12, e1006423.
- 558 16. Hemani, G., Bowden, J., and Davey Smith, G. (2018). Evaluating the potential role of pleiotropy in  
559 Mendelian randomization studies. *Hum Mol Genet* 27, R195-R208.
- 560 17. Stacey, D., Fauman, E.B., Ziemek, D., Sun, B.B., Harshfield, E.L., Wood, A.M., Butterworth, A.S.,  
561 Suhre, K., and Paul, D.S. (2019). ProGeM: a framework for the prioritization of candidate causal genes at  
562 molecular quantitative trait loci. *Nucleic Acids Res* 47, e3.

- 563 18. Kettunen, J., Tukiainen, T., Sarin, A.P., Ortega-Alonso, A., Tikkanen, E., Lyytikäinen, L.P., Kangas,  
564 A.J., Soinen, P., Wurtz, P., Silander, K., et al. (2012). Genome-wide association study identifies  
565 multiple loci influencing human serum metabolite levels. *Nat Genet* 44, 269-276.
- 566 19. Suhre, K., and Gieger, C. (2012). Genetic variation in metabolic phenotypes: study designs and  
567 applications. *Nat Rev Genet* 13, 759-769.
- 568 20. Shin, S.Y., Fauman, E.B., Petersen, A.K., Krumsiek, J., Santos, R., Huang, J., Arnold, M., Erte, I.,  
569 Forgetta, V., Yang, T.P., et al. (2014). An atlas of genetic influences on human blood metabolites. *Nat*  
570 *Genet* 46, 543-550.
- 571 21. Kastenmuller, G., Raffler, J., Gieger, C., and Suhre, K. (2015). Genetics of human metabolism: an  
572 update. *Hum Mol Genet* 24, R93-R101.
- 573 22. GTEx Consortium, Laboratory, D.A., Coordinating Center -Analysis Working, G., Statistical  
574 Methods groups-Analysis Working, G., Enhancing, G.g., Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh,  
575 Nih/Nida, et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204-213.
- 576 23. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma  
577 powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*  
578 43, e47.
- 579 24. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for  
580 removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28,  
581 882-883.
- 582 25. Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., and Dermitzakis, E.T. (2017). A  
583 complete tool set for molecular QTL discovery and analysis. *Nat Commun* 8, 15452.
- 584 26. Suhre, K., Shin, S.Y., Petersen, A.K., Mohny, R.P., Meredith, D., Wagele, B., Altmaier, E.,  
585 CardioGram, Deloukas, P., Erdmann, J., et al. (2011). Human metabolic individuality in biomedical and  
586 pharmaceutical research. *Nature* 477, 54-60.



- 587 27. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear  
588 Models via Coordinate Descent. *J Stat Softw* 33, 1-22.
- 589 28. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A.,  
590 Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion  
591 of the heritability for human height. *Nat Genet* 42, 565-569.
- 592 29. Wishart, D.S., Feunang, Y.D., Marcu, A., Guo, A.C., Liang, K., Vazquez-Fresno, R., Sajed, T.,  
593 Johnson, D., Li, C., Karu, N., et al. (2018). HMDB 4.0: the human metabolome database for 2018.  
594 *Nucleic Acids Res* 46, D608-D617.
- 595 30. Burgess, S., Butterworth, A., and Thompson, S.G. (2013). Mendelian randomization analysis with  
596 multiple genetic variants using summarized data. *Genet Epidemiol* 37, 658-665.
- 597 31. Lenich, A.C., and Goodman, S.I. (1986). The purification and characterization of glutaryl-coenzyme  
598 A dehydrogenase from porcine and human liver. *The Journal of biological chemistry* 261, 4090-4096.
- 599 32. Waldmann, P., Meszaros, G., Gredler, B., Fuerst, C., and Solkner, J. (2013). Evaluation of the lasso  
600 and the elastic net in genome-wide association studies. *Front Genet* 4, 270.
- 601 33. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D.,  
602 Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for  
603 transcriptome-wide association studies. *Nat Genet* 51, 592-599.
- 604 34. Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., and Snyder, M.  
605 (2013). Variation and genetic control of protein abundance in humans. *Nature* 499, 79-82.

606

## 607 **Figure Titles and Legends**

608

### 609 **Figure 1. Model $R^2$ comparison of LASSO regression models.**

610 **(A)** Scatterplot comparing variation in gene expression explained by the top eQTL alone and by the

611 multi-SNP LASSO model in skeletal muscle. **(B)** Violin plot showing the fold increase in gene expression

612 variation explained by LASSO models in skeletal muscle. The asterisks in the violin plots denote that the  
613 y-axis is abrogated at a fold change of 10. **(C)** Comparison of LASSO regression models before and after  
614 contribution based filtering of collinear SNPs. The mean model  $R^2$  was reduced by only 1.6%. **(D)** Violin  
615 plot showing the fold increase in gene expression variation explained by filtered LASSO models across  
616 all 43 tissues. The asterisks in the violin plots denote that the y-axis is abrogated at a fold change of 10.  
617

618 **Figure 2. Transcriptome-wide association analysis of 46 metabolites across 43 tissues.**

619 **(A)** Manhattan plot showing all S-PrediXcan associations across 46 metabolites in all 43 tissues analysed  
620 with each point representing a gene-metabolite association. Labels indicate loci where TWAS  
621 associations involve high confidence causal genes. **(B)** Bar plot of the number of significant gene-  
622 metabolite associations observed per tissue.

623

624 **Figure 3. Step-wise conditional analysis of significant associations.**

625 **(A)** Plot showing results from the conditional analysis of S-PrediXcan associations involving multi-SNP  
626 prediction models. The vertical line denotes the significance threshold used for conditional analysis. Only  
627 273 associations remained significant after conditioning on the lead met-QTL SNP, of which, 148  
628 mapped to the *ACADS* locus and influence butyrylcarnitine levels (yellow triangles). **(B)** Locus Zoom  
629 plot showing met-QTLs associating with butyrylcarnitine levels at the *ACADS* locus and their LD relative  
630 to the top met-QTL.

631

632 **Figure 4. Colocalization analysis of eQTL and met-QTL signals in multi-SNP models for**  
633 **metabolite-associated genes.**

634 (A) Colocalization of the single met-QTL and single *cis*-eQTL signal at the *FADS1* gene in esophagus  
 635 mucosa.(B) Partial colocalization at the *SLC16A9* gene in tibial nerve where only one of the two  
 636 independent *cis*-eQTLs in the multi-SNP model is colocalized with the met-QTL at this gene (C) No  
 637 colocalization of *cis*-eQTL and met-QTL for the *FNDC1* gene in skeletal muscle. The red triangles  
 638 denote the SNPs present in the genes' multi-SNP prediction models.

639

640 **Figure 5. Comparison of features of multi-SNP models for bystander genes to those for true causal**  
 641 **genes.**

642 (A) Comparison of the effect sizes of model SNPs for bystander genes and model SNPs for true causal  
 643 genes on metabolite levels in GWAS. (B) Distribution of effects on gene expression for individual SNPs  
 644 in models for bystander and known causal genes. (C) Comparison of the distance from TSS for model  
 645 SNPs in bystander and causal genes. (D) The distribution of colocalization posterior probabilities (CLPP)  
 646 scores for model SNPs in bystander and causal genes.

647

648 **Tables**

Metabolite ID	Metabolite Name	Causal Gene	Number of Associations	Most Significant Tissue	q-value
M35439	glutaryl carnitine	<i>GCDH</i>	38	Whole-Blood	1.88E-39
M01110	arachidonate (20:4n6)	<i>FADS1</i> <sup>a</sup>	27	Thyroid	1.09E-78
M32412	butyryl carnitine	<i>ACADS</i> <sup>a</sup>	26	Lung	5.64E-202
M01110	arachidonate (20:4n6)	<i>FADS2</i>	26	Esophagus-Gastroesophageal-Junction	1.76E-48
M37058	succinyl carnitine	<i>CRAT</i> <sup>a</sup>	23	Cells-EBV-Transformed-Lymphocytes	5.78E-13
M00606	uridine	<i>TYMP</i>	20	Cells-Transformed-Fibroblasts	1.36E-11
M35433	hydroxyisovaleroyl carnitine	<i>MCCC1</i>	16	Skin-Sun-Exposed	1.29E-09
M01604	urate	<i>SLC2A9</i>	14	Muscle-Skeletal	3.57E-34
M03141	betaine	<i>BHMT</i>	11	Brain-Frontal-Cortex	3.21E-12
M32338	glycine	<i>CPS1</i>	10	Brain-Putamen	3.15E-10
M32654	β-dehydrocarnitine	<i>SLC22A5</i> <sup>a</sup>	10	Skin-Not-Sun-Exposed	5.64E-14
M01123	inosine	<i>NT5E</i> <sup>a</sup>	8	Spleen	3.19E-09
M15500	carnitine	<i>SLC16A9</i>	8	Esophagus-Mucosa	1.28E-44
M15140	kynurenine	<i>SLC7A5</i>	8	Adipose-Visceral	1.58E-12
M22138	homocitrulline	<i>SLC7A9</i>	7	Colon-Transverse	0.000202

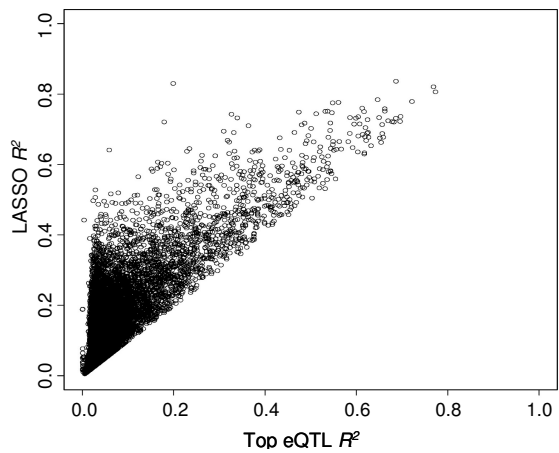
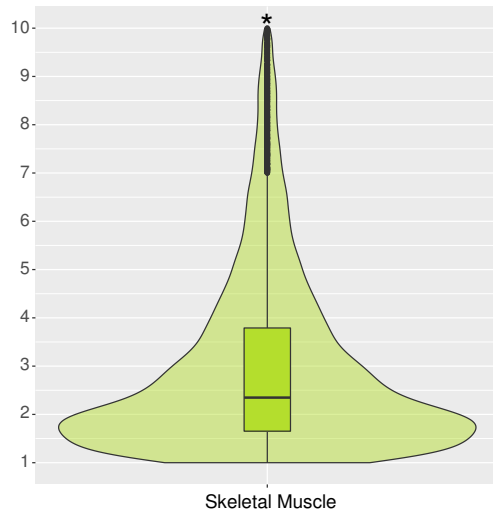
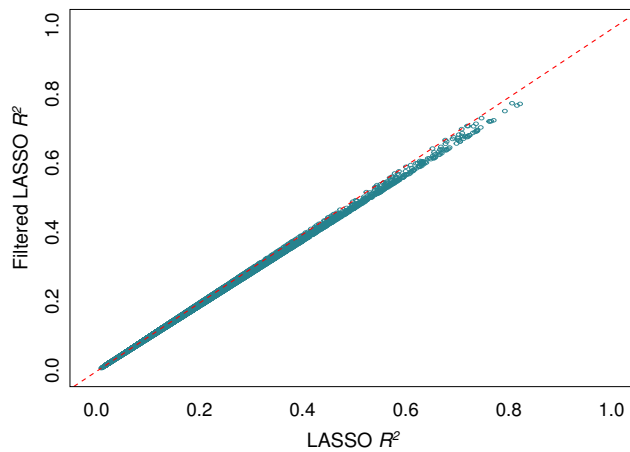
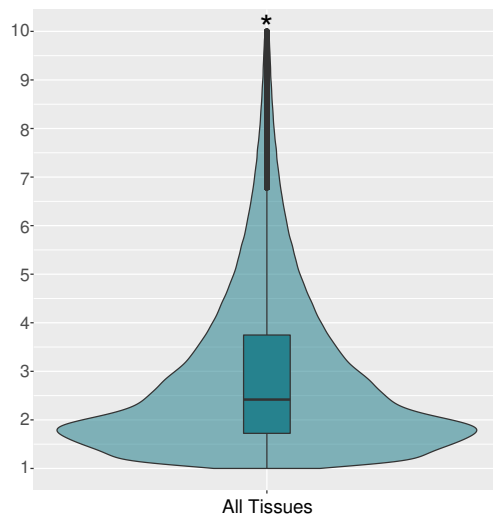
M01110	arachidonate (20:4n6)	<i>FADS3</i> <sup>a</sup>	6	Liver	9.96E-55
M35159	cysteine- glutathione disulfide	<i>GGT1</i>	6	Esophagus-Mucosa	1.88E-08
M35439	glutaryl carnitine	<i>SLC7A6</i>	6	Spleen	4.24E-14
M35439	glutaryl carnitine	<i>CPT2</i>	5	Colon-Sigmoid	8.48E-08
M01494	5-oxoproline	<i>OPLAH</i>	4	Skin-Sun-Exposed	5.08E-98
M02137	biliverdin	<i>UGT1A1</i> <sup>a</sup>	4	Skin-Not-Sun-Exposed	1.16E-49
M32315	serine	<i>PHGDH</i>	3	Colon-Sigmoid	2.73E-13
M33441	isobutyryl carnitine	<i>SLC22A1-2</i>	3	Skin-Not-Sun-Exposed	3.20E-05
M32654	β-dehydrocarnitine	<i>SLC22A4</i>	3	Skin-Sun-Exposed	1.06E-17
M15500	carnitine	<i>SLC22A4</i>	3	Artery-Tibial	2.01E-07
M15500	carnitine	<i>SLC22A5</i>	3	Brain-Cerebellum	0.00104
M37097	tryptophan betaine	<i>SLC22A5</i>	3	Brain-Putamen	2.35E-05
M18349	indolelactate	<i>CCBL1</i>	2	Brain-Cortex	0.000201
M03127	hypoxanthine	<i>GMPR</i>	2	Brain-Cerebellar-Hemisphere	0.00228
M22130	phenyllactate (PLA)	<i>GOT2</i>	2	Brain-Frontal-Cortex	1.05E-08
M35631	1-palmitoyl glycerophosphoethanolamine	<i>LIPC</i> <sup>a</sup>	2	Pancreas	3.96E-06
M03141	betaine	<i>SLC6A12</i>	2	Lung	0.00148
M02132	citrulline	<i>ALDH18A1</i>	1	Skin-Sun-Exposed	0.00807
M33937	alpha-hydroxyisovalerate	<i>HAO2</i>	1	Adrenal-Gland	1.53E-06
M32315	serine	<i>PSPH</i>	1	Esophagus-Mucosa	0.000534
M00054	tryptophan	<i>SLC16A10</i>	1	Brain-Frontal-Cortex	0.00671
M01299	tyrosine	<i>SLC16A10</i>	1	Brain-Frontal-Cortex	0.00058
M32412	butyryl carnitine	<i>SLC16A9</i>	1	Esophagus-Mucosa	0.000185
M32348	2-aminobutyrate	<i>SLC1A4</i>	1	Muscle-Skeletal	1.96E-12
M37097	tryptophan betaine	<i>SLC22A4</i>	1	Artery-Tibial	3.28E-07
M32379	scyllo-inositol	<i>SLC5A11</i>	1	Brain-Hippocampus	0.00474

649

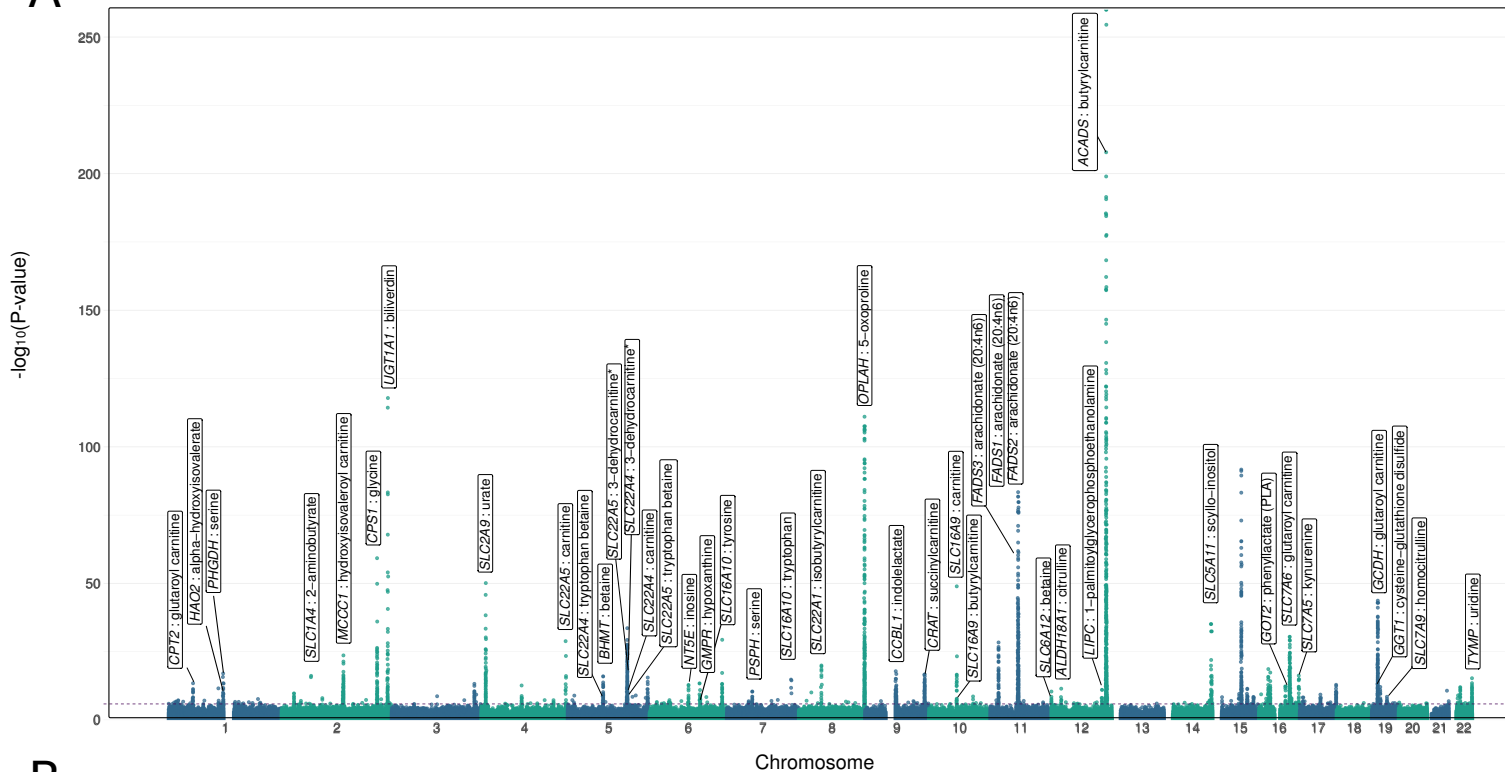
650 **Table 1. Causal genes from the truth set that significantly associated with metabolite levels in a**

651 **TWAS.** Of the 61 high confidence truth set genes, 41 had significant S-PrediXcan associations in at least

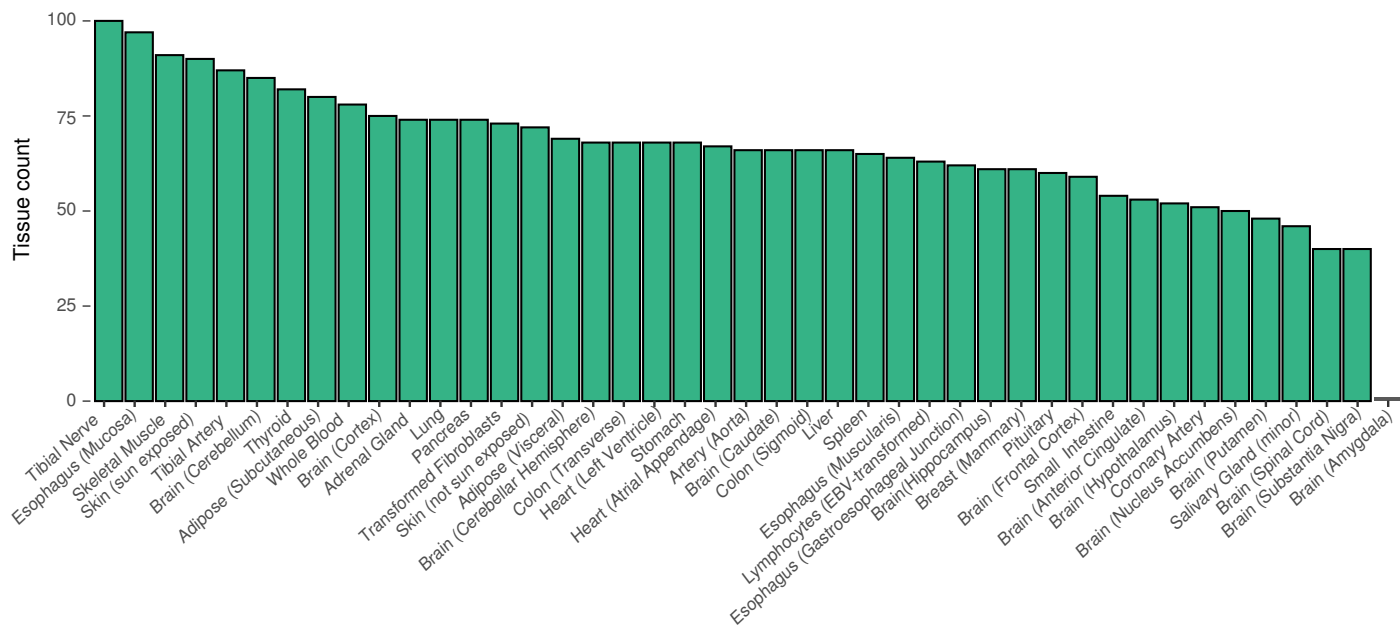
652 one tissue.<sup>a</sup> Eight gene-metabolite pairs that had a significant association in liver.

**A****B****C****D**

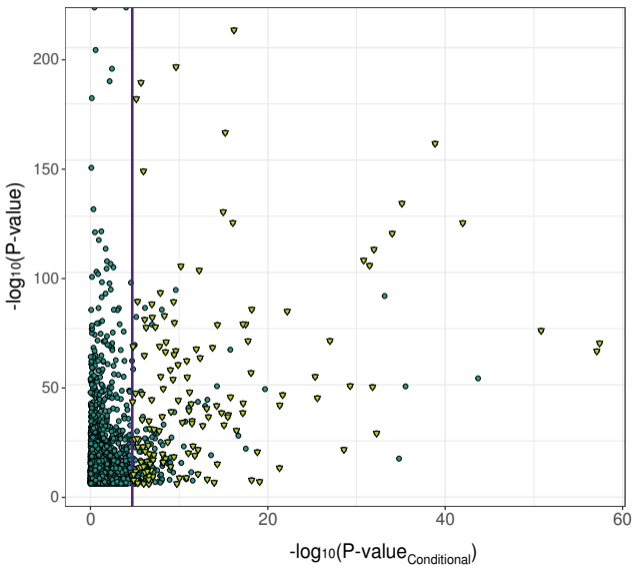
A



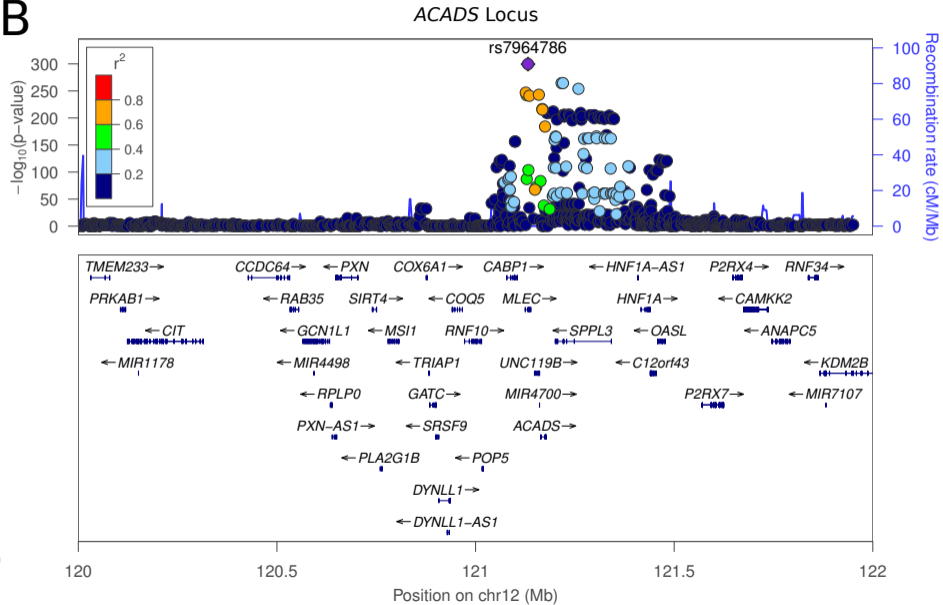
B

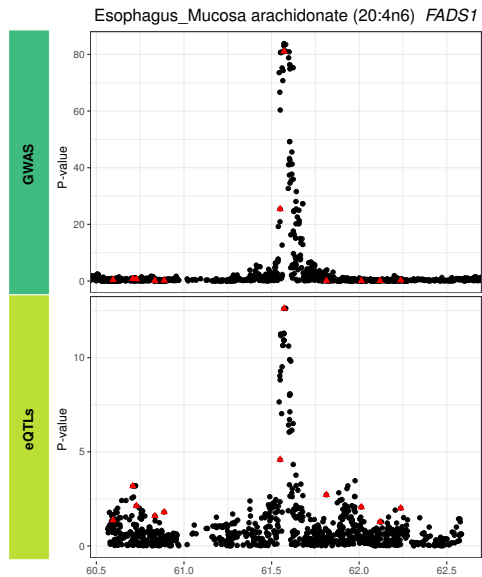
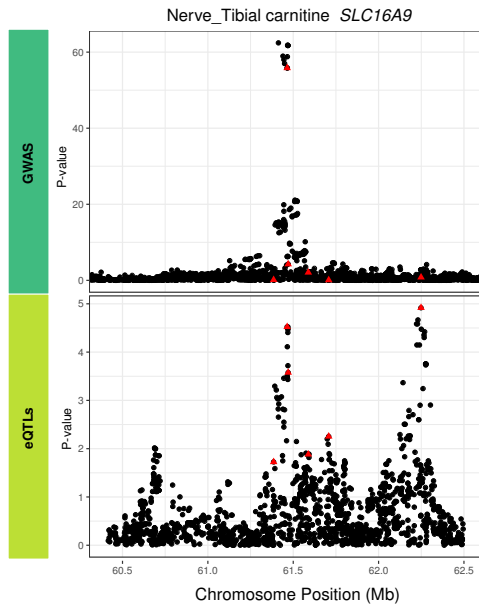
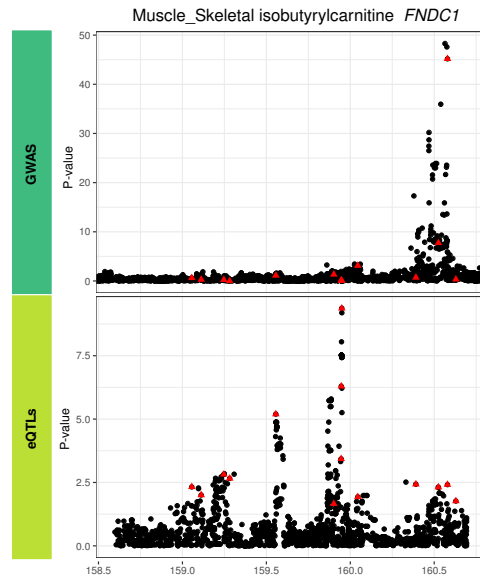


A

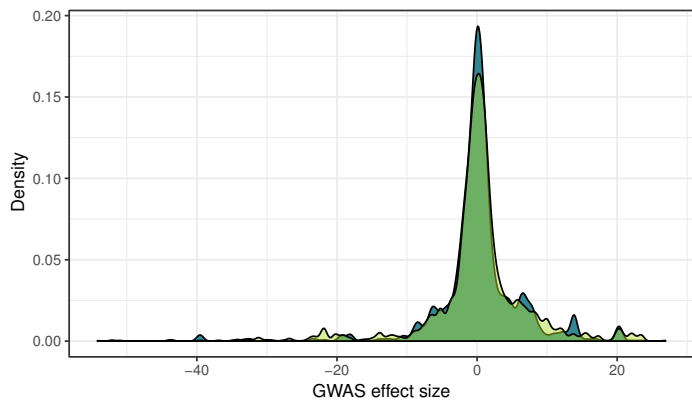
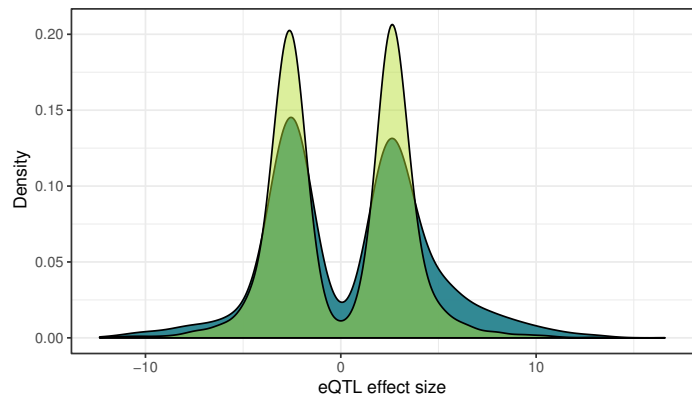
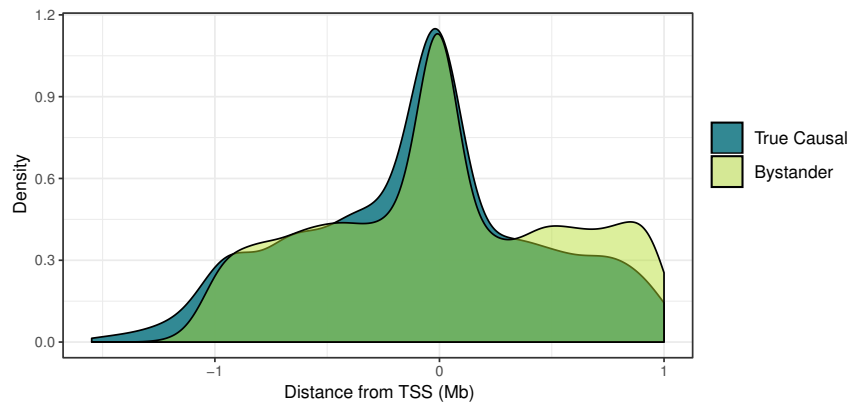


B



**A****B****C**



**A****B****C****D**