# Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression

Diego Calderon,[1,*] Anand Bhaskar,[2,3] David A. Knowles,[2,4] David Golan,[5] Towfique Raj,[6,7] Audrey Q. Fu,[8,10] and Jonathan K. Pritchard[2,3,9,10]

Previous studies have prioritized trait-relevant cell types by looking for an enrichment of genome-wide association study (GWAS) signal within functional regions. However, these studies are limited in cell resolution by the lack of functional annotations from difficult-to-characterize or rare cell populations. Measurement of single-cell gene expression has become a popular method for characterizing novel cell types, and yet limited work has linked single-cell RNA sequencing (RNA-seq) to phenotypes of interest. To address this deficiency, we present RolyPoly, a regression-based polygenic model that can prioritize trait-relevant cell types and genes from GWAS summary statistics and gene expression data. RolyPoly is designed to use expression data from either bulk tissue or single-cell RNA-seq. In this study, we demonstrated RolyPoly's accuracy through simulation and validated previously known tissue-trait associations. We discovered a significant association between microglia and late-onset Alzheimer disease and an association between schizophrenia and oligodendrocytes and replicating fetal cortical cells. Additionally, RolyPoly computes a trait-relevance score for each gene to reflect the importance of expression specific to a cell type. We found that differentially expressed genes in the prefrontal cortex of individuals with Alzheimer disease were significantly enriched with genes ranked highly by RolyPoly gene scores. Overall, our method represents a powerful framework for understanding the effect of common variants on cell types contributing to complex traits.

## Introduction

Identifying the primary subset of cell types or states and genes involved in complex traits is critical to the process of developing mechanistic insights. For example, knowledge that *FTO* (MIM: 610966) acts on *IRX3* (MIM: 612985) and *IRX5* (MIM: 606195) primarily in human adipocyte progenitor cells enabled researchers to rigorously define a novel thermogenesis pathway central for lipid storage and obesity.[1] And, focusing on distinct human human *C4A* (MIM: 120810) and *C4B* (MIM: 120820) isotypes, Sekar et al. highlighted the role of the classical complement cascade (of which *C4* genes are a critical component) and synapse elimination during development in the brains of individuals with schizophrenia.[2]

In addition to estimating disease risk for individual variants, genome-wide association studies (GWASs) have proven useful for identifying trait-relevant cell types or tissues. Assuming that variants affect phenotypes through gene regulation, one can prioritize cell types for further analysis with an enrichment of GWAS signal in cell-type-specific functional genomic regions that affect gene regulation. A series of studies have identified enrichment of GWAS signal in sorted cell-type-specific[3] or tissue-specific[4] expression quantitative trait loci (eQTLs). Other approaches (e.g., assay for transposase-accessible chromatin using sequencing [ATAC-seq], chromatin immunoprecipitation sequencing [ChIP-seq], and RNA sequencing [RNA-seq]) have revealed an enrichment of GWAS signal in cell-type-specific functional annotations.[5–11] However, these analyses are limited in cell-type resolution because they either require samples with population variation (which are infeasible to collect for many cell types) or rely on functional assays that require thousands of cells (which are challenging to collect for rare or uncharacterized cell types). Thus, it remains difficult to evaluate whether disease phenotypes are driven by tissues, broad cell populations, or very specific cell types. Furthermore, an inability to analyze difficult-to-characterize cell types is a concern when scanning for links between traits and cell types in complex tissues composed of many heterogeneous cell types. For example, describing the brain as the primary pathogenic tissue responsible for schizophrenia or Alzheimer disease (AD) is unsatisfying, but it remains difficult to comprehensively collect functional information from the plethora of brain cell types necessary for standard GWAS enrichment analyses.

Meanwhile, single-cell gene expression technology has offered insights into complex cell types.[12–21] Additionally, concerted efforts are underway for the development of comprehensive single-cell atlases of complex human tissues known to be associated with phenotypes of interest, such as immune cell types for autoimmune disease and brain cell types for neuropsychiatric disorders.[22] However, to our knowledge, no existing methods are designed to link novel single-cell-based cell types and phenotypes of interest.

[1]Program in Biomedical Informatics, Stanford University, Stanford, CA 94305, USA; [2]Department of Genetics, Stanford University, Stanford, CA 94305, USA; [3]Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA; [4]Department of Radiology, Stanford University, Stanford, CA 94305, USA; [5]Faculty of Industrial Engineering & Management, Technion, Haifa 3200003, Israel; [6]Department of Neuroscience, Mount Sinai School of Medicine, New York, NY 10029, USA; [7]Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA; [8]Department of Statistical Science, University of Idaho, Moscow, ID 83844, USA; [9]Department of Biology, Stanford University, Stanford, CA 94305, USA
[10]These authors contributed equally to this work
*Correspondence: dcal@stanford.edu
https://doi.org/10.1016/j.ajhg.2017.09.009.

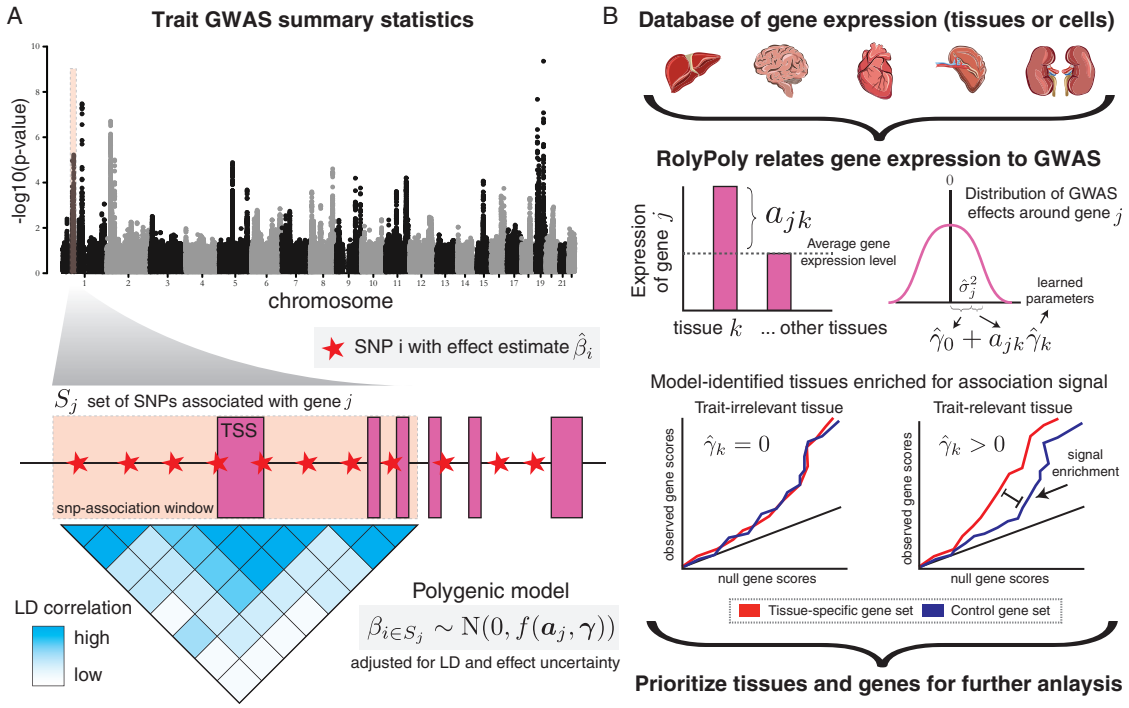


**Figure 1. RolyPoly Detects Trait-Associated Annotations by Using GWAS Summary Statistics and Gene Expression Profiles**
(A) We model the variance of GWAS effect sizes of SNPs associated with a gene as a function of gene annotations, in particular gene expression, while accounting for LD by using population-matched genotype correlation information. (Manhattan plot is based on data from Willer et al.[23])
(B) From a database of functional information (such as tissue or cell-type RNA-seq), we learn a regression coefficient ($\hat{\gamma}_k$) that captures each annotation's influence on the variance of GWAS effect sizes. A deviation from the mean gene expression value of $a_{jk}$ results in an increase of $a_{jk}\hat{\gamma}_k$ to the expected variance of gene-associated GWAS effect sizes. The value $\hat{\gamma}_0$ represents a regression intercept that estimates the population mean variance. To check learned model parameters, we expect to see an enrichment of LD-informed GWAS gene scores for genes that are specifically expressed in a tissue inferred to be trait relevant. Finally, from a model fit, we can prioritize trait-relevant tissues and genes.

Thus, we developed RolyPoly, a model for prioritizing trait-relevant cell types observed from single-cell gene expression assays. Importantly, RolyPoly takes advantage of polygenic signal by utilizing GWAS summary statistics for all SNPs near protein-coding genes, appropriately accounts for linkage disequilibrium (LD), and jointly analyzes gene expression from many tissues or cell types simultaneously. Additionally, our model can utilize signatures of cell-specific gene expression to prioritize trait-relevant genes. Finally, we provide a fast and publicly available implementation of the RolyPoly model.

## Material and Methods

### Overview of the Methods

The primary goals of RolyPoly are to identify and prioritize trait-relevant cell types (or tissues) and genes (Figure 1). Similar models have been developed to identify functional annotations important for complex traits.[7,11] However, unlike RolyPoly, these methods focus on SNPs rather than genes. They require binary input (e.g., whether or not a SNP is associated with a functional annotation) instead of quantitative measurements (such as gene expression). The most closely related method that focuses on genes lacks an underlying model and does not take advantage of the signal from SNPs that do not meet the stringent genome-wide significance threshold, potentially resulting in reduced power.[10] We decided to take a highly polygenic modeling approach to allow for the possibility that many genes might contribute to the trait.[24–26]

At a high level, RolyPoly starts by learning about the relationship between gene expression and estimated GWAS effect sizes from a trait of interest (captured with our $\gamma$ model parameters, described below). For example, we might expect to observe larger GWAS effect sizes for cholesterol regulation at SNPs that affect liver-specific gene expression because the liver is known to regulate cholesterol levels. Thus, on the basis of such an enrichment, RolyPoly would learn that the liver is a trait-relevant tissue. Next, we can use this knowledge to prioritize trait-relevant genes by calculating a score (represented by $h_j^{\mathrm{gene}}$, defined below) that identifies genes upregulated in RolyPoly-inferred relevant tissues. Continuing with our example, once we know that liver-specific gene expression is associated with larger GWAS effect sizes, RolyPoly would prioritize studying liver-specific genes in the context of understanding cholesterol regulation (resulting in larger $h_j^{\mathrm{gene}}$ values). Below, we describe the details of how RolyPoly carries out each of these steps.

### GWAS Summary Statistics

Consider a fully polygenic GWAS model, $y_s = \boldsymbol{x}_s^T\boldsymbol{\beta} + \epsilon_s$, where $y_s$ is the phenotypic measurement from individual $s$, $\boldsymbol{x}_s$ is a vector of genotypes at $p$ SNPs for individual $s$, $\boldsymbol{\beta}$ is a vector of $p$ SNP effects,

and $\epsilon_s \sim N(0, \sigma_e^2)$ represents the stochastic environmental error. Importantly, we assumed that the matrix of genotypes had been scaled and standardized such that the mean was 0 and variance was 1 for each SNP vector (and similarly for the trait $\boldsymbol{y}_s$). As others have pointed out,[27] this implies a biologically plausible and computationally convenient relationship between allele frequency and variance of effect sizes. Because this relationship was not the focus of our study, we left the goal of generalizing this relationship to future work. The main summary statistics released by GWASs are per-variant effect estimates, which we refer to as $\widehat{\boldsymbol{\beta}}$. Researchers typically calculate and report univariate effect-size estimates. These estimates represent the marginal standardized regression coefficient and are calculated as $\widehat{\beta}_i = n^{-1}\boldsymbol{X}_i^T\boldsymbol{y}$, where $\boldsymbol{X}_i$ (note the change in case) represents standardized genotypes for SNP $i$ across $n$ individuals (see Appendix A for derivation). If we substitute the polygenic model for $\boldsymbol{y}$ into the estimation equation (see Appendix A for derivation), the sampling distribution of the estimated SNP effect sizes corresponds to

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{R}\boldsymbol{\beta} + n^{-1}\boldsymbol{X}^T\epsilon, \qquad \text{(Equation 1)}$$

where $\boldsymbol{R}$ is the sample LD matrix (i.e., $r_{ii'}$ is the Pearson correlation values between genotype $i$ and $i'$). We used this definition of estimated GWAS effect sizes to develop a highly polygenic approach that models the variance of these SNP effect sizes as a function of annotation specificity of proximal gene expression.

## Polygenic Model

For notational convenience, let $g(i)$ represent the gene associated with SNP $i$ and $S_j = \{i: g(i) = j\}$ be the set of SNPs associated with gene $j$. We use the notation $\boldsymbol{\beta}_S$ to denote the $\boldsymbol{\beta}$ coordinates whose indices lie in set $S$. We assumed *a priori* that the true GWAS effect sizes of SNPs in gene $j$ follow the normal distribution $\boldsymbol{\beta}_{S_j} \sim \text{MVN}(0, \tau_j I)$, where $I$ is the $|S_j| \times |S_j|$ identity matrix and $\tau_j$ is the prior effect-size variance for all SNPs associated with gene $j$ and is modeled as a linear function. More specifically, $\tau_j$ is a linear function of $N$ annotations $a_{jk}$ (in this case, cell-type-specific gene expression) with annotation coefficients $\gamma_k$ and an intercept term $\gamma_0$:

$$\tau_j = \gamma_0 + \sum_{k=1}^{N} \gamma_k a_{jk}. \qquad \text{(Equation 2)}$$

RolyPoly estimates the parameter vector $\boldsymbol{\gamma}$, which captures the influence of cell-type-specific gene expression on the variance of GWAS effect sizes (see Figure 1B). Intuitively, if we estimate a large coefficient for annotation $k$, then we expect larger GWAS effect sizes around genes that are specifically expressed in annotation $k$. On the other hand, it is possible to estimate negative values for some annotation coefficients $\gamma$. SNPs proximal to genes that are specifically expressed in an annotation with a negative $\gamma$ estimate are expected to have lower effect-size variance than the population mean.

According to this polygenic model, the expected value of the vector of GWAS effect sizes around gene $j$ is $\mathbb{E}[\widehat{\boldsymbol{\beta}}_{S_j}] = \boldsymbol{0}$, and the covariance matrix is given by $\mathbb{V}[\widehat{\boldsymbol{\beta}}_{S_j}] = \tau_j \boldsymbol{R}_{S_j}\boldsymbol{R}_{S_j} + \sigma_e^2 n^{-1}\boldsymbol{R}_{S_j}$, where $\boldsymbol{R}_{S_j}$ denotes the principal $\boldsymbol{R}$ submatrix indexed by the SNPs in $S_j$ (see Appendix A for derivation). This model assumes that the effect size of each SNP around a gene $j$ is drawn from a distribution with a mean of 0 and the same per-SNP variance of $\tau_j$. However, we expected other SNP annotations to affect the variance of a GWAS effect size, such as the minor allele frequency (MAF) of the SNP. Therefore, we included $P$ SNP-level features as covariates while estimating the variance contribution of gene expression. Specifically, we modified our model to use a per-SNP variance $\nu_i$ for SNP $i$, given by

$$\nu_i = \tau_{g(i)} + \sum_{l=1}^{P} \phi_l b_{il}, \qquad \text{(Equation 3)}$$

where $\tau_{g(i)}$ is the previously described (Equation 2) contribution of gene-level annotations to the variance of SNP $i$, $b_{il}$ is the $i^{\text{th}}$ value of SNP-level annotation $l$ for SNP $i$, and $\phi_l$ is the annotation coefficient for annotation $l$. The distribution for the vector of SNP effects associated with a gene becomes

$$\widehat{\boldsymbol{\beta}}_{S_j} \sim \text{MVN}\left(0, \boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j} + \sigma_e^2 n^{-1}\boldsymbol{R}_{S_j}\right), \qquad \text{(Equation 4)}$$

where $\boldsymbol{D} = \text{diag}(\boldsymbol{\nu})$ is a diagonal matrix of SNP effect-size variances. With this modification, we can estimate gene annotation regression coefficients while controlling for the contribution of SNP annotations to the variance of a SNP effect size. We present inferred parameter estimates, including accounting for MAF as a SNP-level covariate. MAF values were downloaded from matched population samples from the phase 3 VCFs of the 1000 Genomes Project.[28]

We chose a 10 kb window centered on the transcription start site (TSS) of a gene because previous work has found that, across a diverse set of cell types and tissues, most eQTLs consistently lie in this region.[29–32] During initial analysis, we observed similar parameter estimates and strength of associations by using a larger window size of 20 kb (Figure S1). Thus, for computational convenience, we used a smaller 10 kb window size for results in this study. However, the model description as presented generalizes to larger window sizes or alternative approaches of SNP-gene association. One could rely on enhancer or chromatin maps from ENCODE to incorporate potentially functional variants that are farther away from the TSS. However, doing so would bias our analysis toward well-characterized cell types; thus, we did not include distal elements. With this definition of SNP-gene association, there are a few SNPs with multiple associated genes. We duplicated these SNPs and treated them as independent SNP-gene pairs. Because RolyPoly infers parameters from hundreds of thousands of SNPs, we do not expect this to contribute significantly to inferred parameters.

## Parameter Inference

To perform maximum-likelihood inference under our model, we would have to compute the determinant and inverse of the potentially high-dimensional covariance matrices involved in Equation 4, which would be computationally challenging. Instead, we adopted a method-of-moments approach in which we fit the gene-level annotation coefficients $\gamma_k$ and, if included, the SNP-level annotation coefficients $\phi_l$. Our model allows for explicit formulas to be derived for the method-of-moments estimators, which also greatly improves computational efficiency and avoids convergence concerns that are common with other inference approaches, such as expectation-maximization. If only gene-level annotations are used, we fit the observed and expected sum of squared SNP effect sizes associated with each gene, where the expected value is given by

$$\mathbb{E}\left[\sum_{i \in S_j} \widehat{\beta}_i^2\right] = \tau_j \text{Tr}\left(\boldsymbol{R}_{S_j}^2\right) + |S_j| \sigma_e^2 n^{-1}, \qquad \text{(Equation 5)}$$

where Tr represents the trace of a matrix (derivation in Appendix A). We derived this expectation while recognizing that the expected value of the squared $\ell_2$ norm of a mean 0 multivariate normal distribution is the trace of the covariance matrix. When we included SNP annotation coefficients such that each SNP effect size had a variance term $v_i$, we performed inference by fitting the observed and expected squared effect size of each SNP, where the expected value is given by

$$\mathbb{E}\left[\hat{\beta}_i^2\right] = \left(\boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j}\right)_{ii} + \sigma_e^2 n^{-1}, \qquad \text{(Equation 6)}$$

where $j = g(i)$ and $(\boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j})_{ii}$ is the diagonal element of the matrix corresponding to SNP $i$. Interestingly, by using an indicator function rather than quantitative features, this model relates to previous work[33] (described in the Appendix A). We performed block bootstrap[34] to estimate standard errors, $\hat{\sigma}_{\gamma_k}$, which we used to compute a $t$-statistic, $\hat{\gamma}_k/\hat{\sigma}_{\gamma_k}$, and corresponding p values. We used a $t$-statistic because we used our bootstrap estimate of the standard error rather than a known value. The purpose of the block bootstrap is to maintain correlations present in the data when sampling from the empirical distribution; thus, we partitioned the genome into 100 non-overlapping blocks and sampled from these blocks with replacement.[35] Additionally, from the bootstrap parameter estimates, we calculated empirical 95% confidence intervals for each $\hat{\gamma}_k$. Unless otherwise specified, for our analyses we performed $10^3$ block bootstrap iterations. After including an intercept term, $\hat{\gamma}_0$, we ranked tissues by strength of association with the $t$-statistic or corresponding p value. As in standard regression, the intercept term estimates the population mean of the response term, which in this case is the per-SNP variance of a GWAS effect size.

## Computing Trait-Relevance Gene Importance Scores and the Proportion of Variance Explained by Individual Annotations

Using a set of inferred gene annotation coefficients, $\hat{\boldsymbol{\gamma}}$, we calculated several quantities that summarize the contributions of gene annotations to the phenotypic variance. First, we computed $h_j^{\text{gene}} = \sum_{k=1}^{N} \hat{\gamma}_k a_{jk}$, which can be used to rank trait-relevant genes. Essentially, $h_j^{\text{gene}}$ is a gene-expression-based prediction of the variance parameter for gene $j$ of a normal distribution from which *cis*-GWAS effect sizes are drawn (Figure 1B). Thus, if $h_j^{\text{gene}}$ is large, we would expect larger *cis*-GWAS effect sizes. Note that this value does not directly rely on GWAS effect-size estimates. Instead, $h_j^{\text{gene}}$ relies on GWASs indirectly through the RolyPoly-inferred parameters. Additionally, we calculated the contribution of an annotation $k$ to a trait as $h_k^{\text{annot}} = |\hat{\gamma}_k| \sum_{j=1}^{M} a_{jk}$, where $M$ is the number of genes. Through simulation, we showed that the true value of $h_k^{\text{annot}}$ affects our power to detect trait-annotation associations. The total contribution explained by all annotations, $h^{\text{total}}$, came from summing the individual annotation values, $h^{\text{total}} = \sum_{k=1}^{N} h_k^{\text{annot}}$. Finally, the proportion of an annotation's unique contribution to the variance of SNP effects, $p_k^{\text{annot}}$, was calculated as $h_k^{\text{annot}}/h^{\text{total}}$.

To validate our gene importance values, $h_j^{\text{gene}}$, we compared them with gene importance estimates based on *cis*-GWAS summary statistics and LD information. This gene score is an estimate of the variance of GWAS effect sizes after inflation due to local LD is accounted for; thus, we refer to it as the LD-informed gene score. For this calculation, we used the methodology described in Lamparter et al.[36] and Liu et al.[37] However, we used the same window size around a gene as was used for RolyPoly. In addition to validating $h_j^{\text{gene}}$, we used the LD-informed gene score to verify GWAS enrichment in specifically expressed genes of model-identified trait-relevant tissues (i.e., quantile-quantile [Q-Q] plots in the Results).

If the main objective is to compute gene values, $h_j^{\text{gene}}$, and unbiased parameter estimates are not required, then we include a penalty on the $\ell_1$ norm of the annotation coefficients. The penalty strength is modulated with a $\lambda$ tuning factor, which is chosen on the basis of cross validation. Regularization has the beneficial effect of shrinking parameter estimates of irrelevant tissues and can result in more accurate gene-score prediction.

## Simulation Setup

For clarity, we denote generated parameters and data with an asterisk. In simulation results reported, we used $2 \times 10^4$ genes, five simulated gene annotations, and one simulated SNP annotation. We generated gene expression, $a^*$, from a standard $\chi^2$ distribution and generated allele frequency as an example SNP annotation, $b^*$, from a standard uniform distribution. Recall that our model annotation coefficients determine the influence that these annotations will have on SNP effect sizes. For each simulated dataset, we fixed annotation effects by sampling from a uniform distribution: $\phi^* \sim \text{uniform}(0, 10^{-5})$ for SNP annotation effects and $\gamma_k^* \sim \text{uniform}(0, 10^{-5})$ for gene annotation effects. We combined the simulated functional information and annotation coefficients to calculate a per-SNP variance term. Thus, for each SNP effect, we computed $v_i^* = \tau_{g(i)}^* + \sum_{l=1}^{P} \phi_l^* b_{il}^*$, where $\tau_{g(i)}^* = \sum_{k=1}^{N} \gamma_k^* a_{jk}^*$. We combined this per-SNP variance term with a per-SNP environmental-error contribution set to $\sigma_e^2 n^{-1} = 10^{-4}$ to arrive at the distribution from which we generated simulated effects,

$$\hat{\boldsymbol{\beta}}_{S_j}^* \sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}^*\boldsymbol{R}_{S_j} + 10^{-4}\boldsymbol{R}_{S_j}\right), \qquad \text{(Equation 7)}$$

where $\boldsymbol{D}^*$ is a diagonal matrix with simulated per-SNP variance values. From this distribution, for each simulated gene, we sampled 20 SNP effects. As input, our inference model takes SNP effects, environmental errors (here set to $10^{-4}$), and annotations and attempts to identify the true annotation effects. From this setup, we determined whether our method implementation could accurately infer generated SNP annotation effects, $\phi_l^*$, and gene annotation effects, $\gamma_k^*$.

Although our method assumes that each SNP effect size is drawn from the model distribution, it is likely that some GWAS effect sizes come from a null distribution. To test robustness to this potential model misspecification, we first sampled per-gene Bernoulli random variables, $\pi_j \sim \text{Bernoulli}(c)$, where $c$ represents the fraction of causal genes (causal here simply implies sampling from the non-null model). We sampled SNP effects for each gene as

$$\hat{\boldsymbol{\beta}}_{S_j}^* \sim \begin{cases} \text{MVN}\left(0, \boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}^*\boldsymbol{R}_{S_j} + 10^{-4}\boldsymbol{R}_{S_j}\right), & \text{if } \pi_j = 1 \\ \text{MVN}\left(0, 10^{-4}\boldsymbol{R}_{S_j}\right), & \text{if } \pi_j = 0. \end{cases} \qquad \text{(Equation 8)}$$

Varying the fraction of causal genes (parameter $c$) across simulated datasets, we studied its effect on model inference.

## Obtaining Gene Expression Databases and GWAS Summary Statistics

We estimated annotation parameters for three gene expression databases. (1) The Genotype-Tissue Expression (GTEx) cohort includes RNA-seq from different individuals at many tissue sites.[38] (2) We downloaded single-cell RNA-seq data on 3,005 single cells from the hippocampus and cerebral cortex of mice from Ziesel et al.[20] (3) We obtained human single-cell RNA-seq data on cortex

samples from Darmanis et al.[39] Within each gene expression database, we standardized the distribution of gene expression across samples with quantile normalization. Expression samples from the same tissue or purified cell population were averaged. Single-cell RNA-seq is notoriously noisy and sparse.[40,41] Thus, to reduce the impact of these effects, we averaged single-cell expression vectors for common previously defined cell-type classes. We did not systematically study how best to cluster single cells into cell types given that there is already substantial literature on the subject.[15,42–45] To compare across genes, we scaled, centered, and then squared expression values across annotations. When using an expression database from mice, we used only orthologous protein-coding genes with a one-to-one functional mapping (based on the definition in Ensembl's BioMart[46]).

We downloaded publicly available GWAS summary statistics from ten traits from their respective publications: schizophrenia,[47] late-onset AD,[48] four metabolic traits (high-density lipoprotein [HDL] cholesterol, low-density lipoprotein [LDL] cholesterol, total cholesterol [TC], and triglyceride levels [TG]),[23] educational attainment (EA),[49] height,[50] extreme body mass index (BMI),[16] and age-related cognitive decline (ACD).[15] We restricted our analysis to the autosomes, removed the major histocompatibility complex (MHC) region for immune traits (chr6: 25–34 Mb), and removed rarer variants (MAF $< 1\%$). We removed rare variants to ensure that the common variant model fit the data appropriately. For late-onset AD and ACD, in addition to using the entire set of GWAS summary statistics, we ran RolyPoly after removing variants from a 1 Mb window centered on the TSS of *APOE* (MIM: 107741; chr19: 44,909,011–45,909,011). All referenced genome coordinates are from UCSC Genome Browser build hg19.

### Differential Gene Expression Analysis

For the analysis of $h_j^{\text{gene}}$ enrichment in differentially expressed (DE) genes of individuals with AD, we downloaded microarray gene expression data from 230 samples of the prefrontal cortex.[51] We used Limma to analyze differential gene expression between case and control tissues.[52] We mapped probes to genes by using a mapping downloaded from Ensembl's BioMart.[46] If multiple probes mapped to a single gene, we took the median expression value across all probes. Unless otherwise specified, we performed Kolmogorov-Smirnov significance tests of gene-value enrichment within DE genes.

### Calculating RolyPoly Gene-Score Enrichment while Accounting for Correlations among Gene Expression Values

To assess the enrichment of RolyPoly gene scores among DE genes, we calculated Spearman's rank-correlation coefficient, $\rho_{\text{obs}}$, between RolyPoly gene scores and differential-expression *t*-statistics. A large $\rho_{\text{obs}}$ value indicates enrichment of large RolyPoly gene scores among DE genes. Assessing the significance of $\rho_{\text{obs}}$ by independently considering each gene is anti-conservative because of the correlation between gene expression levels of co-regulated genes. Therefore, we generated an empirical sampling distribution for $\rho$ under the null of no association between RolyPoly scores and *t*, which accounted for gene expression correlation.

We estimated the variance-covariance matrix of gene expression in healthy individuals, $\Sigma$. Because there were fewer samples than genes, we used singular value decomposition (SVD) to represent the low-rank $\Sigma$ matrix. Under the null hypothesis, we generated a gene expression matrix for both case and control samples by using the same distribution, $\boldsymbol{X}_i \sim \text{MVN}(0, \Sigma)$. We had two sets of individuals: the set of healthy control individuals, *H*, and the set of affected individuals, *A* (of equal size to the true data). For each gene *j*, we computed a *t*-statistic by testing the difference between the means of the healthy and affected simulated expression values,

$$t_j = \frac{\bar{x}_j^A - \bar{x}_j^H}{\sqrt{\frac{s_j^A}{n^A} + \frac{s_j^H}{n^H}}}, \qquad \text{(Equation 9)}$$

where $\bar{x}_j$ is the mean expression of gene *j*, $s_j$ is the sample variance, and *n* is the sample size. We computed Spearman's correlation coefficient, $\rho_{\text{sim}}$, between $t_j$ and $h_j^{\text{gene}}$. We repeated the process of generating expression and calculating $\rho_{\text{sim}}$ $10^3$ times to generate a null distribution, which we then used to evaluate the significance of $\rho_{\text{obs}}$.

### Calculating LD Correlation Values

We downloaded phase 3 VCFs of European individuals from the 1000 Genomes Project.[28] We used PLINK v.1.90b1b to calculate Pearson's r values of SNPs within the default 1 Mb window.[53]

### RolyPoly Implementation and Usage

We implemented our method for use through the `rolypoly` R package, which is freely available and open source via CRAN and at our Git repository (see Web Resources).

## Results

### Simulation

We used simulations (see Material and Methods) to verify our implementation of RolyPoly and characterize properties of parameter estimation and hypothesis testing.

Across 500 data simulations, we found that RolyPoly-inferred $\hat{\gamma}_k$ parameters were unbiased estimates of the true underlying effect $\gamma_k^*$ (see Figure 2A). This is an important property if we aim to accurately quantify the total contribution of an annotation to a trait, $h_k^{\text{annot}} \cdot h_k^{\text{annot}}$ summarizes the amount of signal present in the dataset to detect an association between the trait and annotation *k*. In particular, our power is strongly dependent on $h_k^{\text{annot}}$ (see Figure 2B), where power refers to the probability that we correctly reject the null hypothesis (i.e., $\hat{\gamma} < 0$). It is likely that some fraction of GWAS effect sizes are drawn from a null distribution, which we do not currently model in RolyPoly. Thus, we investigated the effect of varying the fraction of GWAS effects drawn from the model distribution and our power to detect significant annotations. As expected, when the fraction of genes simulated from the causal distribution decreases, we lose power (see Figure 2B). However, even with 25% of genes (and downstream GWAS effect sizes) drawn from the causal distribution, we achieve greater than 50% power for an annotation with $h_k^{\text{annot}} \approx 0.15$. For context, in real data, we consistently observed $h_k^{\text{annot}}$ values greater than 0.1.

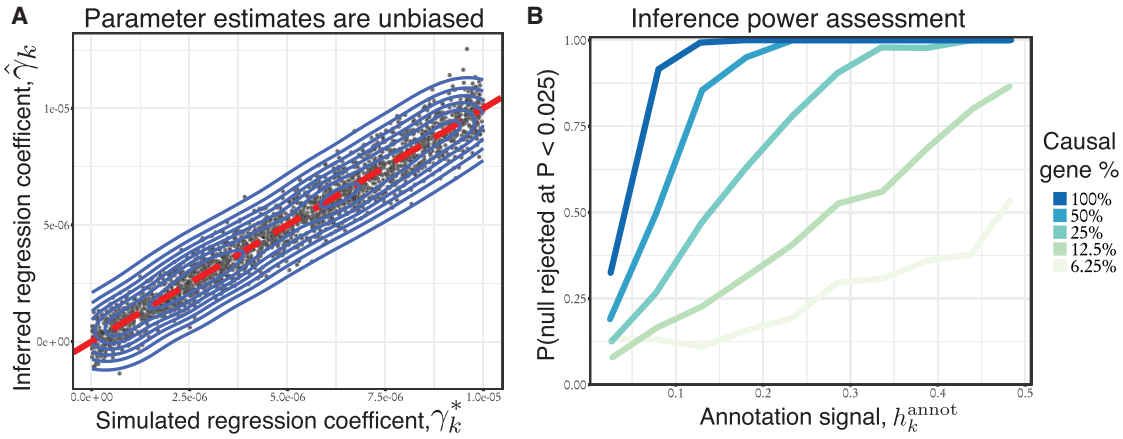


**Figure 2. Simulation Results**
(A) Parameter inference is unbiased and accurate for a range of simulated $\gamma^*$ effects. The red-dashed line represents the identity line. (B) Power as a function of the $\gamma_k^*$ and annotation values defined as $h_k^{annot}$ in the Material and Methods. Even when some SNPs are drawn from the null distribution, we maintain reasonable power to detect associations.

For data generated under the model, we demonstrated that our estimated parameters are unbiased and have low levels of deviation around the true parameter values. Our power to detect significant annotations is modulated by the annotation effect, the annotation values, and the fraction of effects drawn from the model distribution. Furthermore, in the setting where the effects are simulated from a mixture of the model and null distribution, we still have power to detect significant annotations.

### Trait-Relevant Tissues Identified from GTEx Data

As a proof of principle, we ran our method on trait-association data from publicly available GWAS traits and gene expression data from 27 tissues of 544 individuals from the GTEx project (data download and processing are described in the Material and Methods).

In Table 1, we summarize the top two tissue-trait associations that passed a marginal significance threshold ($p < 0.05$) for seven GWAS traits. With a GWAS of extreme BMI, we found associations with kidney ($p = 7 \times 10^{-3}$) and thyroid ($p = 0.03$) gene expression. Obesity is known to negatively affect kidney function; however, from existing literature, it is ambiguous whether the tissue has a causal role in determining BMI.[55] Some studies have demonstrated a correlation between thyroid function and weight.[56,57] Others have observed a link between BMI and brain gene expression.[58,59] We suspect that we did not observe such an association because we composed the GTEx brain annotation by aggregating gene expression across all 13 collected brain tissues (this was the most heterogeneous tissue annotation collected). We observed a significant enrichment of EA signal for genes specifically expressed in the pituitary gland ($p = 0.03$) and brain ($p = 0.04$), which corresponds with recent analysis.[49,60] For height, we detected an association with muscle ($p = 6 \times 10^{-10}$) and pituitary ($p = 6 \times 10^{-7}$). Interestingly, tumors in the pituitary are known to lead to gigantism, characterized by excessive growth and height.[61] Finally, for several metabolic traits (TC, LDL, TG, and HDL), we detected signals for the liver, small intestine, and adrenal gland, all of which follow known biology.

Next, we examined the TC GWAS,[23] given that its association with liver has been unambiguously reported in the literature. For inference, we used 121,312 SNPs that were within 5 kb of a protein-coding gene. With p values from our model, we ranked tissues by the strength of association with TC (Figure 3, left). As expected, liver was the clear top-associated annotation ($p = 2 \times 10^{-4}$), and we estimated an annotation coefficient of $4 \times 10^{-6}$ (Figure 3, right). Thus, we estimated that the variance of TSS-proximal GWAS effect sizes increases by $4 \times 10^{-6}$ as normalized gene expression in the liver increases by one unit (see Material and Methods for a description of gene expression normalization). The small intestine was marginally associated ($p = 0.01$), which follows from the fact that this organ has a central role in nutrient absorption. Additionally, we observed some signal for spleen ($p = 0.04$) and adrenal gland ($p = 0.05$). Although the spleen is primarily thought of as an immune organ, studies have shown a clear link between splenectomy and lipid metabolism.[62] Whereas the p value for adrenal gland was identified with a *q*-value of 0.3, the 95% confidence interval showed a wide distribution of non-zero parameter estimates of large positive effect. Considering that the adrenal gland plays a central role in the production of hormones (many of which are synthesized from cholesterol or even have an effect on cholesterol levels), this association is biologically plausible.[63]

We wanted to verify that GWAS effect sizes were enriched with association signal near genes that were specifically expressed in tissues with RolyPoly annotation coefficients significantly greater than 0. First, we calculated LD-informed gene scores, which estimate the variance of GWAS effect sizes from a *cis* window around each gene while accounting for LD (see Material and Methods). Next, we visualized the enrichment of these scores in specifically expressed gene sets by using Q-Q plots (Figure 4). To define the set of tissue-specific genes, we sorted

**Table 1. Top Trait-Relevant GTEx Tissues for Seven GWAS Traits and Uncorrected p Values**

| Trait | Tissue | p Value |
|---|---|---|
| Height | muscle | $6 \times 10^{-10}$ |
| Height | pituitary | $6 \times 10^{-7}$ |
| TC | liver | $2 \times 10^{-4}$ |
| TC | small intestine | $1 \times 10^{-2}$ |
| LDL | liver | $2 \times 10^{-3}$ |
| LDL | small intestine | $2 \times 10^{-2}$ |
| TG | adrenal gland | $7 \times 10^{-7}$ |
| TG | liver | $2 \times 10^{-2}$ |
| BMI | kidney | $7 \times 10^{-3}$ |
| BMI | thyroid | $3 \times 10^{-2}$ |
| HDL | liver | $7 \times 10^{-3}$ |
| EA | pituitary | $3 \times 10^{-2}$ |
| EA | brain | $4 \times 10^{-2}$ |

Corresponding GWAS abbreviations are as follows: TC, total cholesterol;[23] LDL, LDL cholesterol;[23] TG, triglyceride levels;[23] BMI, extreme body mass index;[54] HDL, HDL cholesterol;[23] and EA, educational attainment.[49]

normalized expression values for the tissue of interest and identified the top 20% of genes as the tissue-specific gene set. Correspondingly, we refer to the bottom 20% of genes sorted by expression as the control set. We observed clear enrichment of TC *cis*-GWAS signal within the set of genes that were upregulated in the liver (Figure 4A). As a negative control, we employed the same Q-Q plot approach to determine whether there was GWAS signal around genes specifically expressed in a tissue not found to contribute significantly to TC. Within specifically expressed genes of the skin tissue (Figure 4B), we did not observe an enrichment of GWAS signal.

## Neuropsychiatric Diseases and Single-Cell Gene Expression

We next analyzed cell types identified from publicly available single-cell expression data from the human brain[39] and several neuropsychiatric traits: ACD, late-onset AD, EA, and schizophrenia. In total, we used 477 human single cells from which gene expression data had been collected. Using a principal-component analysis (PCA)-based clustering approach, the original authors grouped the single cells into six primary cell types and two clusters of fetal cortical cells representing quiescent and replicating cell states. For each gene, we averaged gene expression counts for all cells within a cell-type cluster, thus reducing the noise across single-cell measurements (see Material and Methods). Using our model, we tested the association between each of the traits and eight clustered cell types (Figure 5).

ACD is a trait characterized by a decline in cognitive capability and decreases in brain volume, both thought to be normal functions of aging. However, evidence suggests that the rate at which cognitive decline occurs is a precursor to late-onset AD, hinting at a shared genetic architecture.[64] Thus, we were interested in whether significant overlap of trait-associated cell types existed between the two traits. For ACD, we observed a significant association with fetal quiescent cells ($p = 0.03$), which primarily consist of neurons. Quiescent fetal cells differ from replicating fetal cells in that they have begun to downregulate neuronal growth factors such as EGR1.[39] On the other hand, we found an association between AD and microglia ($p = 0.03$) and astrocytes ($p = 0.03$) but no enrichment for fetal neurons ($p = 0.8$). To rule out an association driven by the *APOE* locus, we reran RolyPoly while removing a 1 Mb window centered on the *APOE* TSS. The significant microglia association persisted ($p = 0.03$), whereas the astrocyte association did not ($p = 0.1$). Thus, although the connection between astrocytes and AD is well studied,[65] this connection appears to be driven by few loci of large effect. Notably, there is mounting evidence for a more central role of microglia in AD;[66,67] our analyses provide genetic evidence in humans to support this hypothesis. Additionally, our results suggest a role for microglia in AD but not ACD. This finding is consistent with recent work demonstrating that although lipid regulation pathways are enriched with GWAS signal for both traits, immune pathways tend to show AD-specific signal.[68] Therefore, one could hypothesize microglial involvement during the transition between ACD and AD.

For schizophrenia, we found a significant relationship with oligodendrocytes ($p = 0.02$) and fetal replicating cells ($p = 0.01$). The genetic basis of schizophrenia is even less well understood than that of AD, but there is a significant body of literature on oligodendrocyte dysfunction and schizophrenia.[69,70] Moreover, recent genetic association studies have shown an enrichment of schizophrenia GWAS signal within pathways of development.[71–73]

To validate these associations between traits and single-cell cell-type clusters, we processed a single-cell dataset (see Material and Methods) from mouse brains,[20] which included seven major brain cell types that had been previously identified. Because it includes only one-to-one human and mouse orthologs, we consider this dataset to be an independent pseudo-human-brain single-cell dataset. Thus, we used this dataset to validate our previous findings. We limited our analysis to cell types overlapping between the human and mouse datasets; these included microglia and oligodendrocytes. For AD, we replicated the significant association with microglia ($p = 0.01$). Of note, there was a cluster that included astrocytes and ependymal cells; however, there was no significant association with this cluster. With schizophrenia, there was a suggestive association with the mouse-derived oligodendrocyte cell-type cluster ($p = 0.09$). Thus, from our analysis of mouse single-cell data, we replicated two of our initial trait and cell-type associations. Furthermore, we demonstrated that if human data are not available, one could swap in similar mouse data to guide initial analyses.

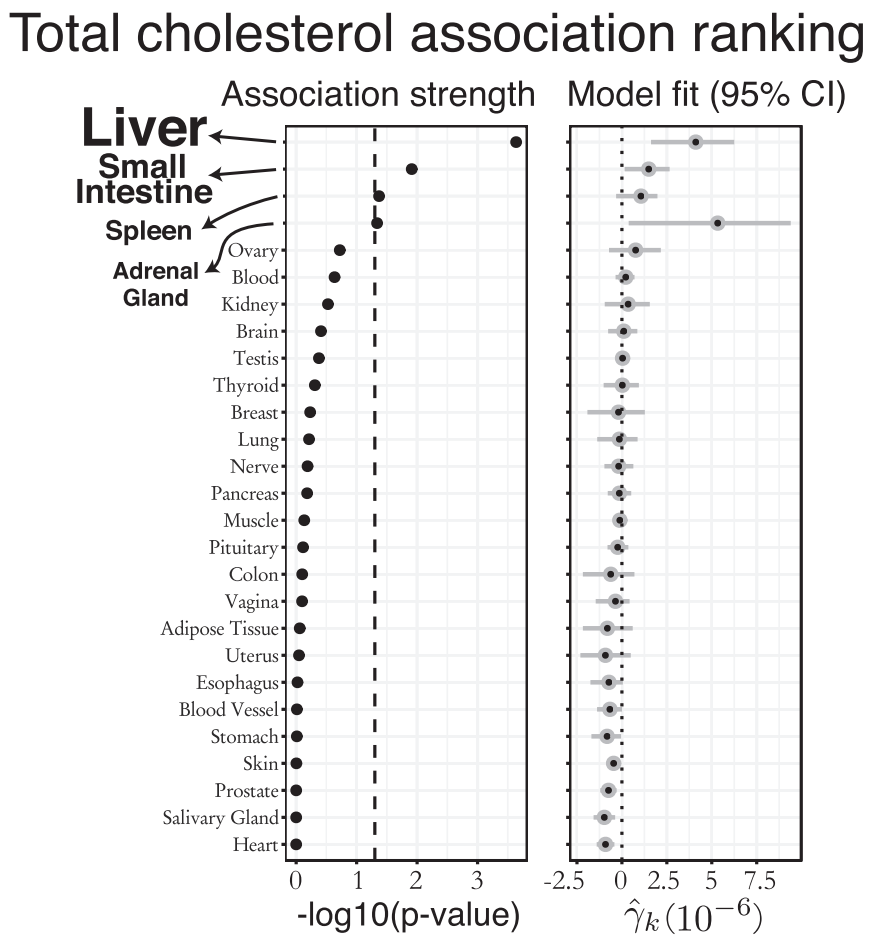


**Figure 3. TC and GTEx Tissue Ranking**
(Left) Tissues are ranked by p value, which represents the strength of association with TC.
(Right) Corresponding parameter estimates and 95% confidence intervals.

We also ran another closely related method, snpSEA,[10] on these data. This method also integrates gene expression with GWAS and aims to identify relevant tissue or cell types. However, it does not take advantage of genome-wide polygenic signal, and therefore its power is reduced. As a result, snpSEA identified an association between microglia and AD in humans but not in mice (Figure S2).

### RolyPoly Gene Scores Correlate with DE Genes in Individuals with AD

We were interested in studying whether RolyPoly-inferred model parameters could predict trait-relevant genes from an independent dataset. Thus, we downloaded and processed gene expression data from human prefrontal cortex samples of 101 control individuals and 129 individuals with AD (see Material and Methods and Zhang et al.[50]). A total of 9,228 genes were DE with a $q$-value $< 0.1\%$ (6,324 genes did not meet this threshold). Such a differential-expression study represents a data-driven approach to identifying AD-associated genes (independently from GWAS results). Additionally, we used summary statistics from this experiment to test the ability of our model parameters to identify trait-relevant genes.

To establish a baseline, we tested the enrichment of LD-informed gene-score estimates within DE genes. We computed these values by taking the variance of GWAS effect sizes within 5 kb of a gene and incorporating information about LD (see Material and Methods). Compared with genes not found to be significantly expressed, the set of DE genes showed only weakly suggestive ($p = 0.09$) enrichment of these values.

As a first step to incorporating information from RolyPoly-inferred model parameters, we tested whether genes that were specifically expressed in a RolyPoly-inferred trait-relevant cell type were enriched with larger differential-expression test statistics. We identified the top 10% of genes specifically expressed in the microglia cell type (which our model identified as significantly associated with AD). Compared with a control set of genes, this set of genes showed a significant ($p = 1 \times 10^{-8}$) enrichment of positive values of the differential-expression test statistic (Figure 6A, right). We performed a similar analysis with a cell type for which RolyPoly did not find evidence of AD association. There was no enrichment of DE summary statistics within the set of genes specifically expressed in fetal quiescent cells (Figure 6A, left).

From these observations, we reasoned that we could rank the trait relevance of genes on the basis of RolyPoly-inferred parameter estimates, $\hat{\boldsymbol{\gamma}}$, and gene expression. As an example for AD, a gene that is specifically expressed in microglia and astrocytes would be ranked higher than a housekeeping gene. Thus, we defined the RolyPoly trait-relevance gene score $h_j^{\text{gene}}$ as a linear combination of $\hat{\boldsymbol{\gamma}}$ and normalized gene expression values (see Material and Methods). Using the model from the AD-specific panel of Figure 5 and human brain single-cell gene expression, we computed estimates of $h_j^{\text{gene}}$. Furthermore, we hypothesized that $h_j^{\text{gene}}$ values could predict DE genes. We found that $h_j^{\text{gene}}$ scores were significantly enriched within the DE genes ($p = 7 \times 10^{-18}$; Figure S3). However, it is possible that correlations among co-regulated genes could result in uncalibrated p values. Therefore, we designed a test that accounts for the covariance structure between genes (see Material and Methods). Using this test, we still identified a significant association ($p = 4 \times 10^{-3}$) between DE genes and $h_j^{\text{gene}}$ values (see Figure 6B).

For validation, we were interested in replicating our enrichment of $h_j^{\text{gene}}$ in DE genes in an independent dataset.

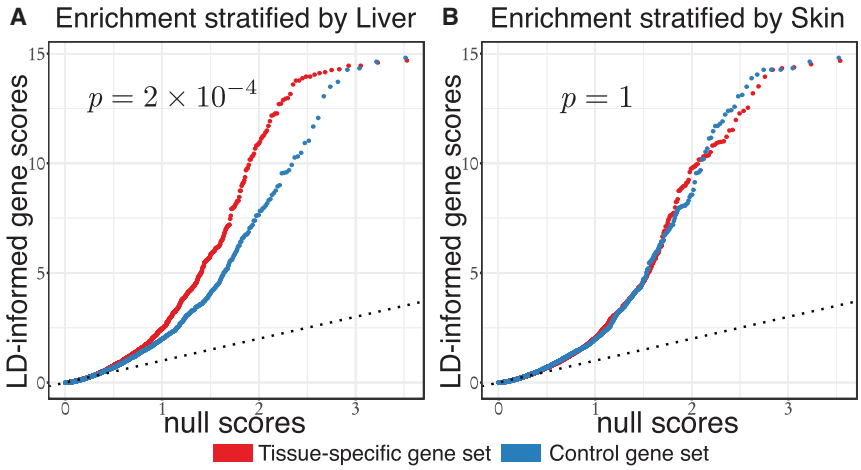


**Figure 4. TC and GTEx Q-Q Plot Comparing Enrichment of LD-Informed Gene Scores**

Both plots show the p value from RolyPoly for the association between the respective tissue and TC.

(A) Q-Q plot comparing enrichment of LD-informed gene scores in genes that are uniquely expressed in the liver. To select gene sets, we sorted genes by their normalized expression in the liver and took the top 20% of genes (red) and the bottom 20% of genes (blue).

(B) A similar plot stratifying gene values by skin-specific gene expression (skin is not predicted to have a role in cholesterol regulation).

Sekar et al. performed laser-capture microdissection to isolate astrocytes from ten healthy control individuals and individuals with AD and then identified 227 DE genes.[74] Of those genes, we predicted RolyPoly gene scores for 150 (70 of the 227 genes were not annotated as protein coding; therefore, they were not measured in the single-cell experiment). We replicated our previous result and identified a significant ($p = 1 \times 10^{-3}$) enrichment within DE genes (Figure S4). We were unable to perform our enrichment test that accounts for gene correlations because gene expression data were not available for this dataset.

Thus, we conclude that from GWAS and gene expression data of healthy individuals, our model parameters capture information about the relevance of a gene to a trait on the basis of which cell types express the gene. Still, we cannot discount the possibility that observed enrichments of differential-expression test statistics are a result of changes in cell-type proportions. However, in such a scenario, we would have identified trait-relevant cell types that increase or decrease in proportion and thus would be consistent with our conclusion about RolyPoly parameters.

## Discussion

We have described a polygenic model for analyzing single-cell gene expression and GWAS summary statistics. Our results demonstrate that we can identify trait-relevant cell types from complex tissues and prioritize genes for further analysis.

Here, we discuss the following assumptions underlying RolyPoly: (1) we focused on *cis*-GWAS effects (as opposed to *trans*) because *cis*-SNPs tend to more consistently have genome-wide effects, and larger effects, on the regulation of gene expression.[31,75–78] (2) Our model treats neighboring genes independently even though some might share *cis*-SNPs, which could result in a correlation among nearby SNP effect sizes. However, we corrected for this effect by performing block bootstrap when computing standard errors and empirical confidence intervals. (3) Because this was a joint analysis (we estimated all annotation parameters at the same time), including or excluding gene expression data that are causal or correlated with causal cell types can have an effect on inference (i.e., result in different model parameter estimates). However, joint analysis is necessary because analyzing each cell type separately would not control for potential overlap of specifically expressed genes. To mitigate these effects, we suggest several approaches. First, one could re-analyze a trait GWAS as more data become available. Second, we recommend a cautious interpretation of model parameters, which should be guided by domain knowledge. Third and finally, with highly correlated annotations, one could carry out an initial round of feature selection before performing standard inference or include regularization (described in the Material and Methods). Even with these model assumptions, our results are well supported by known biology, as shown in the analysis of tissues and brain cell types.

This is a promising step toward connecting single-cell gene expression and GWAS summary statistics to identify relevant cell types and genes. Although there is evidence linking the immune system and microglia to AD in mice,[66] we have identified an enrichment of genetic trait-association signal near genes specifically expressed in human microglia. More generally, single-cell technologies represent an opportunity to discover and characterize novel cell types and cell states.[22] Thus, there is a need for methods, such as RolyPoly, that can prioritize novel cell types relevant to human phenotypes for further study. Here, we focused on single cells clustered into cell types; however, numerous alternative groupings could be examined. For example, during cell stimulation, there exists significant cell heterogeneity even within classical marker-defined immune cell populations.[79,80] Using RolyPoly, one could link these novel subpopulations to autoimmune disease phenotypes. These analyses should increase only as single-cell data become more commonly available.

It is challenging to pinpoint causal genes from GWASs because LD-related correlations among SNP effects

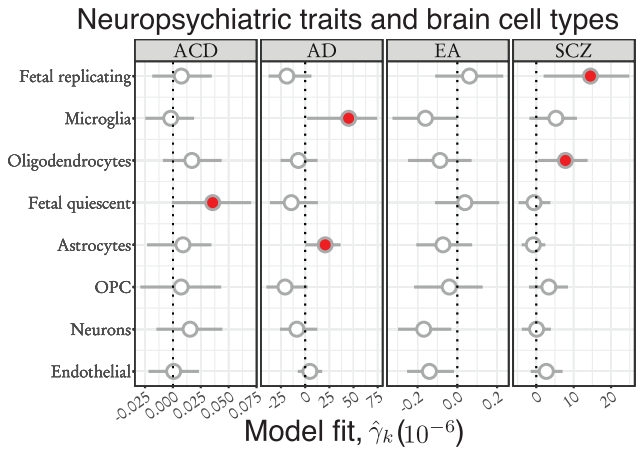


**Figure 5. Neuropsychiatric Trait Associations with Single-Cell-Based Cell Types**
Parameter estimates for age-related cognitive decline (ACD), Alzheimer disease (AD), educational attainment (EA), schizophrenia (SCZ), and single-cell-based cell-type clusters from the human brain dataset.[39] Range specifies the empirical bounds of the 95% confidence interval. Estimates highlighted in red represent significant associations ($p < 0.05$).

confound the identification of causal variants. Moreover, it is difficult to identify the target gene modulated by a regulatory variant. Statistical methods that integrate GWASs and eQTLs while accounting for the effects of LD[81,82] have proven useful. However, the eQTL data might not be specific to the disease-relevant tissue or cell type. To supplement these approaches, we suggest using the signature of gene expression and parameters from our model to prioritize genes proximal to significant GWAS variants for further analysis. Consider a region with complex LD structure and significant trait-association signal; ideally, one would rely on overlapping eQTL information to identify the causal SNP and gene. But, without knowledge of the causal tissue, GWAS-eQTL overlap with a non-causal tissue could be misleading and complicate the task of collecting relevant eQTL information. Instead, one could use annotation parameter estimates from RolyPoly with tissue- or cell-type-specific gene expression to calculate $h_j^{\text{gene}}$ trait importance values and prioritize genes within the local GWAS region. Additionally, as we have shown, our method can identify significantly associated tissues, which one could prioritize for collection of population samples for eQTL analysis.

## Appendix A

### Derivation of Univariate Effect Estimates

We follow much of the notation and derivation from Shi et al.[25] Starting with the definition of the annotation coefficients (recalling that the genotype matrix has been scaled),

$$\widehat{\beta}_i = \frac{1}{n}\boldsymbol{X}_i^T\boldsymbol{y},$$

we substitute the GWAS model,

$$\begin{aligned}\widehat{\beta}_i &= \frac{1}{n}\boldsymbol{X}_i^T(\boldsymbol{X}\boldsymbol{\beta}+\epsilon)\\ &= \frac{1}{n}\boldsymbol{X}_i^T\boldsymbol{X}_1\beta_1 + \ldots + \frac{1}{n}\boldsymbol{X}_i^T\boldsymbol{X}_p\beta_p + \frac{1}{n}\boldsymbol{X}_i^T\epsilon,\end{aligned}$$

and use the definition of Pearson's correlation coefficient once again by relying on the fact that the genotype matrix has been scaled and centered,

$$\widehat{\beta}_i = \sum_{i'=1}^{p} r_{ii'}\beta_{i'} + \frac{1}{n}\boldsymbol{X}_i^T\epsilon.$$

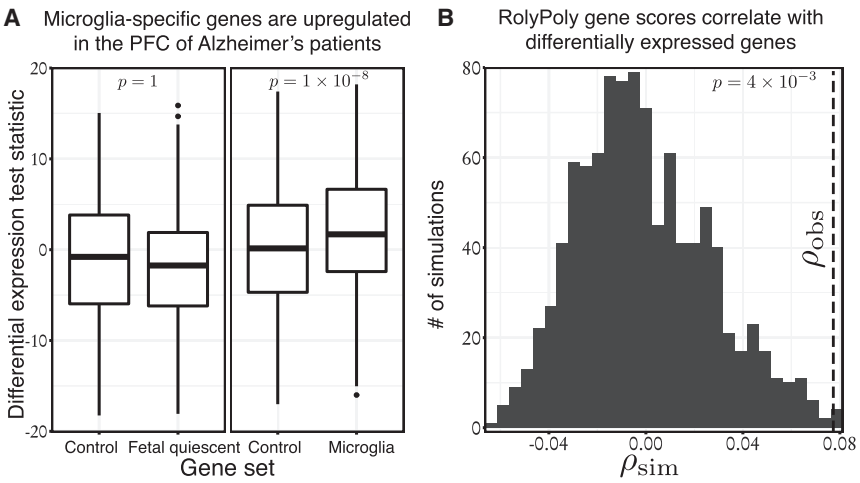


**Figure 6. RolyPoly-Inferred Model Parameters Predict DE Genes in the Prefrontal Cortex (PFC) of Individuals with AD**
(A) Differential-expression test statistics (a larger value represents genes that are upregulated in the brains of affected individuals) were significantly larger in the set of genes specifically expressed in the microglia cell type than in a control gene set (right). We define the set of cell-type-specific genes as the top 10% specifically expressed genes. We compared them with the control gene set, which includes genes that deviate the least from average gene expression. The differential-expression test statistic was not enriched in genes specifically expressed in the fetal quiescent cell type (left).
(B) Controlling for the effect of correlation between gene expression values of co-regulated genes, we observed an enrichment of $h_j^{\text{gene}}$ values in DE genes. The significance of the observed Spearman's rank-correlation coefficient between $h_j^{\text{gene}}$ and the differential-expression test statistic was evaluated with a null distribution generated from simulations, which accounted for the gene expression covariance structure (full details of this test can be found in the Material and Methods).

In the Material and Methods, we write the above expression with matrix notation. Others have described a similar relationship among estimated effects, LD, and the true effect sizes.[83]

### Derivation of Distribution Parameters of Effect Estimates

Here, we describe the mean and variance of the estimated SNP effects by using our polygenic model. The expected value is computed as follows,

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\right] &= \mathbb{E}\left[\boldsymbol{R}\boldsymbol{\beta} + (1/n)\boldsymbol{X}^T\epsilon\right] \\ &= \boldsymbol{R}\mathbb{E}[\boldsymbol{\beta}] + (1/n)\boldsymbol{X}^T\mathbb{E}[\epsilon],\end{aligned}$$

and because we model the genetic and environmental effects with normal distributions with mean 0, we conclude that $\mathbb{E}[\widehat{\boldsymbol{\beta}}] = 0$. Next,

$$\begin{aligned}\mathbb{V}\left[\widehat{\boldsymbol{\beta}}\right] &= \mathbb{V}\left(\boldsymbol{R}\boldsymbol{\beta} + (1/n)\boldsymbol{X}^T\epsilon\right) \\ &= \boldsymbol{R}\mathbb{V}(\boldsymbol{\beta})\boldsymbol{R} + \left(1/n^2\right)\boldsymbol{X}\mathbb{V}(\epsilon)\boldsymbol{X}^T \\ &= \boldsymbol{R}\boldsymbol{D}\boldsymbol{R} + \left(\sigma_e^2/n\right)\boldsymbol{R},\end{aligned}$$

where $\boldsymbol{D}$ refers to the diagonal matrix of SNP effect-size variances; in the second equality, we use the fact that $\boldsymbol{R} = \boldsymbol{R}^T$. We use these values of the expectation and variance to parameterize the multivariate normal distribution that describes the estimated GWAS effect sizes.

### Derivation of Expected SNP Variance

Note that the distribution of the squared $\ell_2$ norm of a random vector drawn from a multivariate normal distribution with mean 0 is the trace of the covariance matrix.[36,84] Thus, the expected value of the sum of squared SNP effect sizes near gene $j$ is given by

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in S_j}\widehat{\beta}_i^2\right] &= \mathrm{Tr}\left(\tau_j\boldsymbol{R}_{S_j}\boldsymbol{R}_{S_j} + \sigma_e^2 n^{-1}\boldsymbol{R}_{S_j}\right) \\ &= \tau_j\mathrm{Tr}\left(\boldsymbol{R}_{S_j}^2\right) + |S_j|\,\sigma_e^2 n^{-1}.\end{aligned}$$

We derive this by using the linearity of the trace and recalling that $\boldsymbol{R}$ is a correlation matrix, and hence the diagonal elements are 1. When SNP annotations are included, we model the expected value of the squared marginal SNP effect size. The marginal distribution of the squared SNP effect size around gene $j$ is $\widehat{\beta}_i \sim \mathrm{N}(0, \sigma_e^2 n^{-1} + (\boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j})_{ii})$. Finally,

$$\begin{aligned}\mathbb{E}\left[\widehat{\beta}_i^2\right] &= \mathrm{Tr}\left(\sigma_e^2 n^{-1} + \left(\boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j}\right)_{ii}\right), \quad \text{where } i\in S_j \\ &= \sigma_e^2 n^{-1} + \left(\boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j}\right)_{ii}.\end{aligned}$$

### Relationship to Previous Work

Rewriting $(\boldsymbol{R}\boldsymbol{D}\boldsymbol{R})_{ii}$ as $\sum_{i'}\nu_{i'}r_{ii'}^2$ and substituting quantitative feature values with an indicator function that signifies whether a SNP is within a discrete annotation class, we arrive at an equation similar to the basic LD Score regression model,

$$\begin{aligned}\mathbb{E}\left[\beta_i^2\right] &= \left(\sigma_e^2/n\right) + \sum_{i'}\nu_{i'}r_{ii'}^2 \\ &= \left(\sigma_e^2/n\right) + \sum_{l=1}^{P}\phi_l\sum_{i'}1_l(b_{i'})r_{ii'}^2 + \sum_{k=1}^{N}\gamma_k\sum_{i'}1_k\left(a_{g(i')}\right)r_{ii'}^2.\end{aligned}$$

Note that we go from the first to the second line by substituting $\nu$ from Equation 3. Although the models share some similarities, our model is derived independently to utilize the full quantitative data from single-cell gene expression assays.

## Supplemental Data

Supplemental Data include four figures and can be found with this article online at https://doi.org/10.1016/j.ajhg.2017.09.009.


## Acknowledgments

We thank Z. Gao, N. Latorraca, N. Sinnott-Armstrong, and anonymous reviewers for comments; A. Raj for technical assistance; and N. Telis for producing organ images. Support for D.C. was provided by National Library of Medicine Training Grant T15LM007033. Support for A.Q.F. was provided by NIH grant R00HG007368. This work was supported by NIH grant 1R01HG008140-01A1 and by the Howard Hughes Medical Institute.

Received: May 10, 2017
Accepted: September 13, 2017
Published: October 26, 2017


## Web Resources

OMIM, http://omim.org/
`rolypoly`: Identifying Trait-Relevant Functional Annotations, https://cran.r-project.org/package=rolypoly
`rolypoly` source code, https://github.com/dcalderon/rolypoly
UCSC Genome Browser, https://genome.ucsc.edu/

## References


1. Claussnitzer, M., Dankel, S.N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I.S., Beaudry, J.L., Puviindran, V., et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med. *373*, 895–907.
2. Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177–183.
3. Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., et al. (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science *344*, 519–523.
4. Ongen, H., Brown, A.A., Delaneau, O., Panousis, N., Nica, A.C., GTEx Consortium, and Dermitzakis, E.T. (2016).

Estimating the causal tissues for complex traits and diseases. bioRxiv. https://doi.org/10.1101/074682.
5. Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.
6. Finucane, H., Reshef, Y., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Genovese, G., Saunders, A., et al. (2017). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. bioRxiv. https://doi.org/10.1101/103069.
7. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., et al.; ReproGen Consortium; Schizophrenia Working Group of the Psychiatric Genomics Consortium; and RACI Consortium (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet. *47*, 1228–1235.
8. Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am. J. Hum. Genet. *89*, 496–506.
9. Pickrell, J.K. (2014). Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am. J. Hum. Genet. *94*, 559–573.
10. Slowikowski, K., Hu, X., and Raychaudhuri, S. (2014). SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. Bioinformatics *30*, 2496–2497.
11. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat. Genet. *45*, 124–130.
12. Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol. *33*, 155–160.
13. Fuzik, J., Zeisel, A., Máté, Z., Calvigioni, D., Yanagawa, Y., Szabó, G., Linnarsson, S., and Harkany, T. (2016). Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. Nat. Biotechnol. *34*, 175–183.
14. Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251–255.
15. Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.
16. Junker, J.P., and van Oudenaarden, A. (2014). Every cell is special: genome-wide studies add a new dimension to single-cell biology. Cell *157*, 8–11.
17. Kowalczyk, M.S., Tirosh, I., Heckl, D., Rao, T.N., Dixit, A., Haas, B.J., Schneider, R.K., Wagers, A.J., Ebert, B.L., and Regev, A. (2015). Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. Genome Res. *25*, 1860–1872.
18. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.
19. Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature *509*, 371–375.
20. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.
21. Campbell, J.N., Macosko, E.Z., Fenselau, H., Pers, T.H., Lyubetskaya, A., Tenen, D., Goldman, M., Verstegen, A.M., Resch, J.M., McCarroll, S.A., et al. (2017). A molecular census of arcuate hypothalamus and median eminence cell types. Nat. Neurosci. *20*, 484–496.
22. Regev, A., et al. (2017). The Human Cell Atlas. bioRxiv. https://doi.org/10.1101/ 121202.
23. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al.; Global Lipids Genetics Consortium (2013). Discovery and refinement of loci associated with lipid levels. Nat. Genet. *45*, 1274–1283.
24. Boyle, E.A., Li, Y.I., and Pritchard, J.K. (2017). An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell *169*, 1177–1186.
25. Shi, H., Kichaev, G., and Pasaniuc, B. (2016). Contrasting the genetic architecture of 30 complex traits from summary association data. Am. J. Hum. Genet. *99*, 139–153.
26. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
27. Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.
28. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
29. Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., et al. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390–394.
30. Gaffney, D.J., Veyrieras, J.-B., Degner, J.F., Pique-Regi, R., Pai, A.A., Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). Dissecting the regulatory architecture of gene expression QTLs. Genome Biol. *13*, R7.
31. Montgomery, S.B., and Dermitzakis, E.T. (2011). From expression QTLs to personalized transcriptomics. Nat. Rev. Genet. *12*, 277–282.
32. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.
33. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes

confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
34. Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statist. Sci. *1*, 54–75.
35. Kunsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. Ann. Statist. *17*, 1217–1241.
36. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS Comput. Biol. *12*, e1004714.
37. Liu, J.Z., McRae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G., Macgregor, S.; and AMFS Investigators (2010). A versatile gene-based test for genome-wide association studies. Am. J. Hum. Genet. *87*, 139–145.
38. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al.; GTEx Consortium (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585.
39. Darmanis, S., Sloan, S.A., Zhang, Y., Enge, M., Caneda, C., Shuer, L.M., Hayden Gephart, M.G., Barres, B.A., and Quake, S.R. (2015). A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. USA *112*, 7285–7290.
40. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The technology and biology of single-cell RNA sequencing. Mol. Cell *58*, 610–620.
41. Saliba, A.-E., Westermann, A.J., Gorski, S.A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. *42*, 8845–8860.
42. Grün, D., and van Oudenaarden, A. (2015). Design and analysis of single-cell sequencing experiments. Cell *163*, 799–810.
43. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.
44. Shekhar, K., Lapan, S.W., Whitney, I.E., Tran, N.M., Macosko, E.Z., Kowalczyk, M., Adiconis, X., Levin, J.Z., Nemesh, J., Goldman, M., et al. (2016). Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell *166*, 1308–1323.e30.
45. Vallejos, C.A., Marioni, J.C., and Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput. Biol. *11*, e1004333.
46. Kinsella, R.J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) *2011*, bar030.
47. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
48. Lambert, J.-C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al.; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; and Cohorts for Heart and Aging Research in Genomic Epidemiology (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.
49. Okbay, A., Beauchamp, J.P., Fontana, M.A., Lee, J.J., Pers, T.H., Rietveld, C.A., Turley, P., Chen, G.-B., Emilsson, V., Meddens, S.F., et al.; LifeLines Cohort Study (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature *533*, 539–542.
50. Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z., et al.; Electronic Medical Records and Genomics (eMEMERGEGE) Consortium; MIGen Consortium; PAGEGE Consortium; and LifeLines Cohort Study (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet. *46*, 1173–1186.
51. Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A.A., Zhang, C., Xie, T., Tran, L., Dobrin, R., et al. (2013). Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell *153*, 707–720.
52. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.
53. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
54. Berndt, S.I., Gustafsson, S., Mägi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C., Day, F.R., et al. (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat. Genet. *45*, 501–512.
55. Hall, J.E. (2003). The kidney, hypertension, and obesity. Hypertension *41*, 625–633.
56. Knudsen, N., Laurberg, P., Rasmussen, L.B., Bülow, I., Perrild, H., Ovesen, L., and Jørgensen, T. (2005). Small differences in thyroid function may be important for body mass index and the occurrence of obesity in the population. J. Clin. Endocrinol. Metab. *90*, 4019–4024.
57. Reinehr, T., and Andler, W. (2002). Thyroid hormones before and after weight loss in obesity. Arch. Dis. Child. *87*, 320–323.
58. Locke, A.E., Kahali, B., Berndt, S.I., Justice, A.E., Pers, T.H., Day, F.R., Powell, C., Vedantam, S., Buchkovich, M.L., Yang, J., et al.; LifeLines Cohort Study; ADIPOGen Consortium; AGEN-BMI Working Group; CARDIOGRAMplusC4D Consortium; CKDGen Consortium; GLGC; ICBP; MAGIC Investigators; MuTHER Consortium; MIGen Consortium; PAGE Consortium; ReproGen Consortium; GENIE Consortium; and International Endogene Consortium (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature *518*, 197–206.
59. Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S., Lindgren, C.M., Heid, I.M., Berndt, S.I., Elliott, A.L., Jackson, A.U., Lamina, C., et al.; Wellcome Trust Case Control Consortium; and Genetic Investigation of ANthropometric Traits Consortium (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. Nat. Genet. *41*, 25–34.
60. Rietveld, C.A., Medland, S.E., Derringer, J., Yang, J., Esko, T., Martin, N.W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., et al.; LifeLines Cohort Study (2013). GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. Science *340*, 1467–1471.

61. Eugster, E.A., and Pescovitz, O.H. (1999). Gigantism. J. Clin. Endocrinol. Metab. *84*, 4379–4384.
62. Fatouros, M., Bourantas, K., Bairaktari, E., Elisaf, M., Tsolas, O., and Cassioumis, D. (1995). Role of the spleen in lipid metabolism. Br. J. Surg. *82*, 1675–1677.
63. Nussey, S.S., and Whitehead, S.A. (2013). Endocrinology: an integrated approach (CRC Press).
64. De Jager, P.L., Shulman, J.M., Chibnik, L.B., Keenan, B.T., Raj, T., Wilson, R.S., Yu, L., Leurgans, S.E., Tran, D., Aubin, C., et al.; Alzheimer's Disease Neuroimaging Initiative (2012). A genome-wide scan for common variants affecting the rate of age-related cognitive decline. Neurobiol. Aging *33*, 1017.e1–1017.e15.
65. Verkhratsky, A., Olabarria, M., Noristani, H.N., Yeh, C.-Y., and Rodriguez, J.J. (2010). Astrocytes in Alzheimer's disease. Neurotherapeutics *7*, 399–412.
66. Gjoneska, E., Pfenning, A.R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., and Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature *518*, 365–369.
67. Rogers, J., Mastroeni, D., Leonard, B., Joyce, J., and Grover, A. (2007). Neuroinflammation in Alzheimer's disease and Parkinson's disease: are microglia pathogenic in either disorder? Int. Rev. Neurobiol. *82*, 235–246.
68. Raj, T., Chibnik, L.B., McCabe, C., Wong, A., Replogle, J.M., Yu, L., Gao, S., Unverzagt, F.W., Stranger, B., Murrell, J., et al. (2016). Genetic architecture of age-related cognitive decline in African Americans. Neurol. Genet. *3*, e125.
69. Tkachev, D., Mimmack, M.L., Ryan, M.M., Wayland, M., Freeman, T., Jones, P.B., Starkey, M., Webster, M.J., Yolken, R.H., and Bahn, S. (2003). Oligodendrocyte dysfunction in schizophrenia and bipolar disorder. Lancet *362*, 798–805.
70. Uranova, N.A., Vostrikov, V.M., Vikhreva, O.V., Zimina, I.S., Kolomeets, N.S., and Orlovskaya, D.D. (2007). The role of oligodendrocyte pathology in schizophrenia. Int. J. Neuropsychopharmacol. *10*, 537–545.
71. Gulsuner, S., Walsh, T., Watts, A.C., Lee, M.K., Thornton, A.M., Casadei, S., Rippey, C., Shahin, H., Nimgaonkar, V.L., Go, R.C., et al.; Consortium on the Genetics of Schizophrenia (COGS); and PAARTNERS Study Group (2013). Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell *154*, 518–529.
72. Jaffe, A.E., Shin, J., Collado-Torres, L., Leek, J.T., Tao, R., Li, C., Gao, Y., Jia, Y., Maher, B.J., Hyde, T.M., et al. (2015). Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. Nat. Neurosci. *18*, 154–161.
73. Kahn, R.S., Sommer, I.E., Murray, R.M., Meyer-Lindenberg, A., Weinberger, D.R., Cannon, T.D., O'Donovan, M., Correll, C.U., Kane, J.M., van Os, J., and Insel, T.R. (2015). Schizophrenia. Nat Rev Dis Primers *1*, 15067.
74. Sekar, S., McDonald, J., Cuyugan, L., Aldrich, J., Kurdoglu, A., Adkins, J., Serrano, G., Beach, T.G., Craig, D.W., Valla, J., et al. (2015). Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. Neurobiol. Aging *36*, 583–591.
75. Grundberg, E., Small, K.S., Hedman, Ä.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al.; Multiple Tissue Human Expression Resource (MuTHER) Consortium (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. *44*, 1084–1089.
76. Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. *7*, e1001317.
77. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.
78. Veyrieras, J.-B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet. *4*, e1000214.
79. Arsenio, J., Kakaradov, B., Metz, P.J., Kim, S.H., Yeo, G.W., and Chang, J.T. (2014). Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. Nat. Immunol. *15*, 365–372.
80. Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240.
81. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., Sul, J.H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL Signals Detects Target Genes. Am. J. Hum. Genet. *99*, 1245–1260.
82. Kumasaka, N., Knights, A.J., and Gaffney, D.J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. Nat. Genet. *48*, 206–213.
83. Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A.L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. *10*, e1004722.
84. Mathai, A.M., and Provost, S.B. (1992). Quadratic forms in random variables: theory and applications (M. Dekker), pp. 49–51.