Multi-tissue polygenic models for transcriptome-wide association studies

```
Yong jin Park, 1, 2, * Abhishek Sarkar, 1, 2, * Kunal Bhutani, 3 and Manolis Kellis 1, 2, †
            ^1Computer Science and Artificial Intelligence Laboratory,
            Massachusetts Institute of Technology,
5
            Cambridge, MA,
6
            USA
            <sup>2</sup>Broad Institute of MIT and Harvard,
8
            Cambridge, MA,
9
            USA
10
            <sup>3</sup>Department of Bioinformatics & Systems Biology,
11
            University of California,
12
            San Diego, CA,
13
            USA
14
```

1 I. ABSTRACT

Transcriptome-wide association studies (TWAS) have proven to be a powerful tool to identify genes associated with human diseases by aggregating *cis*-regulatory effects on gene expression. However, TWAS relies on building predictive models of gene expression, which are sensitive to the sample size and tissue on which they are trained. The Gene Tissue Expression Project has produced reference transcriptomes across 53 human tissues and cell types; however, the data is highly sparse, making it difficult to build polygenic models in relevant tissues for TWAS. Here, we propose fQTL, a multi-tissue, multivariate model for mapping expression quantitative trait loci and predicting gene expression. Our model decomposes eQTL effects into SNP-specific and tissue-specific components, pooling information across relevant tissues to effectively boost sample sizes. In simulation, we demonstrate that our multi-tissue approach outperforms single-tissue approaches in identifying causal eQTLs and tissues of action. Using our method, we fit polygenic models for 13,461 genes, characterized the tissue-specificity of the learned *cis*-eQTLs, and performed TWAS for Alzheimer's disease and schizophrenia, identifying 107 and 382 associated genes, respectively.

14 II. INTRODUCTION

A fundamental barrier to interpreting the role of non-coding genetic variation identified by genome-wide association studies (GWAS) is understanding how specific single nucleotide polymorphisms (SNPs) cause changes in gene expression, and how those changes in expression lead to downstream phenotypes. Recently, transcriptome wide association studies (TWAS) have proven to be a powerful tool to predict the impact of *cis*-regulatory regions on expression and directly associate genes with downstream disease phenotypes 1,2 the key idea of TWAS is to train multivariate expression quantitative trait loci (eQTL) models on reference expression panels, use these models to predict (impute) unobserved gene expression in large scale GWAS cohorts, and compute association statistics by regressing phenotype directly onto imputed gene expression. The success of TWAS is dependent on the sample size of the expression reference panel and the tissue in which expression was measured. The Gene Tissue Expression (GTEx) Project has produced reference expression profiles for 450 individuals across 53 tissues; however, only a subset of individuals were measured for each tissue³. The GTEx dataset is highly sparse: 70% of individual-tissue-gene expression observations

^{*} These authors contributed equally.

[†] manoli@mit.edu

²⁷ are missing. Here, we seek to improve polygenic eQTL modeling by pooling information across tissues to both overcome limited sample size and improve prediction accuracy.

29 Previously proposed multi-tissue (multi-trait) QTL approaches have primarily focused on identifying the 30 tissue of action using Bayesian model comparison^{4,5}. However, combinatorial searches over models quickly become intractable; for example, to find the best model for 20 tissues, we need to compare millions of candidate models for each gene. Moreover, existing approaches still assume independence of SNPs and train univariate eQTL models. In parallel, several multivariate regression methods have been proposed to identify causal eQTLs within single tissues, based either on forward selection⁶ or Bayesian variable selection^{7–9}. 35 Here, we propose factored polygenic QTL analysis (fQTL), a multi-response, multivariate regression model 36 for eQTL mapping, to jointly identify causal eQTLs and tissues of action. Our model has several advantages 37 over existing methods: (1) We fit a multivariate regression model jointly modeling the effect of all cis-SNPs on expression while performing variable selection over SNPs in linkage disequilibrium (LD). (2) We jointly model multiple gene expression vectors across tissues within a single unified model to pool information across 40 relevant tissues. (3) We factorize the SNP by tissue effect size matrix into genomic location-dependent and 41 tissue-dependent components, improving the interpretability of the fitted model. We demonstrate that our 42 approach outperforms existing methods in a variety of realistic simulations. We trained fQTL models for 13,641 genes on the GTEx reference cohort, used them to characterize multi-tissue regulatory contexts in ⁴⁴ TWAS analysis of Alzheimer's disease (AD) and schizophrenia (SCZ), and identified 107 AD-associated and 45 382 SCZ-associated genes. Several of the genes provide interesting clues to disease mechanisms and should be 46 further investigated in follow-up studies. We have made software implementing the method, trained model

47 weights, and full TWAS summary statistics publicly available (https://github.com/ypark/fqtl).

48 III. METHODS

49 A. Factored QTL model specification

⁵⁰ We propose fQTL, a multi-response, multivariate regression approach to jointly model gene expression across ⁵¹ tissues and individuals. For each gene, we assume the columns of the $n \times m$ (individual-by-tissue) gene ⁵² expression matrix Y are normalized to be approximately Gaussian. We regress Y onto the $n \times p$ genotype ⁵³ matrix X, which is centered (but not scaled). We fit a linear model:

$$\begin{split} Y_{it} &\sim \mathcal{N}(\mathbb{E}[Y_{it} \mid \cdot], \mathbb{V}[Y_{it} \mid \cdot]) \\ \mathbb{E}[Y_{it} \mid X, \cdot] &= \sum_{j=1}^p X_{ij} \Theta_{jt} \\ \mathbb{V}[Y_{it} \mid \cdot] &= \underbrace{\frac{1}{n} \sum_{i=1}^n (Y_{it} - \mathbb{E}[Y_{it} \mid \cdot])^2}_{\text{empirical error}} \\ &+ \underbrace{V_{\min}}_{\text{for numerical stability}} + \underbrace{V_{\max} \sigma(\theta_{\text{var},t})}_{\text{unexplained variance}} \end{split}$$

⁵⁴ We define Θ_{jt} to be the effect of SNP j on gene expression of that gene in tissue t. We define ⁵⁵ $\sigma(z) = 1/(1 + \exp(-z))$ and we set $V_{\sf max} = \mathbb{V}[y]$ and $V_{\sf min} = 10^{-4} V_{\sf max}$.

⁵⁶ The key idea of fQTL is that we assume the eQTL effect size matrix Θ can be decomposed into tissue-invariant components $\theta^{\sf snp}$ and tissue-dependent components $\theta^{\sf tis}$:

$$\Theta_{jt} = \sum_{r=1}^K \theta_{jr}^{\mathsf{snp}} (\theta^{\mathsf{tis}})_{rt}^\top,$$

- Here, we assume K=1 for ease of interpreting the results, although our inference algorithm (described below) supports fitting arbitrary $K \leq m$.
- $_{60}$ The fundamental problem in fitting a multivariate regression on genotypes X is co-linearity of the columns
- $_{61}$ of X due to LD. Co-linearity causes ordinary regression to be ill-posed (infinitely many solutions), but a
- 62 Bayesian regression with the spike and slab prior can successfully solve the variable selection problem $^{10-15}$.
- 63 Here, we assume the spike and slab (point-normal mixture) prior on each column of θ^{snp} and θ^{tis} . Considering
- 64 one of the columns θ :

$$p(\theta_i \mid \tau, \pi) \sim \pi \mathcal{N}(\theta_i; 0, \tau^{-1}) + (1 - \pi)\delta_0(\theta_i)$$

65 where $\mathcal{N}(x;\cdot,\cdot)$ denotes the Gaussian density and $\delta(x)$ denotes the Dirac delta function (point mass at zero).

66 B. Stochastic variational inference

For our model, the spike and slab prior is non-conjugate to the likelihood. Therefore, estimating the posterior distribution of θ requires either Markov Chain Monte Carlo (MCMC)^{16,17}, expectation propagation^{18,19}, or variational inference^{13,20}. Based on our preliminary experiments, we chose not to use MCMC due to convergence problems. Instead, we perform variational inference: we re-cast the problem as an optimization problem over variational parameters $\Psi = (\alpha, \beta, \gamma)$, finding the best mean-field approximating distribution $q(\theta \mid \Psi)$ to the intractable posterior $p(\theta \mid X, Y)$.

$$q(\theta \mid \Psi) = \prod_{j} q(\theta_{j} \mid \alpha_{j}, \beta_{j}, \gamma_{j})$$
$$q(\theta_{j} \mid \alpha_{j}, \beta_{j}, \gamma_{j}) = \alpha_{j} \mathcal{N}(\beta_{j}, \gamma_{j}^{-1}) + (1 - \alpha_{j}) \delta(\theta_{j})$$

73 Under the variational approximation, the parameters θ_j are mutually independent and their means and 74 variances (approximate posterior means and posterior variances) can be written in terms of the variational 75 parameters:

$$\mathbb{E}_{q}[\theta_{j}] = \alpha_{j}\beta_{j}$$

$$\mathbb{V}_{q}[\theta_{j}] = \alpha_{j}\gamma_{j}^{-1} + \alpha_{j}(1 - \alpha_{j})\beta_{j}^{2}$$

⁷⁶ Our goal is to find the q with minimum KL-divergence with p by optimizing over Ψ ; however, in general ⁷⁷ the KL-divergence does not have an analytic form. Instead, we solve an equivalent optimization problem, ⁷⁸ maximizing the evidence lower-bound (ELBO):

$$\mathcal{L} = \mathbb{E}_q[\ln p(\mathbf{y}, \theta \mid X)] - \mathbb{E}_q[\ln q(\theta)]$$

$$= \underbrace{\mathbb{E}_q[\ln p(\mathbf{y} \mid X, \theta)]}_{\text{reconstruction error}}$$

$$- \underbrace{\mathcal{K}\mathcal{L}(q(\theta \mid \Psi) || p(\theta \mid \cdot))}_{\text{regularizer}}$$

Prior work derived coordinate ascent updates to maximize the ELBO assuming hyperparameters (π, τ, σ^2) were fixed¹³. The method is scalable due to fast convergence, but is sensitive to the hyperparameters and to the initialization of the variational parameters. In practice, the algorithm requires two outer loops of importance sampling steps over the hyperparameter space: one to find an optimal initialization, and one to actually perform the optimization from a warm start.

- 84 We instead used stochastic optimization to optimize the variational objective, an approach known as stochastic
- 85 variational inference (SVI)^{21,22}. The key idea is to perform gradient ascent, taking steps along the natural
- 86 gradient of the objective function²³ with respect to the variational parameters:

$$\nabla_{\Psi} \mathcal{L} = \nabla_{\Psi} \mathbb{E}_{q}[\ln p(\mathbf{y} \mid X, \theta)] - \nabla_{\Psi} \mathcal{K} \mathcal{L}(q(\theta) || p(\theta))$$
$$\Psi^{(t+1)} = \Psi^{(t)} + \lambda \nabla_{\Psi} \mathcal{L}$$

For our choice of approximating family q, the regularizer term above has an analytic form, as previously derived¹³:

$$\mathcal{KL}(q(\theta \mid \Psi) || p(\theta \mid \cdot)) = \frac{1}{2} \sum_{j} \alpha_{j} \left(1 + \ln \tau - \ln \gamma_{j} - \tau(\gamma_{j}^{-1} + \beta^{2}) \right)$$
$$- \sum_{j} \left(a_{j} \ln \left(\frac{a_{j}}{\pi} \right) + (1 - \alpha_{j}) \ln \left(\frac{1 - \alpha_{j}}{1 - \pi} \right) \right)$$

89 Although we can analytically take gradients of the regularizer (not shown), our use of the spike and slab 90 prior means it is not possible to take gradients of $\mathbb{E}_q[\ln p(y,|X,\theta)]$. We use the log-derivative trick to obtain

91 an unbiased estimator for the natural gradients (see Supplementary Methods)²¹:

$$\nabla_{\Psi} \mathbb{E}_{q}[\ln p(\mathbf{y} \mid X, \theta)]$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} \ln p(\mathbf{y} \mid X, \theta^{(s)}) \nabla_{\Psi} \ln q(\theta^{(s)}), \theta^{(s)} \sim q(\theta^{(s)})$$

Naive implementation of SVI is slow to converge because of the variance of the stochastic gradient. To speed up convergence, we used a variance reduction technique known as control variates 21,22 , developing a novel control variate for our problem (see Supplementary Methods). We additionally re-parameterized the model in terms of the genetic values $\eta = X\theta$, a parameterization which has been previously studied 24,25 . This parameterization has several advantages: (1) We avoid sampling p-vectors θ , reducing the computation time for n < p (which is true in our setting). (2) We further reduce the variance of the stochastic gradient 25 . Even with variance reduction techniques, the convergence rate and sparsity of the final fitted model is highly sensitive to the hyperparameters and learning rate. We developed a variational approximation for the hyperparameters (see Supplementary Methods) which allowed us to simultaneously fit the model parameters θ and hyperparameters (see Supplementary Methods) which allowed us to simultaneously fit the model parameters of θ and hyperparameters (θ). We tuned the learning rate on simulated datasets, finding that a fixed value θ 0.01 worked well.

103 C. Factored QTL model inference

¹⁰⁴ In order to fit fQTL using SVI, we only need to characterize the distribution of η under the variational ¹⁰⁵ approximation, relying on a previously derived result for the variance of the product of two random vectors²⁶:

$$\begin{split} \eta_{it} &\sim \mathcal{N}\left(\sum_{j=1}^{p} X_{ij} \mathbb{E}[\Theta_{jt}], \sum_{j=1}^{p} X_{ij}^{2} \mathbb{V}[\Theta_{jt}]\right) \\ \mathbb{E}[\Theta] &= \mathbb{E}[\theta^{\mathsf{snp}}] \mathbb{E}[\theta^{\mathsf{tis}}]^{\top} \\ \mathbb{V}[\Theta] &= \mathbb{V}[\theta^{\mathsf{snp}}] \mathbb{V}[\theta^{\mathsf{tis}}]^{\top} \\ &+ (\mathbb{E}[\theta^{\mathsf{snp}}] \circ \mathbb{E}[\theta^{\mathsf{snp}}]) \mathbb{V}[\theta^{\mathsf{tis}}]^{\top} \\ &+ \mathbb{V}[\theta^{\mathsf{snp}}] (\mathbb{E}[\theta^{\mathsf{tis}}] \circ \mathbb{E}[\theta^{\mathsf{tis}}])^{\top} \end{split}$$

106 where o denotes elementwise multiplication.

107 D. Simulation study

We simulated multi-tissue gene expression matrices Y using genotypes in 450 individuals from the specified cis-regulatory window around each gene, varying the number of causal eQTLs, the number of tissues of action, and the proportion of variance of gene expression h^2 explained by SNP-specific effects. For each setting, we simulated 175 randomly chosen genes. We matched the number of missing observations in each tissue to the number of missing observations in each GTEx tissue. We then quantile-normalized the generated expression values to match real GTEx expression profiles.

We sampled SNP-specific effects $\theta_j^{\sf snp}$ and tissue-specific effects $\theta_t^{\sf tis}$ from the standard normal distribution. We then generated gene expression within tissues with non-zero tissue effect (i.e., tissues within which gene expression could also be genetically controlled) using a linear model 17,27:

$$\begin{split} y_{it} &= \eta_{it} + \epsilon_i \\ \eta_{it} &= \sum_{j=1}^p X_{ij} \theta_j^{\mathsf{snp}} \theta_t^{\mathsf{tis}} \\ \epsilon_i &\sim \mathcal{N}\bigg(0, \left(\frac{1}{h^2} - 1\right) \mathbb{V}[\boldsymbol{\eta}_t]\bigg) \end{split}$$

117 For tissues with zero tissue effect, we sampled expression values from $\mathcal{N}(0, \mathbb{V}[\eta_t])$.

118 E. Identification of tissues of action in GTEx

We obtained individual genotypes and gene-level RNA-seq read counts from the GTEx consortium (version 6). We restricted our analysis to coding genes annotated by GENCODE v19 and tissues with sample size $121 \text{ n} \geq 50$. Within each tissue, we further restricted to genes having read count greater than 10 in at least 122 50% of samples. To convert read counts to logarithmic scale, rlog transformed the data using the DESeq2 package²⁸. We confirmed after the rlog transformation there was no mean-variance correlation.

For GTEx gene expression, significant PVE is attributable to experimental and technical confounding variables. For each tissue, we jointly fit a polygenic QTL model and performed sparse matrix factorization of the residual to identify non-genetic, technical confounders on the 1000 most variable genes in each tissue. We modeled the k-th most variable gene's expression in i-th sample as:

$$\eta_{ik} = \sum_{r=1}^{K} U_{ir} V_{rk}^{\top} + \sum_{j=1}^{p} X_{ij} \Theta_{jk}$$

Here, the matrix of genetic effects Θ_{jk} denotes the effect of SNP j on gene k. We fixed $\Theta_{jk}=0$ if SNP j was not located within the specified cis-regulatory window of gene k. We enforced rank sparsity of the factorization by assuming a group spike and slab prior 18 on U and V (see Supplementary Methods). We defined learned covariates as columns of U with column posterior inclusion probability PIP > 0.5. We included these learned covariates as well as known covariates in subsequent analysis by introducing another linear term in our original model:

$$\mathbb{E}[Y_{it} \mid X, \cdot] = \sum_{j=1}^{p} X_{ij} \Theta_{jt} + \sum_{k=1}^{q} C_{ik} \Lambda_{kt}$$

To account for correlations between the columns of C, we assume the spike and slab prior on the elements Λ_{kt} , and incorporate the entire matrix into our stochastic variational inference algorithm described above.

136 F. TWAS using factored QTL effects

We perform tissue-specific TWAS by defining tissue-specific QTL effects $\theta_t = \theta^{\sf snp} \theta_t^{\sf tis}$. Prior work proposed a summary-based TWAS statistic $T = \mathbf{z}^{\top} \theta_t$ and characterized its null distribution. However, the proposed statistic depends only on a point estimate of θ_t , while our factored QTL model estimates both a posterior mean and a posterior variance of θ_t .

We derive the null distribution of T, accounting for the posterior variance of the estimated effects. As before, we assume the null distribution of GWAS z-scores follows the multivariate Gaussian distribution, $\mathbf{z} \sim \mathcal{N}(\hat{\mathbf{z}}, R)$ where the covariance matrix $R = X^{\top}X/n$ is estimated from the GWAS cohort genotypes (or approximated using a reference cohort).

Using previously derived results for the mean and variance of the inner product of two stochastic vectors²⁶, we can define $\mathbb{V}[\theta^{\mathsf{snp}}\theta_t^{\mathsf{tis}}]$, and use the same result to define $\mathbb{V}[\mathbf{z}^{\top}\theta_t]$.

$$\begin{split} \mathbb{E}[T] &= \hat{\mathbf{z}}^{\top} \mathbb{E}[\theta_t] \\ \mathbb{V}[T] &= \mathbb{E}[\theta_t]^{\top} R \mathbb{E}[\theta_t] + \hat{\mathbf{z}}^{\top} \mathbb{V}[\theta_t] \hat{\mathbf{z}} + \mathsf{tr}(\mathbb{V}[\theta_t] R) \end{split}$$

147 Compared to the previously derived null distribution, the null distribution of our TWAS statistic has two
148 additional terms in the variance which additionally penalize the model complexity of the polygenic QTL
149 model. This modified variance leads to a more conservative hypothesis test.

150 We performed summary-based tissue-specific TWAS using our method for Alzheimer's disease and schizophre-151 nia. We downloaded summary statistics for Alzheimer's disease from the International Genomics of 152 Alzheimer's Project (http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php) and statistics 153 for schizophrenia from the Psychiatric Genetics Consortium (http://www.med.unc.edu/pgc).

154 International Genomics of Alzheimer's Project (IGAP) is a large two-stage study based upon genome-wide association studies (GWAS) on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyze four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative – EADI the Alzheimer Disease Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium – GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 & 2.

163 IV. RESULTS

164 A. Factored QTL model specification and inference

We propose fQTL, a multi-response, multivariate regression approach to jointly model gene expression variation across individuals and tissues. For each gene, we fit a linear model regressing expression in individuals i and tissues t against cis-SNPs j:

$$\mathbb{E}[Y_{it} \mid X, \cdot] = \sum_{j=1}^{p} X_{ij} \Theta_{jt}$$

168 The key idea of fQTL is to assume the eQTL effect size matrix Θ can be decomposed into tissue-invariant 169 components θ^{snp} and tissue-dependent components θ^{tis} :

$$\Theta_{jt} = \sum_{r=1}^K heta_{jr}^{\sf snp}(heta^{\sf tis})_{rt}^{ op}$$

170 In this study, we focus on K=1 for intuitive interpretation, although our method supports arbitrary K.
171 In this case, the intuition behind the factorization is that the estimated effect of a specific SNP on gene
172 expression in a specific tissue could be explained by several possible mechanisms. (1) The variant might
173 disrupt a regulatory motif and alter transcription factor binding, altering gene expression. (2) The region
174 harboring the variant might not be active in the specific tissue, due to reduced chromatin accessibility or
175 altered epigenomic modifications. (3) The upstream transcription factor which mediates the effect of the
176 SNP might be differentially expressed in the specific tissue. Explanation (1) implies the effect is invariant
177 across tissues since it is determined solely by sequence, while explanations (2) and (3) can be tissue-specific.
178 By factorizing the matrix of effects Θ , we allow our model to fit both types of effects. As K increases, we
179 posit the existence of different causal eQTLs for different subsets of tissues, but do not explore this possibility
180 in this study.

The challenge in fitting polygenic QTL models is that the columns of X are co-linear due to LD, and we need to select a sparse set of relevant predictors using the spike and slab prior ¹³. This prior is nonconjugate to the likelihood, preventing its widespread use in polygenic modeling. We developed a general computational framework to efficiently perform approximate Bayesian inference with this prior in large-scale sparse regression models. The key ideas of our approach are: (1) We use techniques from from black-box variational inference to allow rapid model specification and implementation²². (2) We develop novel variance reduction techniques to improve stochastic variational inference for these models²¹. (3) We use our framework to implement sparse regression and matrix factorization, which are the building blocks of fQTL. We have made C++ source code implementing the framework and an R package wrapping the models publicly available (https://github.com/ypark/fqtl).

Our analysis pipeline is shown in Fig. 1. For each tissue, we jointly performed matrix factorization and multi-response (across genes rather than across tissues), multivariate *cis*-eQTL mapping to learn non-genetic covariates on the 1,000 most variable genes. We then corrected both known and learned covariates in all genes and fit our multi-tissue fQTL model including SNPs from 500kb upstream of the transcription start site to 500kb downstream of the transcription end site. Finally, we used the estimated SNP and tissue effects to perform tissue-specific TWAS, accounting for the posterior variance of the estimated QTL effects.

197 B. Simulation study

201

202

205

206

198 We evaluated fQTL in realistic simulations and found it outperformed existing methods for a variety of 199 scenarios. We estimated the statistical power for two inference problems: (1) identifying the causal SNPs 200 and (2) identifying the tissues of action. We compared fQTL to the following methods:

- Single-tissue polygenic QTL models, which we refer to as sQTL (equivalent to fQTL with one tissue and $\theta^{\mathsf{tis}} = 1$)
- Lasso regression²⁹, implemented in glmnet R package³⁰ (LASSO). We tuned the regularization hyperparameter λ by cross validation.
 - Elastic-net³¹ regression, implemented in glmnet R package³⁰ (ElasticNet). We tuned the regularization hyperparameter λ by cross validation, but we fixed the ridge-lasso tradeoff parameter to $\alpha = 0.5$, which was used in previous TWAS analysis².

We additionally sought to compare fQTL to the Bayesian sparse linear mixed effect model (BSLMM) implemented in GEMMA^{17,32}. However, in preliminary experiements BSLMM could require an infeasible number of Markov Chain Monte Carlo iterations depending on the number of *cis*-SNPs. Furthermore, the

211 convergence rate of the algorithm varied from gene to gene, so we did not include BSLMM in our analysis.
212 In our experiments, when the model converged it performed equivalently to Elastic-net and sQTL.

213 A key problem in performing the proposed inference tasks is setting thresholds for discovery on the estimated 214 model effect sizes. For causal variant prediction, we ranked SNP-tissue pairs according to the magnitude 215 of regression coefficients $|\Theta_{jt}|$ in decreasing order and calculated false discovery rate (FDR) and power 216 using the PRROC package in $\mathbb{R}^{33,34}$. For tissue of action prediction, we calculated the proportion of variance 217 explained (PVE) of gene expression by tissue-specific effects $\mathbb{V}[X\theta^{\mathsf{snp}}\theta_t^{\mathsf{tis}}]/\mathbb{V}[\mathbf{y}_t]$. We then ranked tissues in 218 the decreasing order of PVE and calculated FDR and power.

We simulated multi-tissue expression datasets matching the level of missingness in GTEx using *cis*-genotypes in windows around randomly chosen genes varying the number of tissues of action, the number of causal eQTLs, and the heritability of gene expression. We first estimated the power to identify the causal SNP and found that fQTL had higher power except when there was only one tissue of action (Fig. 2a). In this case, fQTL pools information across irrelevant tissues, leading to poorer predictive models. Interestingly, as we increased the number of causal eQTLs (but still fixed a single tissue of action), fQTL power increased in causal variant identification, eventually matching the power achieved by Elastic-Net. As expected, sat the number of tissues of action increases, fQTL outperforms single-tissue methods which are limited by sample size (missingness) in the training data. fQTL pools information across multiple tissues while favoring sparse solutions, allowing it to robustly identify causal SNPs without over-fitting and without pre-screening irrelevant tissues.

We then investigated the power to detect the tissue of action (Fig. 2b). The results generally agree with those for causal variant identification. However, the number of causal SNPs had less impact on tissue prediction. Multi-tissue fQTL clearly dominated in all settings and gained more power as the number of tissues of action increased.

We finally investigated the sensitivity of our method to mis-specification of K. We first simulated scenarios where causal eQTLs were shared between K=2 and K=3 sets of tissues of action, but with different effect sizes in each subset (Supp. Fig. 7, K=2; Supp. Fig. 8, K=3). We found that fQTL assuming K=1 performed only slightly worse than fQTL with the correct rank specification. In this study we did not explore learning K=1 from the data because the GTEx data is too sparse (i.e., many tissues do not have enough individuals) to perform cross validation or hold-out validation. Importantly, fQTL with rank one still outperformed other single-tissue methods in these scenarios.

We then simulated a worst-case scenario for our model with arbitrary K by randomly assigning elements of 242 Θ to be non-zero (Supp. Fig. 9). In this scenario, expression in different subsets of tissues of action is driven 243 by different causal eQTLs. As expected, fQTL performed worse than single-tissue methods in identifying 244 the causal variant, as it tries to pool information across irrelevant tissues. However, fQTL was still the most 245 powerful method to identify the tissue of action when the SNP effect-mediated expression heritability $h^2 \leq .3$ 246 and there were only few tissues of action. In this case, fQTL was able to capture some of the causal tissues 247 and therefore successfully identify the causal eQTLs by pooling across those tissues.

248 C. Identification of tissues of action in GTEx

Unlike our simulated data, significant PVE of real gene expression, such as that measured by the GTEx project, is attributable to experimental and technical confounding variables. For each gene in each tissue, we performed sparse matrix factorization to identify non-genetic, technical confounders on the 1000 most variable genes. We restricted our analysis to 48 tissues with sample size $n \geq 50$, and found 2-15 non-genetic confounding factors in 40 of the 48 tissues. In the remaining 8 tissues, we found factors which were essentially 0. The learned factors explain a wide range of gene expression across tissues: for example, 15 factors explain 67% of whole blood gene expression variance, where 2 factors explain only 11% of spleen expression variance. Across the 48 tissues, 30% of variability is explained by non-genetic factors on average.

²⁵⁷ We found many of the inferred factors are highly correlated with known covariates: tissue ischemic time, age, body mass index, and gender (Supp. Fig. 10). However, in our analysis we included both known and learned ²⁵⁹ covariates, again using the spike and slab prior to perform variable selection among the correlated covariates.

We next defined a set of 13,641 active genes (FDR 1%) by estimating the PVE of each gene's expression explained by its cis-SNPs $\mathbb{V}[X\theta^{\mathsf{snp}}\theta_t^{\mathsf{tis}}]/\mathbb{V}[y]$. We took the union of heritable genes identified across tissues. control the FDR of active genes, we constructed the null distribution of PVE by permuting the samples (Supp. Fig. 11). For comparison, we estimated the PVE explained by \mathfrak{sQTL} by fixing $\theta^{\mathsf{tis}} = 1$ and used the same procedure to identify active genes. As expected, the number of active genes per tissue was highly correlated with the sample size in that tissue (Fig. 3). However, \mathfrak{fQTL} increased power to detect active genes over \mathfrak{sQTL} , finding on average 3.8 times as many active genes across the 48 tissues.

We then used the estimated tissue posterior inclusion probability $p(\theta_t^{\text{tis}} \neq 0)$ to compute a correlation matrix between tissue activity patterns across all genes (Fig. 4). Our results broadly agree with those previously reported³⁵: the brain tissues form a separate cluster, but non-brain tissues cluster together with no clear sub-clustering based on global tissue activity patterns. It is possible that non-brain tissues will form sub-clusters when examining tissue activity for subsets of genes in specific gene pathways; however, we do not explore that possibility in this study.

273 D. TWAS using factored QTL effects

274 For each gene, we combined the estimated SNP and tissue effects from the fQTL model to estimate tissue-275 specific QTL effects and perform tissue-specific TWAS for two psychiatric disorders: Alzheimer's disease³⁶ 276 (AD), and schizophrenia (SCZ)³⁷. We have made all estimated fQTL model parameters and full TWAS 277 summary statistics publicly available (https://github.com/ypark/fqtl).

278 We found 107 AD-associated genes at FDR 1% (1fdr calculation with normal null distribution by ashr 279 package³⁸; Fig. 5), of which only 10 have been previously reported in the NHGRI GWAS catalog³⁹ as 280 associated with AD. We found 40 genes significantly associated in brain tissues, of which 6 have been included 281 in the GWAS catalog: BCAM, CR1, HLA-DRB1, HLA-DRB5, MRPL10, PVRL2.

We compared our tissue-specific TWAS results to single-tissue TWAS based on our sQTL models and found that fQTL greatly increased the power to detect both constitutively active and brain-specific genes associated with AD. Using sQTL, we found 282 AD-associated genes, including 15 known AD genes. However, 191 of the 282 genes are significant only in a single tissue, and only 11 are associated in brain tissues. In comparison, 5 of the 191 genes were found to be significant in more than one tissue by fQTL. None of the 15 single tissue TWAS genes in the GWAS catalog were significantly associated in brain, likely due to limited sample sizes. In comparison, 3 of these 15 genes were found by fQTL in brain tissues (PVRL2, HLA-DRB1, HLA-DRB5).

Two of the brain-specific AD TWAS genes are already known to be associated with AD. BCAM (basal cell adhesion molecule) and PVRL2 (poliovirus receptor-related 2) are both located in the APOE locus^{40–42}. However, BCAM is also confirmed as an AD-associated gene in follow up meta-analysis⁴³. Similarly, PVRL2 remains significantly associated with AD after adjusting for LD structure⁴¹.

Our brain-specific TWAS identified SERPINH1 (Serine/Cysteine Proteinase Inhibitor 1), which has been previously associated with height in the GWAS catalog. SERPINH1 is elevated in microglial cells after treatment with amyloid beta, the classical hallmark of AD⁴⁴. Differential expression of SERPINH1 is also linked with other psychiatric disorders such as major mood disorders⁴⁵.

Our brain-specific TWAS also identified Claudin-5 (CLDN5), which is important for the function of the blood-brain barrier (BBB). In mice, brain injury followed by down-regulation of CLDN5 increased the permeability of BBB⁴⁶. Depletion of CLDN5 in the blood-cerebrospinal fluid barrier may also lead to increased permeability, imbalance of the immune system, and neurodegeneration⁴⁷.

we found 382 tissue-specific TWAS genes in SCZ using fQTL (FDR 1%, Fig. 6), of which 45 have been previously reported in the GWAS catalog. 137 of the 382 genes are associated in brain tissues, of which 16 are in the GWAS catalog. As was the case for AD, we found that tissue-specific TWAS using fQTL greatly increased power to detect brain-specific TWAS genes. We found 845 single-tissue TWAS genes using sQTL, of which 82 have been previously reported in the GWAS catalog. However, 507 of the 845 genes are associated in only one tissue, only 61 genes of the 845 genes are significantly associated in brain tissues, and only 1 of the 61 brain TWAS genes (CYP2D6 associated in cerebellar hemisphere) is also included in the GWAS catalog.

- We found a number of brain-specific SCZ TWAS genes which play known roles in other SCZ-relevant phenotypes: APOLD1, C1orf51 (CIART), CARS, and UNC119 (Supp. Fig. 14). APOLD1 (apolipoprotein L subunit D1), also known as VERGE (vascular early response gene), regulates vascular formation in the brain brain and is correlated with stress responses 19.
- 313 C1ORF51 (circadian-associated transcriptional repressor; CIART) is important for circardian rhythm and 314 is robustly expressed in the prefrontal cortex⁵⁰. Circadian rhythms are disrupted in schizophrenia cases 315 compared to controls⁵¹.
- ³¹⁶ CARS (cysteinyl-tRNA synthetase) is associated with alcohol dependence and substance abuse^{52,53}. Many reports document co-occurrence of SCZ and substance abuse^{54–56}.
- 318 UNC119 (uncoordinated 119) regulates an essential pathway of immunological synapse formation and T cell activation 57,58 . In mice, immature T cells trigger cognitive decline and behavioral abnormalities, but can be 320 ameliorated by T-cell activation 59 .

321 V. DISCUSSION

Here, we proposed fQTL, a multi-response multivariate method for eQTL mapping and tissue-specific expression prediction. We showed that our approach improves power over existing methods to detect both causal eQTLs and tissues of action by pooling information across tissues. We used our model to decompose tissue-invariant and tissue-specific QTL effects in the GTEx reference cohort, then combined these effects with summary statistics for AD and SCZ to perform tissue-specific TWAS and identify hundreds of new genes associated with these diseases.

There are a number of modeling and algorithmic improvements which can be made to our approach. Most importantly, in this study we assumed the eQTL-by-tissue effect matrix was rank one for intuitive interpretation: we fit a model with one vector of SNP-specific effects and one vector of tissue-specific effects. Our simulation revealed that mis-specification of this parameter leads to slight loss of power in our model. However, the problem of finding the optimal k requires some attention. We could determine k through cross-validation (although this is not possible in GTEx due to high missingness), optimizing the Bayesian information criterion 60 , or Bayesian model averaging over K.

335 We motivated our study of the rank one model by suggesting that the tissue-specific effect sizes of eQTLs
336 observed in single-tissue approaches could be decomposed into tissue-invariant and tissue-specific components.
337 In particular, we suggested that the tissue-specific component could be explained by variation across tissues
338 in epigenomic state, or by altered expression of the upstream regulator in specific tissues. Distinguishing
339 these two cases is necessary to interpret the mechanism of cis-regulation for genes identified by downstream
340 TWAS and translate TWAS genes into therapeutics. However, the model we have proposed here cannot yet
341 distinguish between these two cases. Future work should incorporate epigenomic information, transcription
342 factor binding, and transcription factor expression to make specific testable biological predictions regarding
343 transcriptional regulation of TWAS genes.

In this study, we assumed gene expression was Gaussian, and coerced the GTEx expression data to be approximately Gaussian. However, a more principled approach would be to use a more appropriate distribution (such as the negative binomial distribution) to directly model the process which generated the observed read counts. Existing methods cannot be easily extended to implement this approach because they make strong assumptions (like Gaussianity) and implement model-specific inference algorithms. Our SVI framework makes implementing extensions of this sort straightforward, without modification of the core inference algorithm. Future work should estimate the improvements in statistical power and false discovery rate when analyzing diverse data types in their observed, un-transformed distribution.

352 VI. ACKNOWLEDGMENTS

353 We thank Alexander Gusev, Benjamin Iriarte, Alkes Price, and Luke Ward for helpful discussions.

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i-Select chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Universite de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant 503480), Alzheimer's Research UK (Grant 503176), the Wellcome Trust (Grant 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01-AG-12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC-10-196728.

368 VII. FIGURES

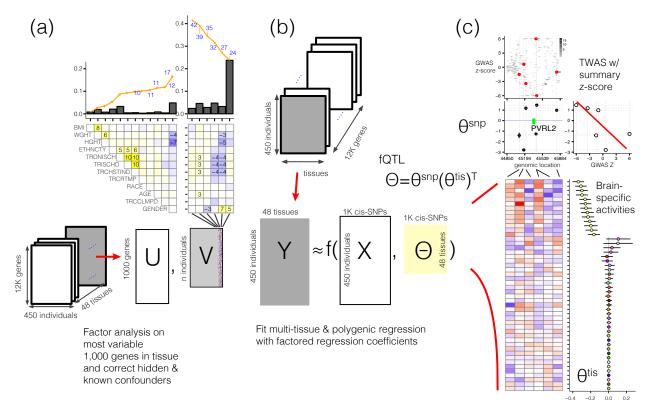


Figure 1: Illustration of multi-tissue fQTL approach to identify tissue contexts in TWAS analysis. (a) Sparse matrix factorization to learn and correct hidden confounders in each tissue. (b) Multi-tissue factored QTL (fQTL) estimation decomposing SNP-specific and tissue-specific effects. (c) TWAS with estimated SNP-specific effects and estimated tissue-specific context.

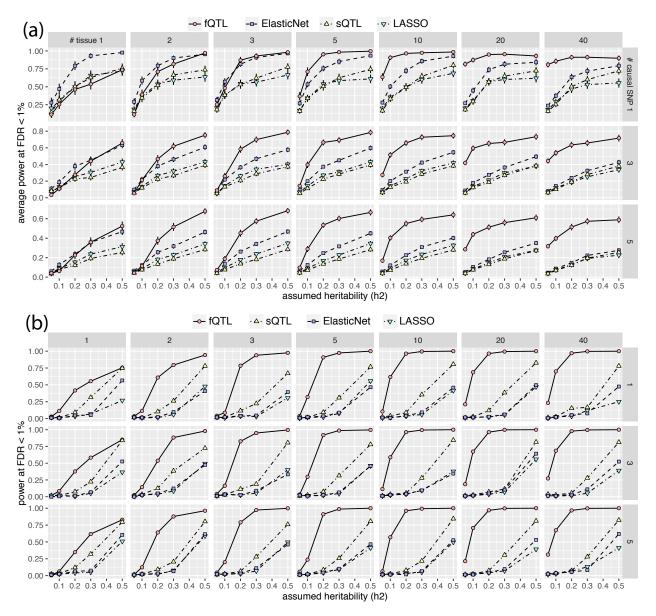


Figure 2: Power calculation of different methods on simulated multi-tissue data. Subpanels are arranged by different number of causal SNPs (rows; 1, 3, 5) and tissues (columns; 1 to 40). X-axis: assumed heritability. Y-axis: statistical power at FDR < 1%. (a) Power calculation of causal SNP prediction averaged over 175 repetitions. Error bars indicate 95% confidence intervals of the mean values (2 standard error). (b) Power calculation of tissue of action prediction.

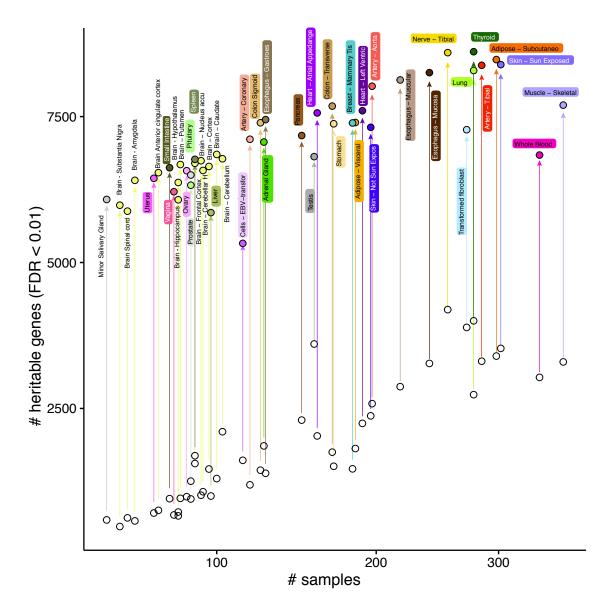


Figure 3: Number of heritable genes across tissues. Each tissue is represented as two points: empty circles denote the number of heritable genes within each tissue found by the single tissue polygenic model, where filled circles denote the number of heritable genes found by the multi-tissue polygenic model. Arrows match the points corresponding to the same tissue.

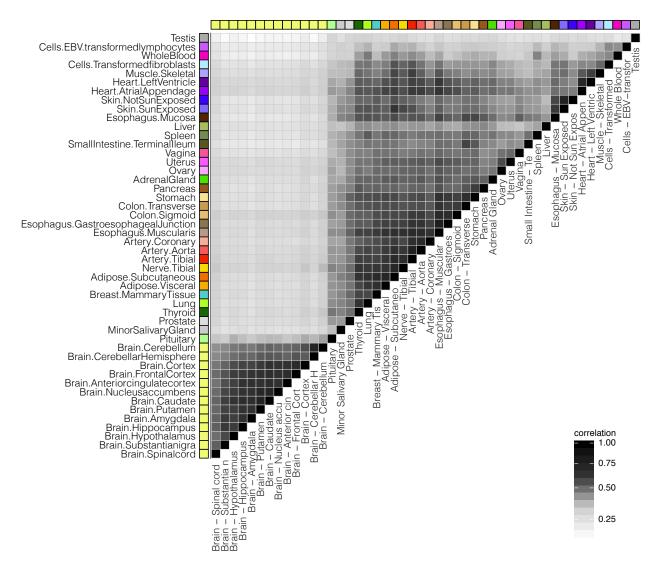


Figure 4: Correlation between tissue activities in fQTL models.

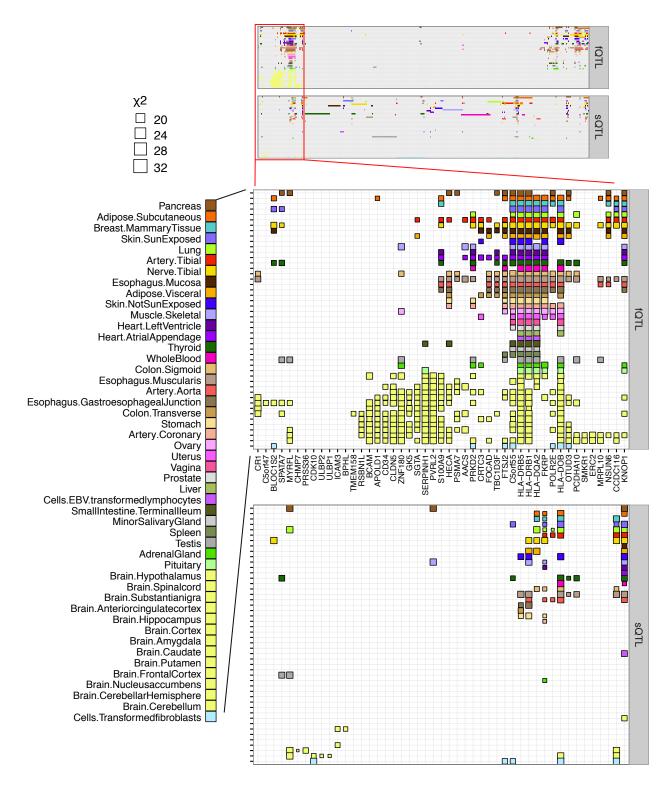


Figure 5: **Tissue-specific AD association statistics.** Columns include 346 genes significantly associated with AD at FDR 1% (107 fQTL; 282 sQTL); rows include 48 GTEx tissues. Sizes of blocks are proportional to χ^2 statistics. Different colors indicate different tissue types.

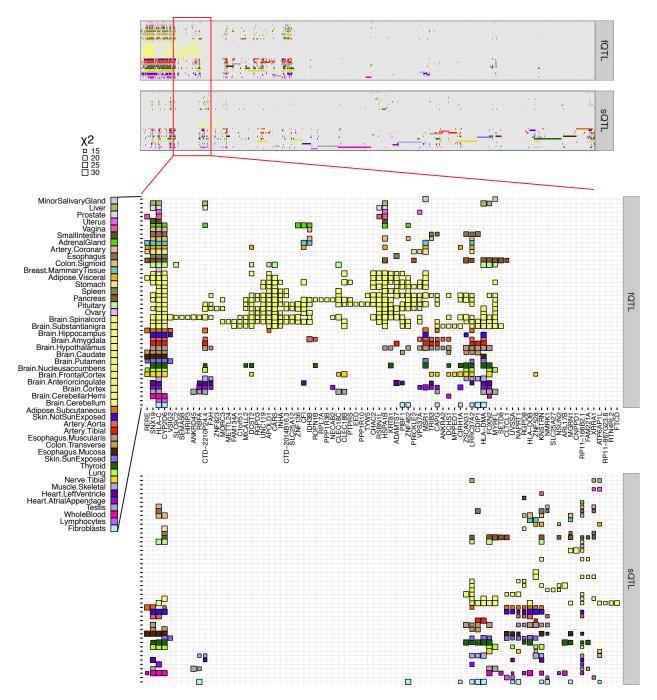


Figure 6: **Tissue-specific SCZ association statistics.** Columns include 1,007 genes significantly associated with AD at FDR 1% (368 fQTL; 845 sQTL); rows include 48 GTEx tissues. Sizes of blocks are proportional to χ^2 statistics. Different colors indicate different tissue types.

369 VIII. SUPPLEMENTARY METHODS

370 A. Stochastic variational inference updates for sparse multivariate regression

Recall we assume the spike and slab prior on the parameter θ , and seek to find the optimal $q(\theta)$ to approximate the intractable posterior $p(\theta \mid X, Y, \cdot)$. Under the variational approximation:

$$\theta_j \sim \alpha_j \mathcal{N}(\beta_j, \gamma_j^{-1}) + (1 - \alpha_j) \delta(\theta_j)$$

373 In order to find the optimal q, we need to optimize the evidence lower bound, which requires estimating a 374 stochastic gradient using the log-derivative trick²¹:

$$\begin{split} \nabla_{\Psi} \mathbb{E}_{q} [\ln p(\mathbf{y} \mid X, \theta)] &= \sum_{s=1}^{S} f(\theta^{(s)}) z_{j}^{(s)} \nabla \ln q_{j} \\ \frac{\partial \ln q_{j}}{\partial \alpha_{j}} &= \frac{(z_{j}^{(s)} - \alpha_{j})}{\alpha_{j} (1 - \alpha_{j})} \\ \frac{\partial \ln q_{j}}{\partial \beta_{j}} &= \gamma_{j} (\theta_{j}^{(s)} - \beta_{j}) \\ \frac{\partial \ln q_{j}}{\partial \gamma_{j}} &= -\frac{1}{2} (\theta_{j}^{(s)} - \beta_{j})^{2} + \frac{1}{2\gamma_{j}} \end{split}$$

Using this approach, we can update (α, β, γ) by directly sampling $\theta^{(s)} \sim q(\theta)$, but this naive algorithm does not scale to high-dimensional models (p > 100 SNPs): sampling scales linearly in p and a large number of samples are required to accurately estimate the joint distribution of θ . In order to avoid sampling $\theta^{(s)}$ directly and speed up the inference, we re-parameterize:

$$q(\eta_i|\alpha,\beta,\nu) = \mathcal{N}(\mu_i,v_i)$$

$$\mu_i = \sum_{j \in [p]} X_{ij} \alpha_j \beta_j$$

$$v_i = \sum_{j \in [p]} X_{ij}^2(\alpha_j \gamma_j^{-1} + \alpha_j (1 - \alpha_j) \beta_j^2)$$

Now, we sample $\epsilon^{(s)} \sim N(0, I)$, re-parameterize again to get a differentiable term⁶¹ $\eta^{(s)} = \mu + \epsilon_s \sqrt{\nu}$, and take gradients with respect to μ, v :

$$\ln q_i = -\frac{1}{2} \ln v_i - \frac{1}{2v_i} (\eta_i^{(s)} - \mu_i)^2 + C$$

$$\frac{\partial \ln q_i}{\partial \mu_i} = v_i^{-1} (\eta_i^{(s)} - \mu_i),$$

$$\frac{\partial \ln q_i}{\partial v_i} = \frac{1}{2} \left(\frac{\eta_i^{(s)} - \mu_i}{v_i} \right)^2 - \frac{1}{2v_i},$$

- 381 and with respect to α, β, γ : $\partial \mu_i/\partial \beta_j = X_{ij}\alpha_j$, $\partial \mu_i/\partial \alpha_j = X_{ij}\beta_j$, $\partial v_i/\partial \alpha_j = X_{ij}^2(1/\gamma_j + \beta_j^2 2\alpha_j\beta_j^2)$, 382 $\partial v_i/\partial \beta_j = 2X_{ij}^2\alpha_j(1-\alpha_j)\beta_j$ and $\partial v_i/\partial \gamma_j = -X_{ij}^2\alpha_j/\gamma_j^2$.
- 383 We can write the gradients using the chain rule, and efficiently compute them in parallel using matrix
- operations. For every stochastic vector $\epsilon^{(s)}$ we can evaluate log-likelihood vector \mathbf{f}_s , where $F_{is} = p(y_i \mid \eta_i^{(s)})$.
- 385 First we can take derivatives with respect to $\mathbb{E}_q[\theta]$, in short $\mathbf{g}_1 \triangleq \nabla_{\mathbb{E}[\theta]} \mathcal{L}$:

$$X^{\top} \left(\sum_{s=1}^{S} \epsilon_s \circ \mathbf{f}_s \circ \mathbf{v}^{-1/2} / S \right)$$

386 and with respect to $\mathbb{V}_q[\boldsymbol{\theta}]$, in short $\mathbf{g}_2 \triangleq \nabla_{\mathbb{V}[\boldsymbol{\theta}]} \mathcal{L}$:

$$(X \circ X)^{\top} \left(\sum_{s=1}^{S} \mathbf{f}_{s} \circ (\boldsymbol{\epsilon}_{s}^{2} - \mathbf{1}) \circ \mathbf{v}^{-1} / 2S \right).$$

387 Applying the chain rule:

$$\nabla_{\alpha} \mathcal{L} = \mathbf{g}_1 \circ \boldsymbol{\beta} + \mathbf{g}_2 \circ (\boldsymbol{\gamma}^{-1} + \boldsymbol{\beta}^2 - 2\boldsymbol{\alpha} \circ \boldsymbol{\beta}^2)$$

$$\nabla_{\beta} \mathcal{L} = \mathbf{g}_1 \circ \boldsymbol{\alpha} + 2\boldsymbol{g}_2 \circ \boldsymbol{\alpha} \circ (\mathbf{1} - \boldsymbol{\alpha}) \circ \boldsymbol{\beta}$$

$$abla_{\gamma}\mathcal{L} = \mathbf{g}_2 \circ oldsymbol{lpha} \circ oldsymbol{\gamma}^{-2}.$$

388 B. Matrix factorization

389 For matrix factorization $H = UV^{\top}$ we used group spike-slab prior¹⁸:

$$\mathbf{u}_k \sim z_k \mathcal{N} \big(\mathbf{0}, (\tau_k^{\mathsf{u}})^{-1} I \big) + (1 - z_k) \delta_0.$$

390 Under the variational approximation we define

$$\mathbb{E}_q[\mathbf{u}_k|\cdot] = \alpha_k \boldsymbol{\beta}_k,$$

$$\mathbb{V}_q[\mathbf{u}_k|\cdot] = \alpha_k \boldsymbol{\gamma}_k^{-1} + \alpha_k (1 - \alpha_k)(\boldsymbol{\beta}_k^2).$$

392 We can characterize

$$\mathbb{E}_q[H] = \mathbb{E}_q[U]\mathbb{E}_q[V]^\top$$

393 and

$$\mathbb{V}_q[H] = \mathbb{V}_q[U]\mathbb{V}_q[V]^\top + \mathbb{E}_q[U^2]\mathbb{V}_q[V^\top] + \mathbb{V}_q[U]\mathbb{E}_q[V^2].$$

Sampling stochastic matrix $E^{(s)}, E_{ij} \sim \mathcal{N}(0, 1)$, we evaluate

$$H^{(s)} \leftarrow \mathbb{E}_q[H] + E^{(s)} \circ \mathbb{V}_q[H]^{1/2}$$

395 and log-likelihood $F^{(s)} = \ln p(Y|H^{(s)},\cdot)$ and estimate stochastic components of gradients:

$$G_1 \triangleq \frac{1}{S} \sum_{s=1}^{S} F^{(s)} \circ E^{(s)} \circ V^{-1/2}$$

$$G_2 \triangleq \frac{1}{2S} \sum_{s=1}^{S} F^{(s)} \circ ((E^{(s)})^2 - 1) \circ V^{-1}$$

396 Using G_1 and G_2 we can derive gradient matrices straightforwardly

$$\nabla_{\mathbb{E}[U]} = G_1 \mathbb{E}_q[V] + 2(G_2 \mathbb{V}_q[V]) \circ \mathbb{E}_q[U]$$

$$\nabla_{\mathbb{V}[U]} = G_2(\mathbb{V}_q[V] + \mathbb{E}_q[V]^2)$$

$$\nabla_{\mathbb{E}[V]} = G_1^\top \mathbb{E}_q[U] + 2(G_2^\top \mathbb{V}_q[U]) \circ \mathbb{E}_q[V]$$

$$\nabla_{\mathbb{V}[V]} = G_2^{\top}(\mathbb{V}_q[U] + \mathbb{E}_q[U]^2)$$

397 On these partial gradients we applying chain rules with respect to the original variational parameters.

398 C. Control variate

Consider a generic variational approximation problem with surrogate distribution $q(\theta \mid \lambda)$ for the intractable distribution $f(\theta)$. We want to calculate the gradient with respect to the variational parameter λ :

$$\nabla_{\lambda} \mathbb{E}_{q}[f(\theta)] = \mathbb{E}_{q}[f(\theta)\nabla_{\lambda} \ln q(\theta \mid \lambda)]$$

Define $h(\theta) = f(\theta) \nabla_{\lambda} \ln q(\theta \mid \lambda)$ and introduce a control variate $g(\theta)$ such that $\mathbb{E}_q[g(\theta)] = 0$. Then, the gradient is equal to:

$$\mathbb{E}_{q}[h(\theta) - \xi q(\theta)]$$

403 We can the find optimal ξ^* by solving:

$$\xi^* = \min_{\xi} \mathbb{V}_q[h(\theta) - \xi g(\theta)]$$

$$= \min_{\boldsymbol{\xi}} \mathbb{V}_q[h(\boldsymbol{\theta})] - 2\boldsymbol{\xi} \mathsf{Cov}_q[h(\boldsymbol{\theta}), g(\boldsymbol{\theta})] + \boldsymbol{\xi}^2 \mathbb{V}_q[g(\boldsymbol{\theta})]$$

405 Previous work²¹ derived the optimal ξ :

404

407

408

$$\xi^* = \frac{\mathsf{Cov}_q[h(\theta), g(\theta)]}{\mathbb{V}_q[q(\theta)]}$$

406 Here, we define $g(\theta) = \nabla_{\lambda} \ln q(\theta \mid \lambda)$. Then, we have:

$$Cov[h(\theta), g(\theta)] = \mathbb{E}[f(\theta)g(\theta)^2]$$

$$\mathbb{V}[g(\theta)] = \mathbb{E}[g(\theta)^2]$$

$$\xi^* = \frac{\mathbb{E}[f(\theta)g(\theta)^2]}{\mathbb{E}[g(\theta)^2]} = \frac{\mathsf{Cov}[f(\theta),g(\theta)^2] + \mathbb{E}[f(\theta)]\mathbb{E}[g(\theta)^2]}{\mathbb{E}[g(\theta)^2]} = \mathbb{E}[f(\theta)] + \frac{\mathsf{Cov}[f(\theta),g(\theta)^2]}{\mathbb{V}[g(\theta)]}.$$

However, accurately estimating ξ^* requires many samples $\theta^{(s)} \sim q(\theta^{(s)} \mid \lambda)$. Instead, we use $\hat{\xi} = \mathbb{E}[f(\theta)]$, to resulting in a stochastic gradient estimator with variance:

$$\mathbb{V}[f(\theta)g(\theta) - \hat{\xi}g(\theta)] = \mathbb{V}[f(\theta)g(\theta)] - \mathbb{E}[f(\theta)]^2 \mathbb{V}[g(\theta)] - 2\mathbb{E}[f(\theta)] \mathsf{Cov}(f(\theta), g(\theta)^2).$$

411 D. Hyperparameter tuning

- 412 The spike and slab prior is defined by two hyperparameters, the prior inclusion probability π and prior effect
- 413 size precision τ . In our extensive simulation experiments, direct optimization over the hyperparameters
- 414 leads to degenerate solutions. To circumvent degeneracy, previous work used importance sampling or grid
- 415 search 13,38 .

417

416 Here, we included the hyperparameters in the variational surrogate distribution. We re-parameterized:

$$\pi = \boldsymbol{\sigma}(\delta_0)$$

$$\alpha_i = \boldsymbol{\sigma}(\delta_0 + \delta_i)$$

418 In our SVI updates, we updated δ_0 and δ_j simultaneously. Similarly, for the prior precision τ and variational 419 precision γ_j :

$$au =
u_{
m max} oldsymbol{\sigma}(
u_0)$$

420

$$\gamma_j = \nu_{\max} \boldsymbol{\sigma}(\nu_j - \nu_0)$$

The model fit is highly sensitive to the setting of $\nu_{\rm max}$. If the prior precision is too high, the prior dominates the likelihood and pulls effect sizes to zero. We set $\nu_{\rm max}=1000$.

123 IX. SUPPLEMENTARY FIGURES

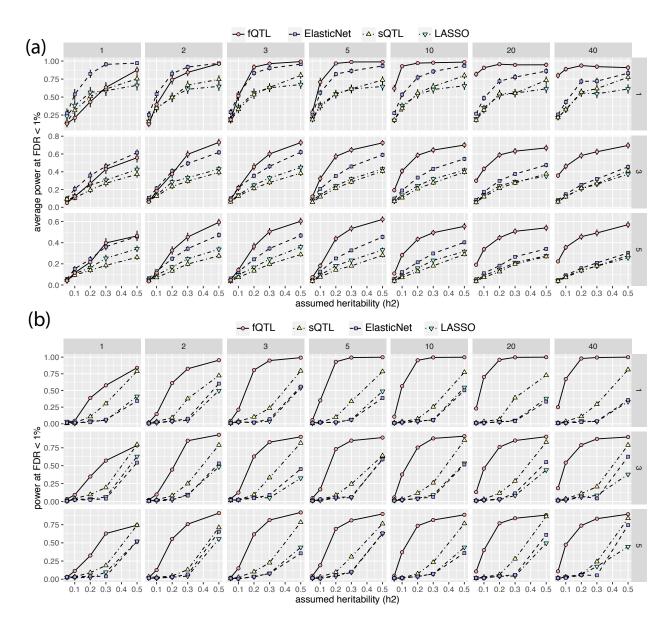


Figure 7: Power calculation of different methods on multi-tissue data with mis-specified rank K=2. Subpanels are arranged by different number of causal SNPs (rows; 1, 3, 5) and tissues (columns; 1 to 40). X-axis: assumed heritability. Y-axis: statistical power at FDR < 1%. (a) Power calculation of causal SNP prediction. Error bars indicate 95% confidence intervals of the mean values (2 standard error). (b) Power calculation of tissue of action prediction.

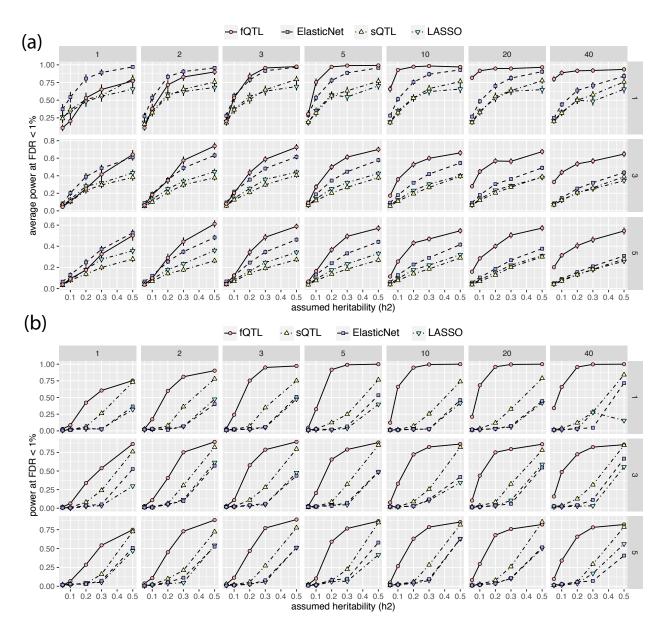


Figure 8: Power calculation of different methods on multi-tissue data with mis-specified rank K=3. Subpanels are arranged by different number of causal SNPs (rows; 1, 3, 5) and tissues (columns; 1 to 40). X-axis: assumed heritability. Y-axis: statistical power at FDR < 1%. (a) Power calculation of causal SNP prediction. Error bars indicate 95% confidence intervals of the mean values (2 standard error). (b) Power calculation of tissue of action prediction.

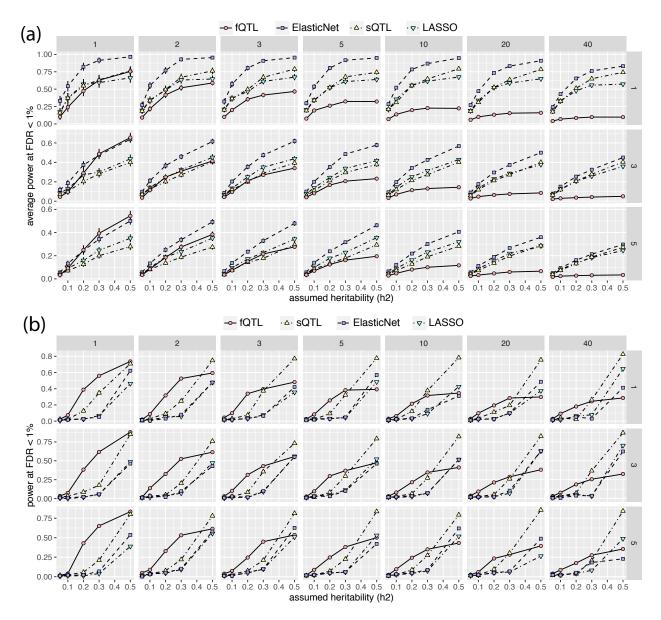


Figure 9: Power calculation of different methods on unstructured multi-tissue data. Subpanels are arranged by different number of causal SNPs (rows; 1, 3, 5) and tissues (columns; 1 to 40). X-axis: assumed heritability. Y-axis: statistical power at FDR < 1%. (a) Power calculation of causal SNP prediction. Error bars indicate 95% confidence intervals of the mean values (2 standard error). (b) Power calculation of tissue of action prediction.

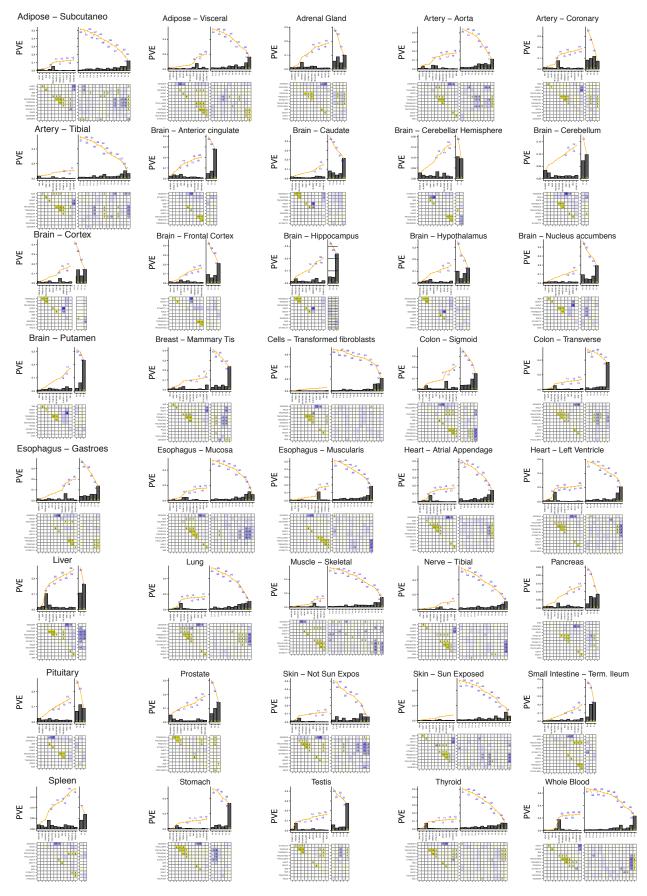


Figure 10: Association of tissue-specific hidden factors with known variables.

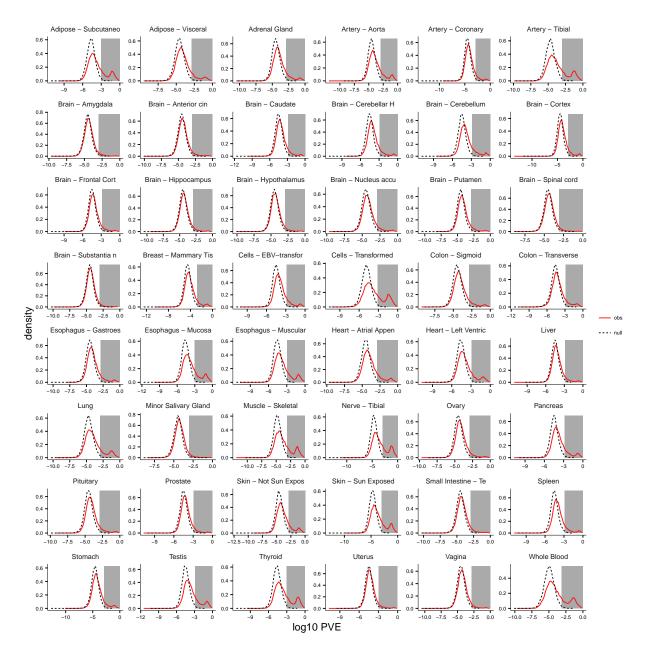


Figure 11: Tissue-specific null distribution sQTL PVE constructed by sample permutation.

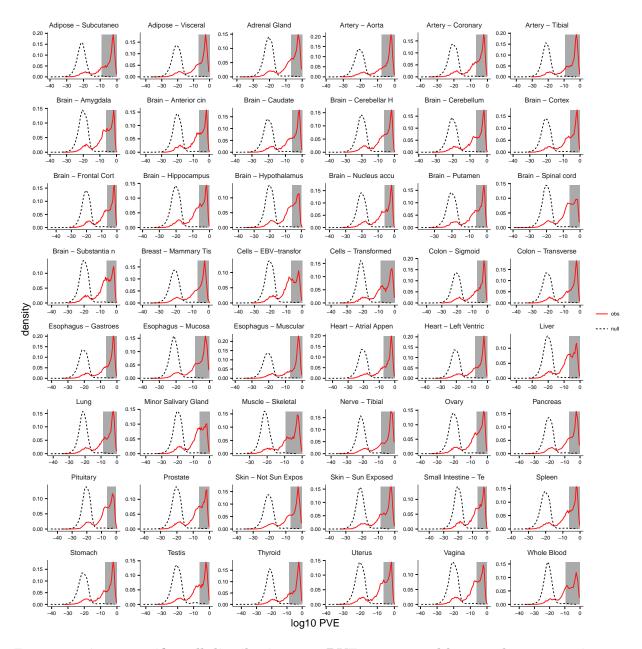


Figure 12: Tissue-specific null distribution fQTL PVE constructed by sample permutation.

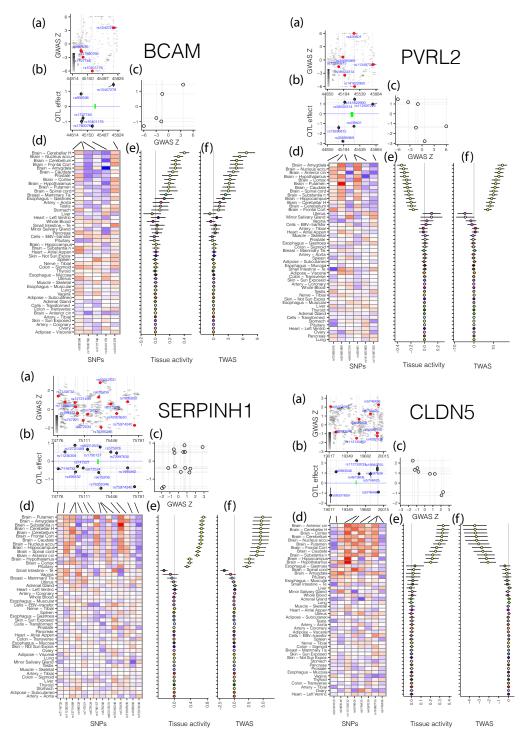


Figure 13: **Examples of brain-specific AD genes.** (a) Distribution of GWAS z-scores in the *cis*-region. (b) Polygenic QTL effect sizes in the same region. (c) Correlation between QTL and z-score vectors. (d) Marginal correlation matrix of SNP-tissue pairs. (e) Tissue activity effects found by fQTL model. (f) TWAS statistics. Errorbars denote 95% confidence intervals.

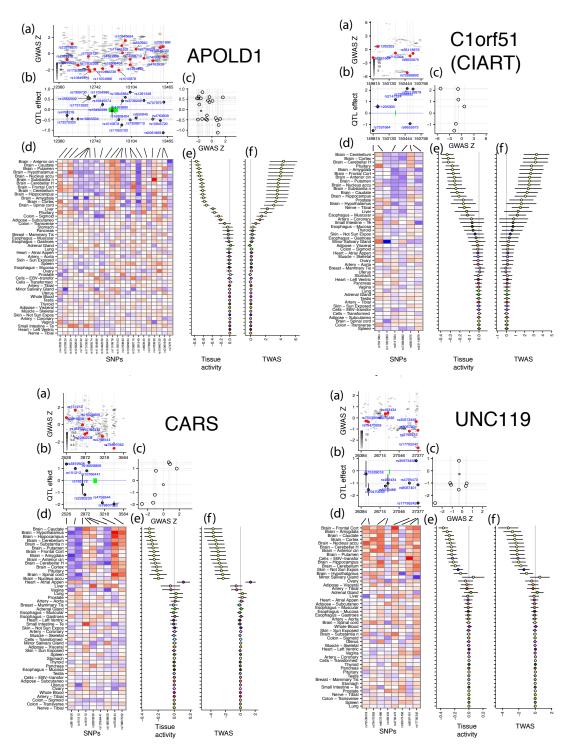


Figure 14: **Examples of brain-specific SCZ genes.** (a) Distribution of GWAS z-scores in the *cis*-region. (b) Polygenic QTL effect sizes in the same region. (c) Correlation between QTL and z-score vectors. (d) Marginal correlation matrix of SNP-tissue pairs. (e) Tissue activity effects found by fQTL model. (f) TWAS statistics. Errorbars denote 95% confidence intervals.

424 X. SUPPLEMENTARY TABLES

Full gzipped text files on TWAS results are made available at https://github.com/YPARK/FQTL/tree/master/GTEx.

427 XI. REFERENCES

- 1. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. Nature quentics 48, 245–252 (2016).
- ⁴³⁰ 2. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome ⁴³¹ data. *Nature genetics* **47**, 1091–1098 (2015).
- 432 3. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue 433 gene regulation in humans. *Science (New York, N.Y.)* **348**, 648–660 (2015).
- 434 4. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS genetics* **9**, e1003486 (2013).
- 436 5. Brown, C. D., Mangravite, L. M. & Engelhardt, B. E. Integrative Modeling of eQTLs and Cis-Regulatory Elements Suggests Mechanisms Underlying Cell Type Specificity of eQTLs. *PLoS genetics* **9**, e1003649 EP 438 (2013).
- 6. Segura, V. et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics 44, 825–830 (2012).
- ⁴⁴¹ 7. Kichaev, G. & Pasaniuc, B. AR TICLE Leveraging Functional-Annotation Data in Trans-ethnic Fine-⁴⁴² Mapping Studies. *The American Journal of Human Genetics* **97**, 260–271 (2015).
- 8. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS genetics* **10**, e1004383 EP (2014).
- 445 9. Hormozdiari, F. et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. bioRxiv 065037 446 (2016). doi:10.1101/065037
- 447 10. Mitchell, T. J. & Beauchamp, J. J. Bayesian variable selection in linear regression. *Journal of the American Statistical* ... **83**, 1023–1032 (1988).
- 449 11. George, E. I. & McCulloch, R. E. Approaches for Bayesian variable selection. *Statistica sinica* (1997). at 450 http://www.jstor.org/stable/24306083
- 451 12. Friedman, L. & Wall, M. Graphical Views of Suppression and Multicollinearity in Multiple Linear 452 Regression. *The American Statistician* **59**, 127–136 (2005).
- 453 13. Carbonetto, P. & Stephens, M. Scalable Variational Inference for Bayesian Variable Selection in Regression, 454 and Its Accuracy in Genetic Association Studies. *Bayesian Analysis* 7, 73–108 (2012).
- 455 14. Ročková, V. & George, E. I. Negotiating multicollinearity with spike-and-slab priors. *METRON* 72, 456 217–229 (2014).
- 457 15. Ghosh, J. & Ghattas, A. E. Bayesian Variable Selection Under Collinearity. *The American Statistician* 458 **69**, 165–173 (2015).
- 459 16. Guan, Y. & Stephens, M. Bayesian variable selection regression for genome-wide association studies and 460 other large-scale problems. *The Annals of Applied Statistics* **5**, 1780–1815 (2011).
- 461 17. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models.

- 462 PLoS genetics 9, e1003264 (2013).
- 463 18. Hernández-Lobato, D., Hernández-Lobato, J. M. & Dupont, P. Generalized Spike-and-Slab Priors for
- 464 Bayesian Group Feature Selection Using Expectation Propagation. Journal of Machine Learning Research
- 465 **14**, 1891–1945 (2013).
- 466 19. Hernández-Lobato, J. M., Hernández-Lobato, D. & Suárez, A. Expectation propagation in linear regression
- 467 models with spike-and-slab priors. Machine Learning 99, 437–487 (2015).
- 468 20. Carbonetto, P. & Stephens, M. Integrated enrichment analysis of variants and pathways in genome-wide
- 469 association studies indicates central role for IL-2 signaling genes in type 1 diabetes, and cytokine signaling
- 470 genes in Crohn's disease. PLoS genetics 9, e1003770 (2013).
- 471 21. Paisley, J., Blei, D. & Jordan, M. Variational Bayesian Inference with Stochastic Search. in Proceedings of
- 472 the 28th international conference on machine learning (eds. Langford, J. & Pineau, J.) 1367–1374 (Omnipress,
- 473 2012). at http://www.worldcat.org/title/proceedings-of-the-twenty-ninth-international-conference-on-machine-learning/
- 474 oclc/810280543>
- 475 22. Ranganath, R., Gerrish, S. & Blei, D. M. Black Box Variational Inference. in Proceedings of the 13th
- 476 international conference on artificial intelligence and statistics (eds. Kaski, S. & Corander, J.) 814-822
- 477 (2014). at http://jmlr.org/proceedings/papers/v33/ranganath14.pdf
- 478 23. Hoffman, M. D. et al. Stochastic variational inference. The Journal of Machine Learning Research 14,
- 479 1303–1347 (2013).
- 480 24. Wang, S. & Manning, C. Fast dropout training. Proceedings of the 30th International Conference on
- 481 Machine Learning 28, 118–126 (2013).
- 482 25. Kingma, D. P., Salimans, T. & Welling, M. Variational Dropout and the Local Reparameterization Trick.
- 483 arXiv.org (2015). at http://arxiv.org/abs/1506.02557v1
- 484 26. Brown, G. G. & Rutemiller, H. C. Means and Variances of Stochastic Vector Products with Applications
- 485 to Random Linear Models. Management Science 24, 210–216 (1977).
- 486 27. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait
- 487 Analysis. The American Journal of Human Genetics 88, 76–82 (2011).
- 488 28. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq
- 489 data with DESeq2. *Genome Biology* **15**, 1–21 (2014).
- 490 29. Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society.
- 491 Series B (Methodological) (1996). doi:10.2307/2346178
- 492 30. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via
- 493 Coordinate Descent. Journal of statistical software 33, 1–22 (2010).
- 494 31. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. Journal of the Royal
- 495 Statistical Society. Series B (Statistical Methodology) 67, 301–320 (2005).
- 496 32. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association
- 497 studies. Nature methods 11, 407–409 (2014).
- 498 33. Keilwagen, J., Grosse, I. & Grau, J. Area under Precision-Recall Curves for Weighted and Unweighted
- 499 Data. *PloS one* **9**, e92209 (2014).
- 500 34. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver
- 501 operating characteristic curves in R. Bioinformatics 31, 2595–2597 (2015).
- 502 35. Urbut, S. M., Wang, G. & Stephens, M. Flexible statistical methods for estimating and testing effects in

- ₅₀₃ genomic studies with multiple conditions. bioRxiv 096552 (2016). doi:10.1101/096552
- ⁵⁰⁴ 36. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for ⁵⁰⁵ Alzheimer's disease. *Nature genetics* **45**, 1452–1458 (2013).
- ⁵⁰⁶ 37. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 ⁵⁰⁷ (2014).
- 508 38. Stephens, M. False discovery rates: a new deal. *Biostatistics (Oxford, England)* kxw041 (2016). 509 doi:10.1093/biostatistics/kxw041
- ⁵¹⁰ 39. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic* ⁵¹¹ *Acids Research* **42**, D1001–D1006 (2014).
- 512 40. Abraham, R. *et al.* A genome-wide association study for late-onset Alzheimer's disease using DNA 513 pooling. *BMC medical genomics* **1**, 44 (2008).
- 514 41. Logue, M. W. et al. A Comprehensive Genetic Association Study of Alzheimer Disease in African 515 Americans. Archives of Neurology 68, 1569–1579 (2011).
- 516 42. Tanzi, R. E. The Genetics of Alzheimer Disease. Cold Spring Harbor Perspectives in Medicine 2, 517 a006296—a006296 (2012).
- 518 43. Pérez-Palma, E. et al. Overrepresentation of Glutamate Signaling in Alzheimer's Disease: Network-Based 519 Pathway Enrichment Using Meta-Analysis of Genome-Wide Association Studies. PloS one 9, e95413 (2014).
- 520 44. Yoo, Y. et al. Amyloid-Beta-Activated Human Microglial Cells Through ER-Resident Proteins. Journal 521 of Proteome Research 14, 214–223 (2015).
- 522 45. Pantazatos, S. P. *et al.* Whole-transcriptome brain expression and exon-usage profiling in major 523 depression and suicide: evidence for altered glial, endothelial and ATPase activity. *Molecular psychiatry* 524 (2016). doi:10.1038/mp.2016.130
- 525 46. Lee, P. et al. Effects of aging on blood brain barrier and matrix metalloproteases following controlled 526 cortical impact in mice. Experimental Neurology 234, 50–61 (2012).
- 527 47. Bergen, A. A., Kaing, S., Brink, J. B. ten, Gorgels, T. G. & Janssen, S. F. Gene expression and functional 528 annotation of human choroid plexus epithelium failure in Alzheimers disease. *BMC Genomics* **16**, 956 (2015).
- 529 48. Regard, J. B. et al. Verge: A Novel Vascular Early Response Gene. Journal of Neuroscience 24, 530 4092–4103 (2004).
- 531 49. Roszkowski, M. *et al.* Rapid stress-induced transcriptomic changes in the brain depend on beta-adrenergic 532 signaling. *Neuropharmacology* **107**, 329–338 (2016).
- 533 50. Chen, C.-Y. et al. Effects of aging on circadian patterns of gene expression in the human prefrontal 534 cortex. Proceedings of the National Academy of Sciences of the United States of America 113, 206–211 535 (2016).
- 536 51. Wulff, K., Dijk, D.-J., Middleton, B., Foster, R. G. & Joyce, E. M. Sleep and circadian rhythm disruption 537 in schizophrenia. *The British Journal of Psychiatry* **200**, 308–316 (2012).
- 538 52. Edenberg, H. J. et al. Genome-Wide Association Study of Alcohol Dependence Implicates a Region on Chromosome 11. Alcoholism: Clinical and Experimental Research 34, 840–852 (2010).
- 540 53. Treutlein, J. & Rietschel, M. Genome-Wide Association Studies of Alcohol Dependence and Substance 541 Use Disorders. *Current Psychiatry Reports* **13**, 147–155 (2011).
- 542 54. Räsänen, P. *et al.* Schizophrenia, alcohol abuse, and violent behavior: A 26-year followup study of an 543 unselected birth cohort. *Schizophrenia Bulletin* **24**, 437–441 (1998).
- 544 55. Alden, A., Brennan, P., Hodgins, S. & Mednick, S. Psychotic Disorders and Sex Offending in a Danish

- 546 56. Fazel, S., Långström, N., Hjern, A., Grann, M. & Lichtenstein, P. Schizophrenia, Substance Abuse, and 547 Violent Crime. *JAMA* **301**, 2016–2023 (2009).
- 548 57. Gorska, M. M., Liang, Q., Karim, Z. & Alam, R. Uncoordinated 119 Protein Controls Trafficking of Lck 549 via the Rab11 Endosome and Is Critical for Immunological Synapse Formation. *The Journal of Immunology* 550 **183**, 1675–1684 (2009).
- 551 58. Rodríguez-Fernández, J. L., Riol-Blanco, L. & Delgado-Martín, C. What is an immunological synapse? 552 *Microbes and Infection* **12**, 438–445 (2010).
- 553 59. Jones, A. L., Mowry, B. J., Pender, M. P. & Greer, J. M. Immune dysregulation and self-reactivity in 554 schizophrenia: Do some cases of schizophrenia have an autoimmune basis? *Immunology and Cell Biology* 83, 555 9–17 (2005).
- 556 60. Schwarz, G. Estimating the Dimension of a Model. The Annals of Statistics 6, 461–464 (1978).
- 557 61. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *International Conference on Learning Representations* (2014). at http://arxiv.org/abs/1312.6114>

