Inferring relevant cell types for complex traits 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] and Jonathan K. Pritchard[2,3,9]

[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, 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*

Abstract

Previous studies have prioritized trait-relevant cell types by looking for an enrichment of 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, hardly any work exists linking single-cell 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 single-cell RNA-seq. We demonstrate RolyPoly's accuracy through simulation and validate previously known tissue-trait associations. We discover a significant association between microglia and late-onset Alzheimer's disease, and an association between oligodendrocytes and replicating fetal cortical cells with schizophrenia. Additionally, RolyPoly computes a trait-relevance score for each gene which reflects the importance of expression specific to a cell type. We found that differentially expressed genes in the prefrontal cortex of Alzheimer's patients were significantly enriched for highly ranked genes 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 the *FTO* locus acts on *IRX3* and *IRX5* primarily in human adipocyte progenitor cells enabled researchers to rigorously define a novel thermogenesis pathway central for lipid storage and obesity [6]. And, focusing on distinct human *C4* isotypes, Sekar et. al., highlighted the role of the classical complement cascade (of which *C4* is a critical component) and synapse elimination during development in the brains of individuals with schizophrenia [57].

In addition to estimating disease risk for individual variants, GWAS have proven useful for identifying trait-relevant cell types or tissues. Assuming 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 regions of the genome that affect gene regulation. A series of studies identified enrichment of GWAS signal in sorted cell type [51] or tissue-specific eQTLs [45]. Other approaches have revealed enrichment of GWAS signal in cell type-specific functional annotations (e.g., ATAC-seq, ChIP-seq, RNA-seq) [25, 66, 47, 15, 61, 14, 12]. However, these analyses are limited in cell type resolution because they either require samples with population variation (infeasible to collect for many cell types) or rely on functional assays that require on the order of 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 heterogenous cell types. For example, describing the brain as the primary pathogenic tissue responsible for schizophrenia or Alzheimer's disease is unsatisfying, but it remains difficult to comprehensively collect functional information from the plethora of brain cell types necessary to perform standard GWAS enrichment analyses.

Meanwhile, single-cell gene expression technology has offered insights into complex cell types [46, 27, 71, 65, 33, 16, 20, 28, 4]. Additionally, there are concerted efforts underway to develop 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 [52]. However, to our knowledge, there are no existing methods designed to link novel single-cell based cell types and phenotypes of interest.

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 genome-wide 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). 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 γ 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, based on 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^{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^{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 we represent the stochastic environmental error with $\epsilon_s \sim N(0, \sigma_e^2)$. Importantly, we assume that the matrix of genotypes has been scaled and standardized such that the mean is 0 and variance 1 for each SNP vector

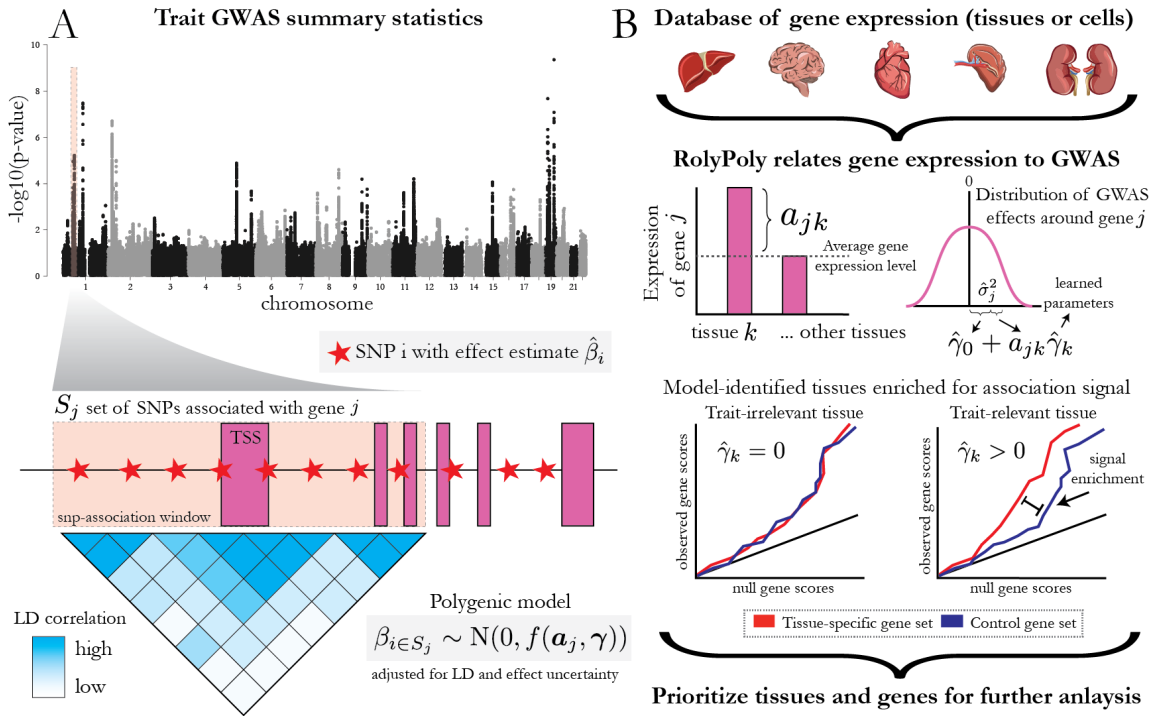

Figure 1: RolyPoly *detects trait-associated annotations using GWAS summary statistics and gene expression profiles. A) We model the variance of GWAS effect sizes of SNPs associated with a gene as a function of gene annotations, in particular gene expression, while accounting for LD using population matched genotype correlation information. (Manhattan plot is based on data from [19].) B) From a database of functional information (such as tissue or cell type RNA-seq) we learn a regression coefficient for each annotation, $\hat{\gamma}_k$, that captures its 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.*

(and similarly for the trait y_s). The main summary statistics released by GWAS are per-variant effect estimates, which we refer to as $\hat{\boldsymbol{\beta}}$. Researchers typically calculate and report univariate effect-size estimates. These estimates represent the marginal standardized regression coefficient and are calculated as $\hat{\beta}_i = n^{-1}\boldsymbol{X}_i^T\boldsymbol{y}$, where $\boldsymbol{X}_i$ (note the change in case) represents standardized genotypes for SNP i across the n individuals (see Appendix for derivation). Substituting the polygenic model for $\boldsymbol{y}$ into the estimation equation (see Appendix for derivation), the sampling distribution

of the estimated SNP effect sizes corresponds to

$$\hat{\boldsymbol{\beta}} = \boldsymbol{R}\boldsymbol{\beta} + n^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{R}$ is the sample LD matrix (i.e. $r_{ii'}$ is the Pearson correlation values between genotype i and i'). Using this definition of estimated GWAS effect sizes, we 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 coordinates of $\boldsymbol{\beta}$ whose indices lie in set S. We assume *a priori* that the true GWAS effect sizes of SNPs in gene j follow a normal distribution $\boldsymbol{\beta}_{S_j} \sim \text{MVN}(\mathbf{0}, \tau_j I)$, where I is the $|S_j| \times |S_j|$ identity matrix, and τ_j is the prior effect size variance for all SNPs associated with gene j and is modeled as a linear function. More specifically, τ_j is a linear function of N annotations a_{jk} (in this case cell-type specific gene expression), with annotation coefficients γ_k and an intercept term γ_0:

$$\tau_j = \gamma_0 + \sum_{k=1}^{N} \gamma_k a_{jk} \,. \tag{2}$$

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 γ. SNPs proximal to genes that are specifically expressed in an annotation with a negative γ estimate are expected to have reduced effect size variance compared with the population mean.

Based on this polygenic model, the expected value of the vector of GWAS effect sizes around gene j is $\mathbb{E}[\hat{\boldsymbol{\beta}}_{S_j}] = \mathbf{0}$, and the covariance matrix is given by $\mathbb{V}[\hat{\boldsymbol{\beta}}_{S_j}] = \tau_j \boldsymbol{R}_{S_j}\boldsymbol{R}_{S_j} + \sigma_e^2 n^{-1}\boldsymbol{R}_{S_j}$, where $\boldsymbol{R}_{S_j}$ denotes the principal submatrix of $\boldsymbol{R}$ indexed by the SNPs in S_j (see Appendix 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 zero and

the same per-SNP variance of τ_j. However, there are other SNP annotations that we expect to affect the variance of a GWAS effect size, such as the minor allele frequency (MAF) of the SNP. Therefore, we include P SNP-level features as covariates while estimating the variance contribution of gene expression. Specifically, we modify our model to use a per-SNP variance ν_i for SNP i, given by

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

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-th value of SNP-level annotation l for SNP i, and ϕ_l is the annotation coefficient for annotation l. The distribution for the vector of SNP effects associated with a gene becomes

$$\hat{\boldsymbol{\beta}}_{S_j} \sim \text{MVN}(\mathbf{0}, \boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j} + \sigma_e^2 n^{-1}\boldsymbol{R}_{S_j}), \tag{4}$$

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 1000 Genomes Project [1].

For results presented here, we used a window size of 10kb centered on the transcription start site of a gene to associate a SNP to a gene. We chose this window-size because previous work has found that, across a diverse set of cell types and tissues, most eQTLs consistently lie in this region [9, 17, 62, 41]. 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 towards 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 duplicate these SNPs and treat them as independent SNP-gene pairs. Since RolyPoly infers parameters from hundreds of thousands of SNPs, we do not expect this contribute significantly to inferred parameters.

Parameter inference

In order 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 (4), which would be computationally challenging. Instead, we adopt a method of moments approach, where we fit the gene-level annotation coefficients γ_k and, if included, the SNP-level annotation coefficients ϕ_l. 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

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

where Tr above represents the trace of a matrix (derivation in Appendix). This expectation was derived recognizing that the expected value of the squared ℓ_2 norm of a mean zero multivariate normal distribution is the trace of the covariance matrix. When we include SNP annotation coefficients such that each SNP effect size has a variance term ν_i, we perform inference by fitting the observed and expected squared effect size of each SNP, where the expected value is given by

$$\mathbb{E}[\hat{\beta}_i^2] = (\boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j})_{ii} + \sigma_e^2 n^{-1}, \quad \text{where } j = g(i), \tag{6}$$

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, we noticed that this model relates to previous work [5] (described in the Appendix). We perform block bootstrap [10] to estimate standard errors, $\hat{\sigma}_{\gamma_k}$, which are used to compute a t-statistic, $\hat{\gamma}_k/\hat{\sigma}_{\gamma_k}$, and corresponding p-values. We use a t-statistic because we use 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 sample from these blocks with replacement [35]. Additionally, from the bootstrap parameter estimates, we calculate 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 rank 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 proportion of variance explained by individual annotations

Using a set of inferred gene annotation coefficients, $\hat{\boldsymbol{\gamma}}$, we calculate several quantities that summarize the contributions of gene annotations to the phenotypic variance. First, we compute $h_j^{\text{gene}} = \sum_{k=1}^{N} \hat{\gamma}_k a_{jk}$, which can be used to rank trait-relevant genes. Essentially, h_j^{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^{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^{gene} relies on GWAS indirectly through the RolyPoly-inferred parameters. Additionally, we calculate 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 show that the true value of h_k^{annot} affects our power to detect trait-annotation associations. The total contribution explained by all annotations, h^{total}, comes 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^{annot}, can be calculated as $h_k^{\text{annot}}/h^{\text{total}}$.

To validate our gene importance values, h_j^{gene}, we compare them to 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 accounting for inflation due to local LD, thus we refer to it as the LD-informed gene score. For this calculation we use the methodology described in [38, 37]. However, we use the same window size around a gene as was used for RolyPoly. In addition to validating h_j^{gene}, we use the LD-informed gene score to verify GWAS enrichment in specifically expressed genes of model-identified trait-relevant tissues (i.e., Q-Q plots in the Results section).

If the main objective is to compute gene values, h_j^{gene}, and unbiased parameter estimates are not required, then we include a penalty on the ℓ_1 norm of the annotation coefficients. The penalty strength is modulated with a λ tuning factor which is chosen based on cross validation. Regularization has the beneficial effect of shrinking parameter estimates of irrelevant tissues and can result in higher gene score prediction accuracy.

Simulation setup

For clarity we denote generated parameters and data with an asterisk ($*$). In simulation results reported we used 2×10^4 genes, five simulated gene annotations and one simulated SNP annotation. We generated gene expression, a^*, from a standard χ^2-distribution, and allele frequency as an example SNP annotation, b^*, from a standard uniform distribution. Recall that our model annotation coefficients determine the influence these annotations will have on SNP effect sizes. For each simulated data set 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 $\nu_i^* = \tau_{g(i)}^* + \sum_{l=1}^{P} \phi_l^* b_{il}^*$, 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,

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

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, ϕ_l^*, and gene annotation effects γ_k^*.

Although our method assumes 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

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

Varying the fraction of causal genes, parameter c, across simulated data sets, 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 [39]. 2) We downloaded single-cell RNA sequencing data from Ziesel et al., containing data for 3005 single cells from the hippocampus and cerebral cortex of mice [71]. 3) We obtained human single-cell RNA sequencing data of cortex samples from Darmanis et al., [7]. 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. In the case of single-cell expression data we took the average of single-cell expression vectors for common previously defined cell type classes. To compare across genes, we scale, center, and then square expression values across annotations. When using an expression database from mice, we only used orthologous protein coding genes with a one-to-one functional mapping (based on the definition in Ensembl's BioMart [31]).

We downloaded publicly available GWAS summary statistics from 10 traits from their respective publications: Schizophrenia [43], late-onset Alzheimer's disease [36], four metabolic traits from [19] (HDL cholesterol, LDL cholesterol, total cholesterol and Tryglyceride levels), educational attainment [44], height [70], extreme body mass index [3], and age-related cognitive decline [8]. We restricted or analysis to the autosomes, removed the *MHC* region for immune traits (chromosome 6 between 25 and 34 Mb), and removed rarer variants (MAF $< 0.1\%$). For late-onset Alzheimer's disease and age-related cognitive decline, 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* (chromosome 19 between 44909011 and 45909011). All referenced genome coordinates are from hg19.

Differential gene expression analysis

For the analysis of h_j^{gene} enrichment in differentially expressed genes of Alzheimer's patients, we downloaded microarray gene expression data from 230 samples of the prefrontal cortex [72]. We used `Limma` to perform a differential gene expression analysis between patient and control tissues [55]. Probes were mapped to genes using a mapping downloaded from Ensembl's BioMart [31]. 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 differentially expressed genes.

Calculating RolyPoly gene score enrichment accounting for correlations among gene expression values

To assess the enrichment of RolyPoly gene scores among differentially expressed genes we calculate the Spearman rank correlation coefficent, ρ_{obs}, between RolyPoly gene scores and a differential expression t-statistics. A large value of ρ_{obs} indicates enrichment of large RolyPoly gene scores among differentially expressed genes. Assessing the significance of ρ_{obs} by considering each gene as independent will be anti-conservative because of correlation between gene expression levels of co-regulated genes. To account for this, we generate an empirical sampling distribution for ρ under the null of no association between RolyPoly scores and t which accounts for gene expression correlation.

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

$$t_j = \frac{\bar{x}_j^A - \bar{x}_j^H}{\sqrt{\frac{s_j^A}{n^A} + \frac{s_j^H}{n^H}}} \tag{9}$$

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

Calculating LD correlation values

We downloaded Phase 3 VCFs of European individuals from the 1000 Genomes Project [1]. We used PLINK v1.90b1b to calculate Pearson's r values of SNPs within the default 1 Mb window [49].

RolyPoly implementation and usage

We implemented our method for use through the `rolypoly` R package, which is made available free 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 γ_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^{annot}. h_k^{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^{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^{annot} values greater than 0.1.

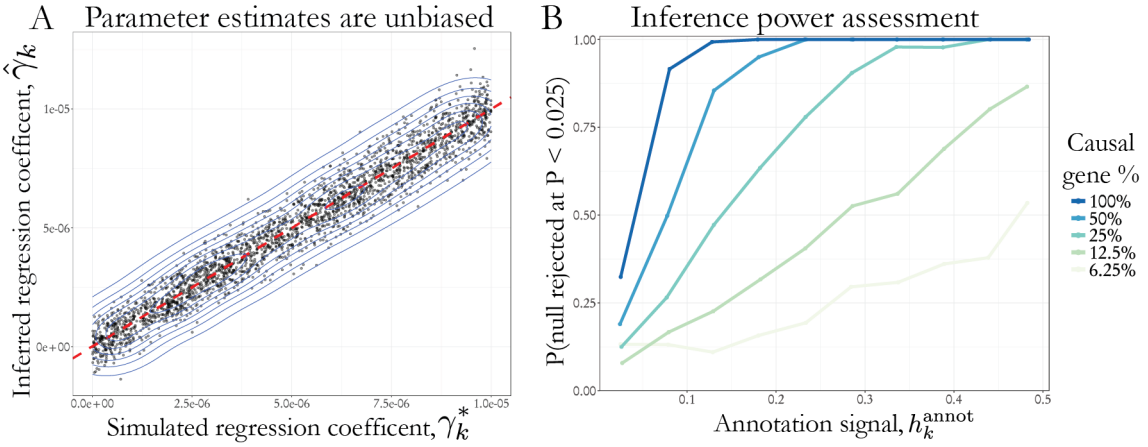

Figure 2: *Simulation results. A) Parameter inference is unbiased and accurate for a range of simulated γ^* effects. Red-dashed line represents the identity line. B) Power as a function of the γ_k^* and annotation values defined as h_k^{annot} in the Material and Methods section. Even when some SNPs 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 consortium (data download and processing described in Material and Methods).

In Table 1, we summarize the top two tissue-trait associations that pass a marginal significance threshold ($p < 0.05$) for seven GWAS traits. With an extreme body mass index GWAS (BMI) we found associations with kidney ($p = 7 \times 10^{-3}$) and thyroid ($p = 0.03$) tissue gene expression. Obesity is known to negatively affect kidney function; however, from existing literature it is ambiguous whether the tissue has a

Trait	Tissue	p-value
Height	Muscle	6×10^{-10}
Height	Pituitary	6×10^{-7}
TC	Liver	2×10^{-4}
TC	Small Intestine	1×10^{-2}
LDL	Liver	2×10^{-3}
LDL	Small Intestine	2×10^{-2}
TG	Adrenal Gland	7×10^{-7}
TG	Liver	2×10^{-2}
BMI	Kidney	7×10^{-3}
BMI	Thyroid	3×10^{-2}
HDL	Liver	7×10^{-3}
EA	Pituitary	3×10^{-2}
EA	Brain	4×10^{-2}

Table 1: *Top trait-relevant GTEx tissue for seven GWAS traits and uncorrected p-values.*

causal role in determining body mass index [23]. There are studies that demonstrate a correlation between thyroid function and weight [53, 32]. We observed a significant enrichment of educational attainment (EA) signal for genes specifically expressed in the pituitary gland ($p = 0.03$) and brain ($p = 0.04$), which corresponds with recent analysis [54, 44]. For height, we detect 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 [11]. Finally, for several metabolic traits (TC, LDL, TG, HDL), there were signals for the liver, small intestine, and adrenal gland, all of which follow known biology.

Next, we examined the total cholesterol (TC) GWAS [19], as its association with liver has been unambiguously reported in the literature. For inference, we used a total of 121,312 SNPs that were within 5kb of a protein coding gene. With p-values from our model we ranked tissues by the strength of association with total cholesterol (see left panel of Figure 3). As expected, liver was the clear top-associated annotation ($p = 2 \times 10^{-4}$), and we estimated an annotation coefficient of 4×10^{-6} (see right panel of Figure 3). Thus, we estimated that the variance of TSS-proximal GWAS effect sizes increase by 4×10^{-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 ($p = 0.05$).

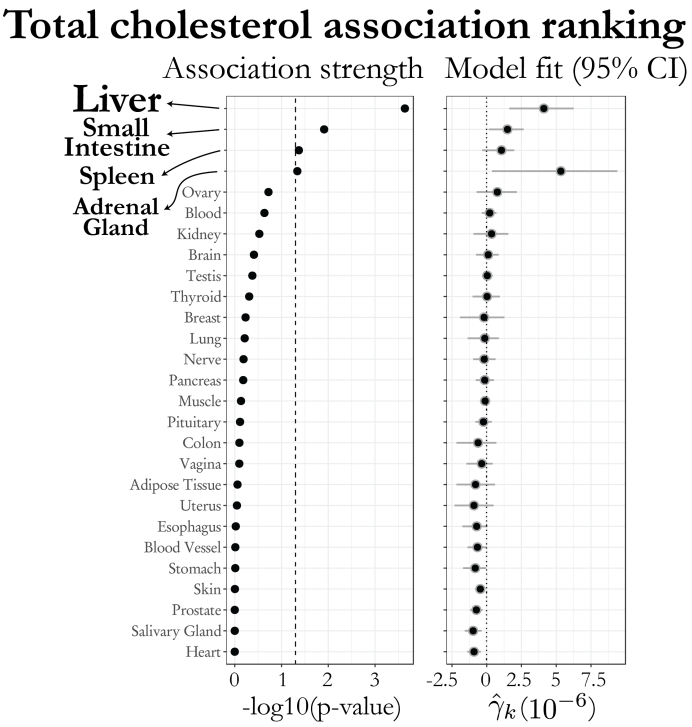

Figure 3: *Total cholesterol and GTEx tissue ranking. Left, tissues were ranked by p-value, which represents the strength of association with total cholesterol. Right, corresponding parameter estimates and 95% confidence intervals.*

While the spleen is primarily thought of as an immune organ, studies show a clear link between splenectomy and lipid metabolism [13]. While the p-value for adrenal gland was identified with a q-value of 0.3, the 95% confidence interval show a wide distribution of non-zero parameter estimates of large positive effect. Considering 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 [42].

We wanted to verify that GWAS effect sizes were enriched for association signal near genes that were specifically expressed in tissues with RolyPoly annotation coefficients significantly greater than zero. 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 using Q-Q plots (Figure 4). To define the set of tissue-specific genes, we sorted normalized expression values for the tissue of

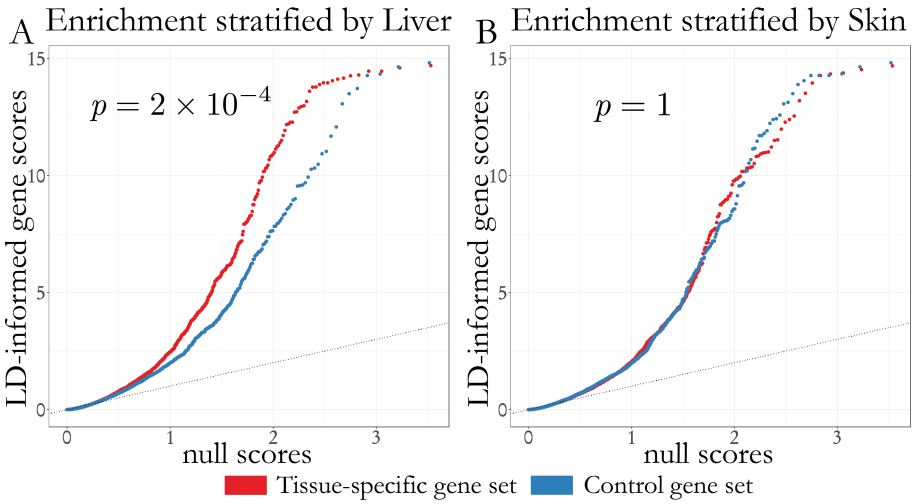

Figure 4: *Total cholesterol and GTEx Q-Q plot comparing enrichment of LD-informed gene scores. Both plots show the p-value from* RolyPoly *for the association between the respective tissue and total cholesterol. A) Q-Q plot comparing enrichment of LD-informed gene scores in genes that are uniquely expressed 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) Similar plot, except stratifying gene values by Skin-specific gene expression, a tissue not predicted to have a role in cholesterol regulation.*

interest by decreasing abundance of normalized gene expression 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 total cholesterol *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 total cholesterol. 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 [7] and several neuropsychiatric traits: age-related cognitive decline, late-onset Alzheimer's disease, educational attainment, and schizophrenia. In total we used 477 human single cells from which gene expression data were

collected. Using a PCA-based clustering approach, the original authors grouped the single cells into 6 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 8 clustered cell types (Figure 5).

Age-related cognitive decline (ACD) is a trait characterized by a decline in cognitive capability and decreases in brain volume, both thought to be a normal function of aging. However, evidence suggests that the rate at which cognitive decline occurs is a precursor to late-onset Alzheimer's disease (AD), hinting at a shared genetic architecture [8]. 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 consists of neurons. Quiescent fetal cells differ from replicating fetal cells in that they have begun to downregulate neuronal growth factors such as EGR1 [7]. On the other hand, we found an association with AD to microglia ($p = 0.03$) and astrocyte ($p = 0.03$) cell types, 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* gene TSS. The significant microglia association persisted ($p = 0.03$) whereas the astrocyte association did not ($p = 0.1$). While the connection between astrocytes and AD is well studied [68], from our analysis this connection appears to be driven by few loci of large effect. Furthermore, there is mounting evidence for a more central role of microglia in AD [18, 56]. However, to our knowledge this is the first human genetics-based enrichment analysis providing evidence for such a connection. Additionally, our results suggest a role for microglia in AD but not ACD. This finding is consistent with recent work demonstrating that while lipid regulation pathways are enriched in GWAS signal for both traits, immune pathways tend to show AD-specific signal [50]. Thus, one could hypothesize microglial involvement during the transition between ACD and AD.

For schizophrenia we found a significant relationship to the oligodendrocyte ($p = 0.02$) and fetal replicating ($p = 0.01$) cell type clusters. The genetic basis of schizophrenia is even less well understood than AD, however there is a significant body of literature studying oligodendrocyte dysfunction and schizophrenia [64, 67]. Moreover, recent genetic association studies have shown an enrichment of schizophrenia GWAS signal within pathways of development [29, 22, 26].

To validate these associations between traits and single-cell cell type clusters, we processed a single-cell data set (see Material and Methods) from mouse brains [71],

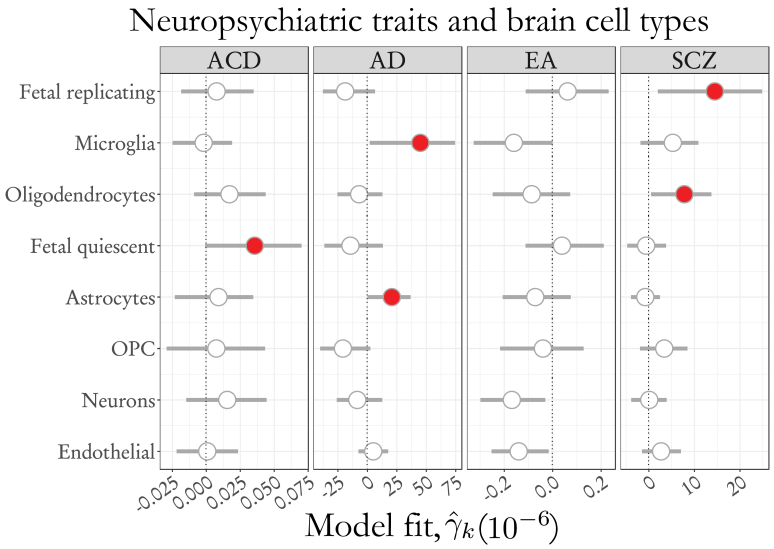

Figure 5: *Neuropsychiatric trait associations with single-cell based cell types. Parameter estimates for age-related cognitive decline (ACD), Alzheimer's disease (AD), Educational attainment (EA) and Schizophrenia (SCZ), and single-cell based cell type clusters from the human brain data set [7]. Range specifies the empirical 95% confidence interval bound. Estimates highlighted in red represent significant associations* ($p < 0.05$).

which included seven major brain cell types that were previously identified. By only utilizing one-to-one human and mouse orthologs, we consider this data set to be an independent pseudo-human brain single-cell data set. Thus, we used this data set to validate our previous findings. We limited our analysis to cell types overlapping in the human and the mouse data set, which 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 demonstrate that if human data is not available one could swap in similar mouse data to guide initial analyses.

RolyPoly gene scores correlate with differentially expressed genes in patients with Alzheimer's

We were interested in studying whether RolyPoly-inferred model parameters could predict trait-relevant genes from an independent data set. Thus, we downloaded and processed gene expression data from human brain samples of 101 control and 129 Alzheimer's disease patients from the prefrontal cortex (see Material and Methods and [72]). A total of 9,228 genes were differentially expressed (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 (independent 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. These values were computed by taking the variance of GWAS effect sizes within 5kb of a gene and incorporating information about LD (see Material and Methods). We detected only weakly suggestive ($p = 0.09$) enrichment of these values within the set of DE genes compared to genes not found to be significantly expressed.

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 for 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 Alzheimer's disease). Within this set of genes we found a significant enrichment ($p = 1 \times 10^{-8}$) of positive values of the differential expression test statistic when compared to a control set of genes (right, Figure 6A). We performed a similar analysis with a cell type for which RolyPoly did not find evidence for AD-association. There was no enrichment of DE summary statistic values within the set of genes specifically expressed in fetal quiescent cells (left, Figure 6A).

From these observations we reasoned we could rank the trait-relevance of genes based on RolyPoly-inferred parameter estimates, $\hat{\boldsymbol{\gamma}}$, and gene expression. As an example for Alzheimer's disease, a gene that is specifically expressed in microglia and astrocyte cells would be higher ranked than a housekeeping gene. Thus, we defined the RolyPoly trait-relevance gene score h_j^{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^{gene}. Furthermore, we hypothesized h_j^{gene} values could predict differ-

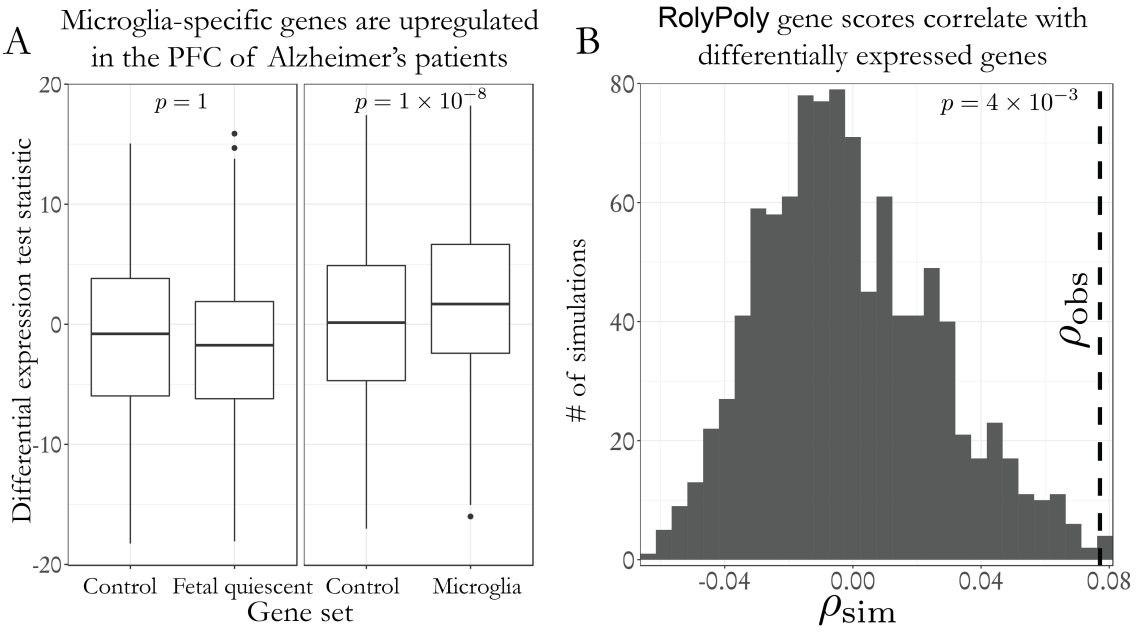

Figure 6: RolyPoly-*inferred model parameters predict differentially expressed genes in the prefrontal cortex (PFC) of Alzheimer's disease patients. A) Differential expression test statistics (a larger value represents genes that are upregulated in the brains of case patients) were significantly larger in the set of genes specifically expressed in the microglia cell type compared with a control gene set (right). We define the set of cell type-specific genes as the top 10% specifically expressed genes. We compared them to the control gene set, which include 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^{gene} *values in differentially expressed genes. The significance of the observed Spearman's rank correlation coefficient between* h_j^{gene} *and the differential expression test statistic was evaluated with a null distribution generated from simulations, which account for the gene expression covariance structure (full details of this test can be found in the Material and Methods).*

entially expressed genes. We found h_j^{gene} scores were significantly enriched within the differentially expressed genes ($p = 7 \times 10^{-18}$, Figure S1). 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 differentially expressed genes and h_j^{gene} values (see Figure 6B).

For validation, we were interested in replicating our enrichment of h_j^{gene} in differen-

tially expressed genes in an independent data set. Sekar et al., performed laser capture microdissection to isolate astrocytes from 10 healthy controls and Alzheimer's patients, and then identified 227 differentially expressed genes [58]. Of those genes, we predicted RolyPoly gene scores for 150 (the others were excluded because they were not measured in the single-cell expression database). We replicated our previous result and identified a significant ($p = 1 \times 10^{-3}$) enrichment within differentially expressed genes (Figure S2). We were unable to perform our enrichment test that accounts for gene correlations, because gene expression data was not available for this data set.

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 based on 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 are increasing or decreasing in proportion and thus would be consistent with our conclusion about RolyPoly parameters.

Discussion

We 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.

We discuss the following assumptions underlying RolyPoly: i) we focused on *cis*-GWAS effects (as opposed to *trans*), because *cis*-SNPs tend to more consistently have effects, and larger effects, on the regulation of gene expression genome-wide [69, 63, 21, 41, 48]. ii) Our model treats neighboring genes independently even though some may have shared *cis*-SNPs, which could result in correlation among nearby SNP effect sizes. However, we corrected for this effect by performing block bootstrap when computing standard errors and empirical confidence intervals. iii) As this is a joint analysis (we estimate all annotation parameters at the same time), inclusion or exclusion of 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, we re-analyze a trait GWAS as more data become available. Secondly, we recommend a cautious interpretation of model

parameters, which should be guided by domain knowledge. Finally, with highly-correlated annotations, one could carry out an initial round of feature selection before performing standard inference or include regularization (described in 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.

To the best of our knowledge, this is the first attempt to connect single-cell gene expression and genome-wide summary statistics from GWAS to identify relevant cell types and genes. While there is evidence linking the immune system and microglia to Alzheimer's disease [18], we identified for the first time 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 [52]. Thus, there is a need for methods such as RolyPoly that can prioritize novel cell types for further study that are relevant to human phenotypes. Here, we focused on single-cells clustered into cell types, however there are numerous alternative groupings to examine. For example, during cell stimulation there exists significant cell heterogeneity even within classical marker-defined immune cell type populations [59, 2]. Using RolyPoly one could link these novel subpopulations to autoimmune disease phenotypes. These analyses should only increase as single-cell data become more commonly available.

It is challenging to pinpoint causal genes from GWAS, because correlations among SNP effects due to LD confound the identification of causal variants. Moreover, it is difficult to identify the target gene modulated by a regulatory variant. Statistical methods that integrate GWAS and eQTLs, while accounting for the effects of LD [34, 24], have proven useful. However, the eQTL data may 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^{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

Derivation of univariate effect estimates

We follow much of the notation and derivation from [60]. Starting with the definition of the annotation coefficients (recalling that the genotype matrix has been scaled),

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

we substitute the GWAS model,

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

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

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

In the Material and Methods section we write the above expression with matrix notation. Others have described a similar relationship between estimated effects, LD, and the true effect sizes [30].

Derivation of distribution parameters of effect estimates

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

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

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

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

where $\boldsymbol{D}$ refers to the diagonal matrix of SNP effect size variances and 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 ℓ_2 norm of a random vector drawn from a mean 0 multivariate normal distribution is the trace of the covariance matrix [37, 40]. Thus, the expected value of the sum of squared SNP effect sizes near gene j is given by,

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

This was derived 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 $\hat{\beta}_i \sim \mathrm{N}(0, \sigma_e^2 n^{-1} + (\boldsymbol{R}_{S_j}\boldsymbol{D}_{S_j}\boldsymbol{R}_{S_j})_{ii})$. Finally,

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

Relationship to previous work

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

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

Note that we went from the first to the second line by substituting ν from equation 3. Although the models share some similarities, our model was derived independently to utilize the full quantitative data from single-cell gene expression assays.

Supplemental Data

Supplemental Information includes two figures.

Acknowledgments

We thank Natalie Telis for producing organ images, Ziyue Gao, Naomi Latorraca, and Nasa Sinnott-Armstrong for feedback and discussion, and Anil Raj for early contributions to method development. Support for D.C. was provided by NLM Training Grant Number T15LM007033. This work was supported by NIH grant 1R01HG008140-01A1 and by the Howard Hughes Medical Institute.

Web Resources

source code repository, `https://github.com/dcalderon/rolypoly`
CRAN page, `https://cran.r-project.org/package=rolypoly`

References

[1] 1000 Genomes Project Consortium, et al. A global reference for human genetic variation. *Nature 526*, 7571 (2015), 68–74.

[2] Arsenio, J., Kakaradov, B., Metz, P. J., Kim, S. H., Yeo, G. W., and Chang, J. T. Early specification of CD8+ T lymphocyte fates during adaptive immunity revealed by single-cell gene-expression analyses. *Nature Immunology 15*, 4 (2014), 365–372.

[3] 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. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nature Genetics 45*, 5 (2013), 501–512.

[4] Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology 33*, 2 (2015), 155–160.

[5] 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., of the Psychiatric Genomics Consortium, S. W. G., et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics 47*, 3 (2015), 291–295.

[6] 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. FTO obesity variant circuitry and adipocyte browning in humans. *New England Journal of Medicine 373*, 10 (2015), 895–907.

[7] Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Gephart, M. G. H., Barres, B. A., and Quake, S. R. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences 112*, 23 (2015), 7285–7290.

[8] 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. A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiology of Aging 33*, 5 (2012), 1017–e1.

[9] 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. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature 482*, 7385 (2012), 390–394.

[10] Efron, B., and Tibshirani, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* (1986), 54–75.

[11] Eugster, E. A., and Pescovitz, O. H. Gigantism. *The Journal of Clinical Endocrinology & Metabolism 84*, 12 (1999), 4379–4384.

[12] 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. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature 518*, 7539 (2015), 337–343.

[13] Fatouros, M., Bourantas, K., Bairaktari, E., Elisaf, M., Tsolas, O., and Cassioumis, D. Role of the spleen in lipid metabolism. *British Journal of Surgery 82*, 12 (1995), 1675–1677.

[14] Finucane, H., Reshef, Y., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.-R., Genovese, G., Saunders, A., et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *bioRxiv* (2017), 103069.

[15] 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. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics 47*, 11 (2015), 1228–1235.

[16] Fuzik, J., Zeisel, A., Máté, Z., Calvigioni, D., Yanagawa, Y., Szabó, G., Linnarsson, S., and Harkany, T. Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nature Biotechnology 34*, 2 (2016), 175–183.

[17] 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. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology 13*, 1 (2012), R7.

[18] Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H., and Kellis, M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature 518*, 7539 (2015), 365–369.

[19] Global Lipids Genetics Consortium, et al. Discovery and refinement of loci associated with lipid levels. *Nature Genetics 45*, 11 (2013), 1274–1283.

[20] Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature 525*, 7568 (2015), 251–255.

[21] Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., Bell, J. T., Yang, T.-P., Meduri, E., Barrett, A., et al. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature Genetics 44*, 10 (2012), 1084–1089.

[22] 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. Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. *Cell 154*, 3 (2013), 518–529.

[23] Hall, J. E. The kidney, hypertension, and obesity. *Hypertension 41*, 3 (2003), 625–633.

[24] Hormozdiari, F., van de Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics 99*, 6 (2016), 1245–1260.

[25] Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *The American Journal of Human Genetics 89*, 4 (2011), 496–506.

[26] 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. Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Neuroscience 18*, 1 (2015), 154–161.

[27] Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science 343*, 6172 (2014), 776–779.

[28] Junker, J. P., and van Oudenaarden, A. Every cell is special: genome-wide studies add a new dimension to single-cell biology. *Cell 157*, 1 (2014), 8–11.

[29] Kahn, R., 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. Schizophrenia. *Nature Reviews Disease Primers 1* (11 2015), 15067 EP.

[30] Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics 10*, 10 (2014), e1004722.

[31] Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database 2011* (2011), bar030.

[32] Knudsen, N., Laurberg, P., Rasmussen, L. B., Bülow, I., Perrild, H., Ovesen, L., and Jørgensen, T. Small differences in thyroid function may be important for body mass index and the occurrence of obesity in the population. *The Journal of Clinical Endocrinology & Metabolism 90*, 7 (2005), 4019–4024.

[33] 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. Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Research 25*, 12 (2015), 1860–1872.

[34] Kumasaka, N., Knights, A. J., and Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics 48*, 2 (2016), 206–213.

[35] Kunsch, H. R. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* (1989), 1217–1241.

[36] Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., Jun, G., DeStefano, A. L., Bis, J. C., Beecham, G. W., et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics 45*, 12 (2013), 1452–1458.

[37] Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS Computational Biology 12*, 1 (2016), e1004714.

[38] 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., et al. A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics 87*, 1 (2010), 139–145.

[39] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. The genotype-tissue expression (GTEx) project. *Nature Genetics 45*, 6 (2013), 580–585.

[40] Mathai, A. M., and Provost, S. B. *Quadratic forms in random variables: theory and applications*. M. Dekker New York, 1992.

[41] Montgomery, S. B., and Dermitzakis, E. T. From expression QTLs to personalized transcriptomics. *Nature Reviews Genetics 12*, 4 (2011), 277–282.

[42] Nussey, S. S., and Whitehead, S. A. *Endocrinology: an integrated approach*. CRC Press, 2013.

[43] of the Psychiatric Genomics Consortium, S. W. G., et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature 511*, 7510 (2014), 421–427.

[44] 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. W., et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature 533*, 7604 (2016), 539–542.

[45] Ongen, H., Brown, A. A., Delaneau, O., Panousis, N., Nica, A. C., Dermitzakis, E. T., Consortium, G., et al. Estimating the causal tissues for complex traits and diseases. *bioRxiv* (2016), 074682.

[46] 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. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science 344*, 6190 (2014), 1396–1401.

[47] Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics 94*, 4 (2014), 559–573.

[48] Price, A. L., Helgason, A., Thorleifsson, G., McCarroll, S. A., Kong, A., and Stefansson, K. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genetics 7*, 2 (2011), e1001317.

[49] 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., et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics 81*, 3 (2007), 559–575.

[50] 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. Genetic architecture of age-related cognitive decline in african americans. *Neurology Genetics 3*, 1 (2017), e125.

[51] Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M. N., Replogle, J. M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science 344*, 6183 (2014), 519–523.

[52] Regev, A., Teichmann, S., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Gottgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S. K., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundeberg, J., Majumder, P., Marioni, J., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Philipakis, A., Ponting, C. P., Quake, S. R., Reik, W., Rozenblatt-Rosen, O., Sanes, J. R., Satija, R., Shumacher, T., Shalek, A. K., Shapiro, E., Sharma, P., Shin, J., Stegle, O., Stratton, M., Stubbington, M. J. T., van Oudenaarden, A., Wagner, A., Watt, F. M., Weissman, J. S., Wold, B., Xavier, R. J., Yosef, N., and Human Cell Atlas Meeting Participants. The Human Cell Atlas. *bioRxiv* (2017).

[53] Reinehr, T., and Andler, W. Thyroid hormones before and after weight loss in obesity. *Archives of Disease in Childhood 87*, 4 (2002), 320–323.

[54] 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. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science 340*, 6139 (2013), 1467–1471.

[55] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. limma powers differential expression analyses for RNA-

sequencing and microarray studies. *Nucleic Acids Research* (2015), gkv007.

[56] Rogers, J., Mastroeni, D., Leonard, B., Joyce, J., and Grover, A. Neuroinflammation in Alzheimer's disease and parkinson's disease: are microglia pathogenic in either disorder? *International review of neurobiology 82* (2007), 235–246.

[57] 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 risk from complex variation of complement component 4. *Nature 530*, 7589 (2016), 177–183.

[58] Sekar, S., McDonald, J., Cuyugan, L., Aldrich, J., Kurdoglu, A., Adkins, J., Serrano, G., Beach, T. G., Craig, D. W., Valla, J., et al. Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiology of Aging 36*, 2 (2015), 583–591.

[59] 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. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature 498*, 7453 (2013), 236–240.

[60] Shi, H., Kichaev, G., and Pasaniuc, B. Contrasting the genetic architecture of 30 complex traits from summary association data. *bioRxiv* (2016), 035907.

[61] Slowikowski, K., Hu, X., and Raychaudhuri, S. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* (2014), 326.

[62] 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. Patterns of cis regulatory variation in diverse human populations. *PLoS Genetics 8*, 4 (2012), e1002639.

[63] 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. Population genomics of human gene expression. *Nature Genetics 39*, 10 (2007), 1217.

[64] 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. Oligodendrocyte dysfunction in schizophrenia and bipolar disorder. *The Lancet 362*, 9386 (2003), 798–805.

[65] 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. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature 509*, 7500 (2014), 371–375.

[66] Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics 45*, 2 (2013), 124–130.

[67] Uranova, N. A., Vostrikov, V. M., Vikhreva, O. V., Zimina, I. S., Kolomeets, N. S., and Orlovskaya, D. D. The role of oligodendrocyte pathology in schizophrenia. *International Journal of Neuropsychopharmacology 10*, 4 (2007), 537–545.

[68] Verkhratsky, A., Olabarria, M., Noristani, H. N., Yeh, C.-Y., and Rodriguez, J. J. Astrocytes in Alzheimer's disease. *Neurotherapeutics 7*, 4 (2010), 399–412.

[69] Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genetics 4*, 10 (2008), e1000214.

[70] 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. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature Genetics 46*, 11 (2014), 1173–1186.

[71] 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. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science 347*, 6226 (2015), 1138–1142.

[72] Zhang, B., Gaiteri, C., Bodea, L.-G., Wang, Z., McElwee, J., Podtelezhnikov, A. A., Zhang, C., Xie, T., Tran, L., Dobrin, R., et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell 153*, 3 (2013), 707–720.