Supplementary Information for "scTWAS: A powerful statistical framework for single-cell transcriptome-wide association studies"

Zhaotong Lin, Chang Su

Contents

S1 Supplementary Methods

S1.1 A detailed description of the IRLS algorithm

In this section, we provide implementation details of the iteratively reweighted least squares (IRLS) algorithm used to estimate the genetic effects on gene expression in our scTWAS framework. Specifically, we describe three key components: (1) preprocessing with `sctransform`, (2) parameter configuration in penalized regression using `cv.glmnet`, and (3) truncation of regression weights to ensure numerical stability.

Algorithm 1 in the main text requires a vector of initial weights $\boldsymbol{\omega}^{(0)}$ and an overdispersion parameter θ as input for a given gene g. In practice, we apply the `SCTransform()` function in the R package `Seurat` to the pseudo-bulk count matrix constructed from single-cell data of a specific cell type, with all genes and individuals to be analyzed. This function fits a regularized negative binomial model, regressing out technical covariates such as sequencing depth [1]. We then extract the estimated intercept β_{0g} and overdispersion parameter θ_g for gene g using the `SCTResults()` function. The gene-level baseline mean expression (with no genetic effects) for individual i is computed as $s_i\mu_{ig}^{(0)}$ for $\mu_{ig}^{(0)} = \exp(\beta_{0g})$, where s_i is the sequencing depth for individual i. The corresponding variance is given by ${\sigma_{ig}^{(0)}}^2 = {\mu_{ig}^{(0)}}^2/\theta_g$ by the definition of the overdispersion parameter [1, 2]. The variance of the observed counts is then given by $s_i\mu_{ig}^{(0)} + s_i^2{\sigma_{ig}^{(0)}}^2$ and further truncated by a lower bound `min_var` returned by `SCTResults()`. Finally, the initial IRLS weights is calculated as $\omega_i^{(0)} = 1/(s_i\mu_{ig}^{(0)} + s_i^2{\sigma_{ig}^{(0)}}^2)$.

To fit the gene expression prediction model, we use elastic net regression via the `cv.glmnet` function in R (line 4 in Algorithm 1). We set `penalty.factor` to be a vector of ones for all cis-SNPs and zeros for covariates, ensuring that penalization is applied only to genetic predictors while leaving covariates unpenalized. We set `intercept = FALSE` as we incorporate the intercept term in the covariate $\mathbf{C}$, and `standardize = FALSE` since the input features (genotype dosages) have already been standardized. The mixing parameter α for elastic net is fixed at 0.5. The tuning parameter λ is selected through five-fold cross-validation.

Finally, during the IRLS updates, estimated weights can become extreme, leading to unstable algorithm performance. To mitigate this, we apply truncation to the weights used in the weighted least squares step (line 7 in Algorithm 1) by ensuring that the variance remains above the `min_var` value returned by `SCTResults()`.

S1.2 Estimation of hidden batch effects in the ROSMAP single-cell data

We performed principal component analysis (PCA) on sctransform-normalized pseudo-bulk gene expression in microglia, and found a clear separation between groups of samples by the value of PC 1 (Supplementary Figure S9a). We found that the separation can be explained by technical batch effects: samples from a subset of batches (B4, B5, and MAP) are separated from samples from other batches (Supplementary Figure S8), suggesting that PC 1 may capture hidden batch effects. Hence, we applied k-means clustering with the number of clusters equal to 2 on PC 1 and constructed a binary covariate to represent the hidden batch effect. Moreover, this covariate also separates samples in other cell types (Supplementary Figure S9b-d), suggesting that the hidden batch effect is universal across cell types. Consequently, we choose to include it as a covariate in the Stage 1 prediction models for all cell types.

S2 Supplementary Tables

(a) Non-null Stage 1 and null Stage 2.

	$\mathrm{CD4}_{\mathrm{NC}}$	$\mathrm{CD8}_{\mathrm{NC}}$	B_{IN}	$\mathrm{Mono}_{\mathrm{C}}$	NK_{R}	Plasma
AG-TWAS	0.038	0.037	0.023	0.01	0.003	0.003
ZJ-TWAS	0.031	0.046	0.043	0.015	0.006	0.001
scTWAS	0.042	0.046	0.048	0.027	0.009	0.005

(b) Null Stage 1 and null Stage 2.

	$\mathrm{CD4}_{\mathrm{NC}}$	$\mathrm{CD8}_{\mathrm{NC}}$	B_{IN}	$\mathrm{Mono}_{\mathrm{C}}$	NK_{R}	Plasma
AG-TWAS	0.001	0.001	0.001	0.001	0.001	0.001
ZJ-TWAS	0	0.003	0.002	0.001	0.001	0.002
scTWAS	0	0.001	0.002	0	0.002	0

Table S1. **Empirical type-I error rates of gene-trait association tests in simulation studies.** (a) Gene expression was simulated as a function of cis-SNPs in Stage 1, and phenotype was simulated to be independent of gene expression in Stage 2. (b) Gene expression was independent of cis-SNPs in Stage 1, and phenotype was also independent of gene expression in Stage 2. A gene-trait association is considered significant if the Stage 1 p-value and Stage 2 p-value are less than 0.05.

S3 Supplementary Figures

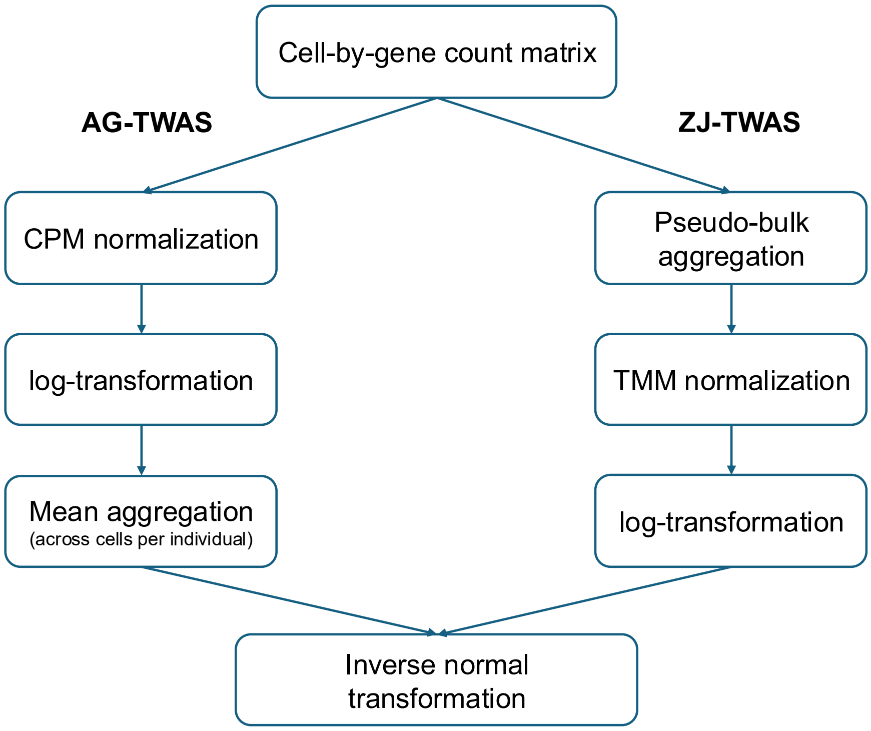

Figure S1. **Normalization steps used in AG-TWAS [3] and ZJ-TWAS [4].** The flowchart illustrates the preprocessing pipelines applied to single-cell or single-nucleus RNA-seq data before applying the traditional TWAS framework in AG-TWAS (left) and ZJ-TWAS (right).

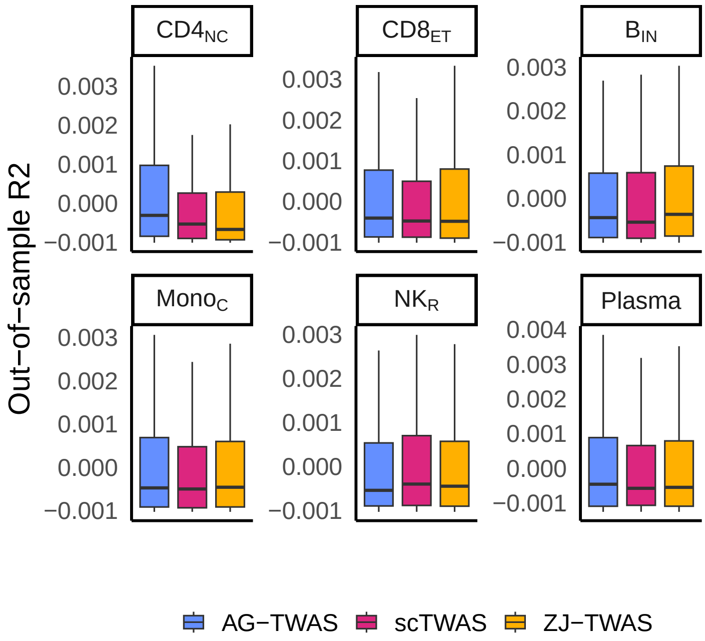

Figure S2. **Out-of-sample R^2 from five-fold cross-validation under the null Stage 1 simulation scenario.** Gene expression was simulated to be independent of cis-SNPs, resulting in no true signal in Stage 1.

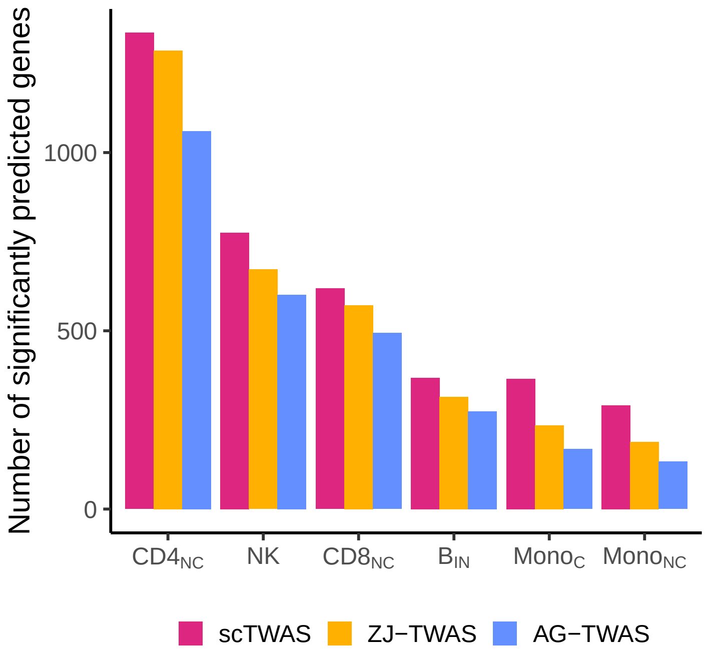

Figure S3. **Number of significantly predicted genes in DICE using GReX models trained on OneK1K scRNA-seq data.** Genes are considered significantly predicted if the nominal p-value for Pearson correlation is less than 0.05.

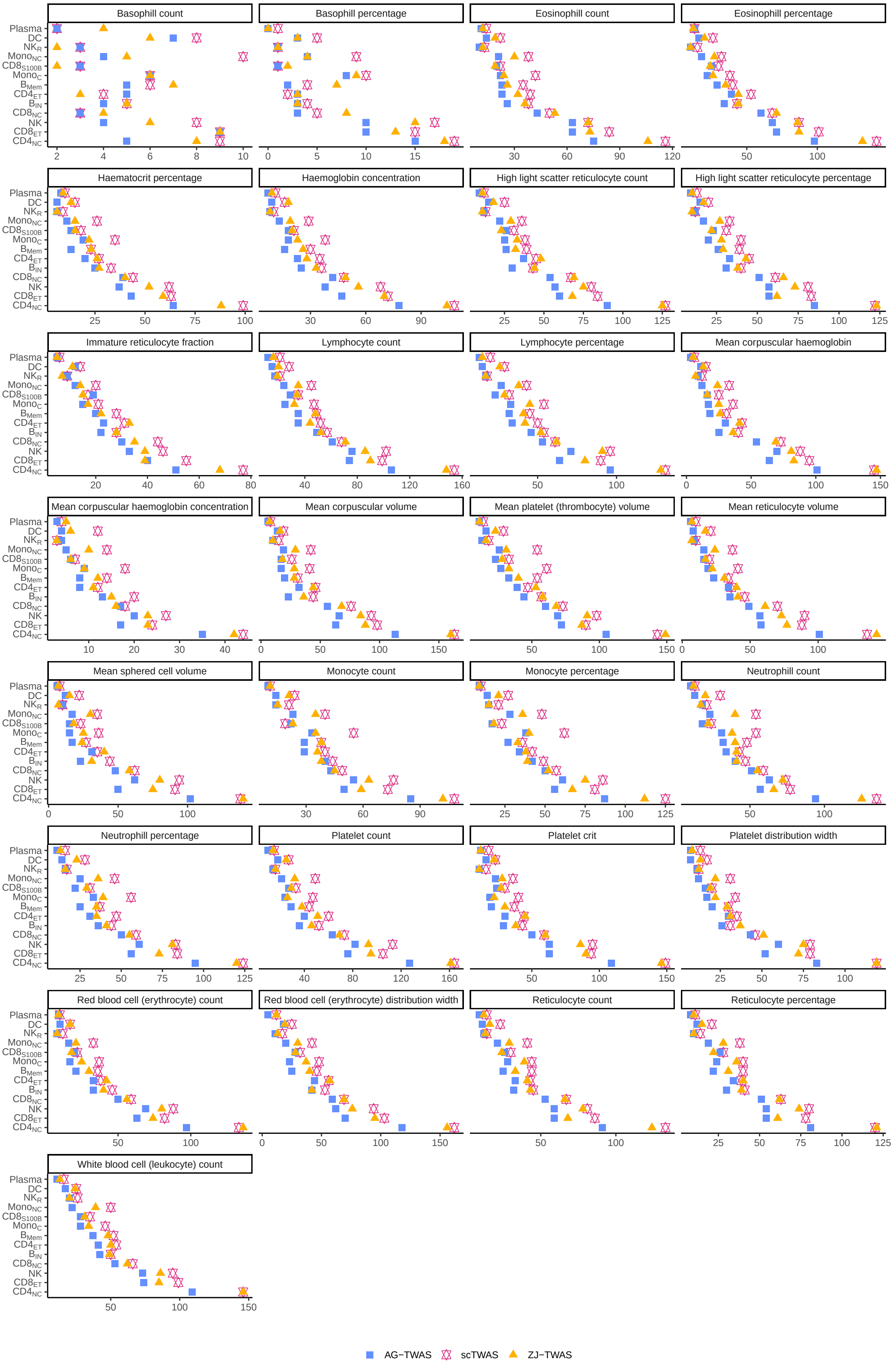

Figure S4. **Number of significant gene-trait associations identified by three methods across 29 blood cell traits.**

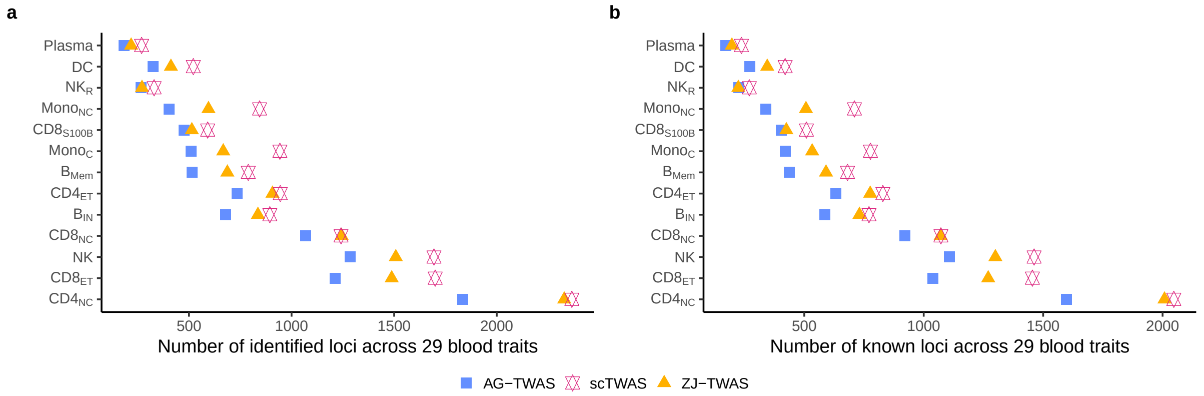

Figure S5. **Results for hematological traits. a.** Total number of TWAS-identified loci and **b.** Number of TWAS loci overlapping known GWAS signals across 29 blood cell traits by three methods.

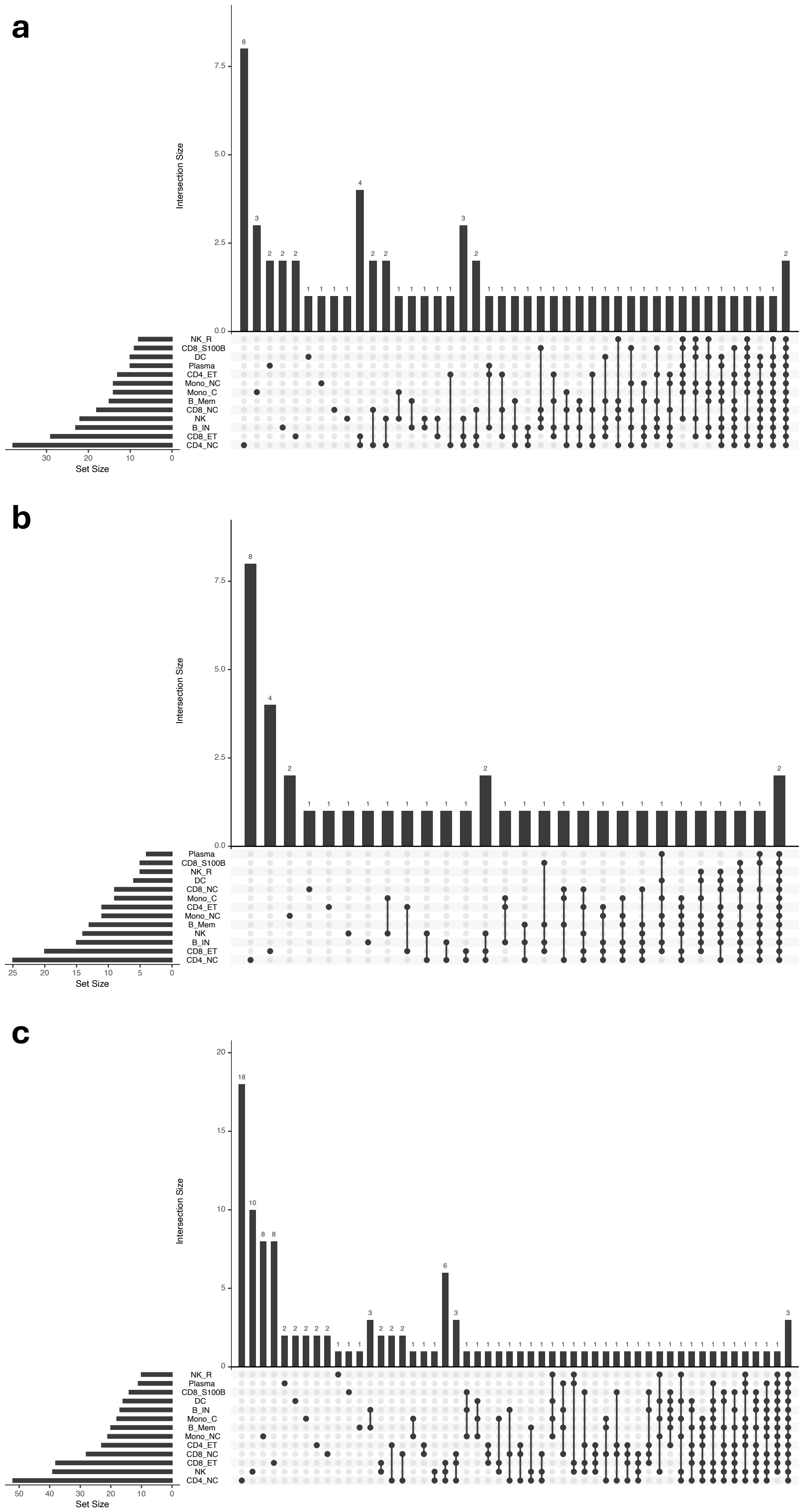

Figure S6. **UpSet plots of scTWAS associations for three immune-related diseases. a.** Rheumatoid arthritis. 21/64 genes are cell-type-specific. **b.** Systemic lupus erythematosus. 20/41 genes are cell-type-specific. **c.** Asthma. 57/111 genes are cell-type-specific.

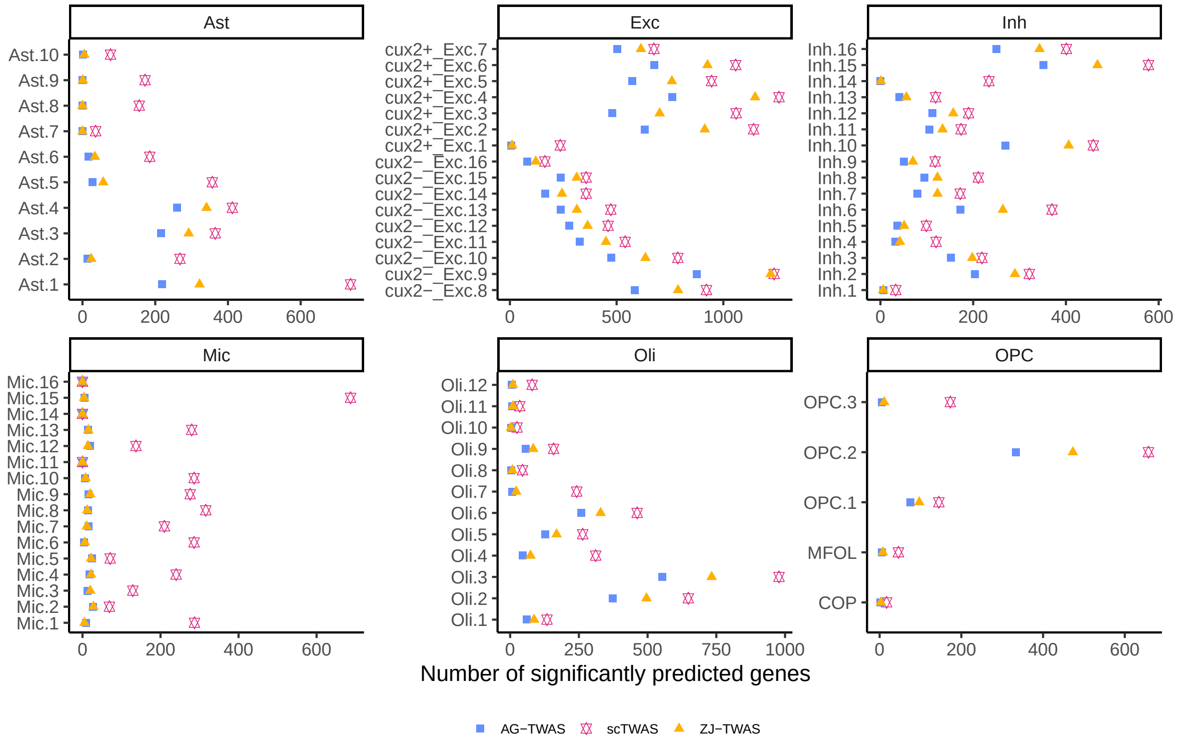

Figure S7. Number of significant Stage 1 models by scTWAS for 75 brain cell subtypes trained with ROSMAP data on dorsolateral prefrontal cortex [5].

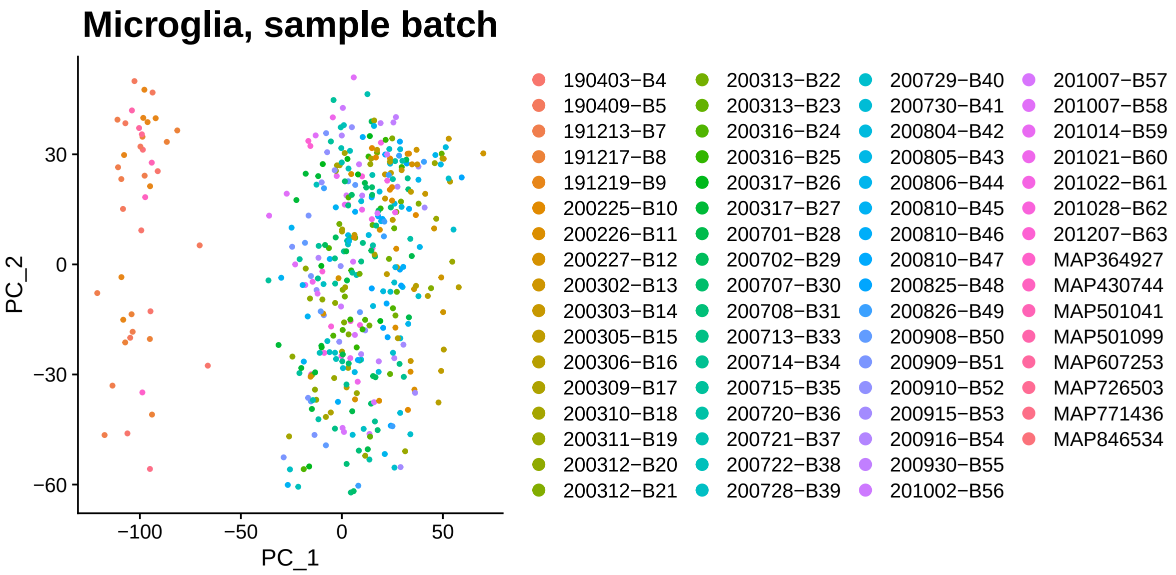

Figure S8. PCA of pseudo-bulk gene expression in microglia colored by technical batches.

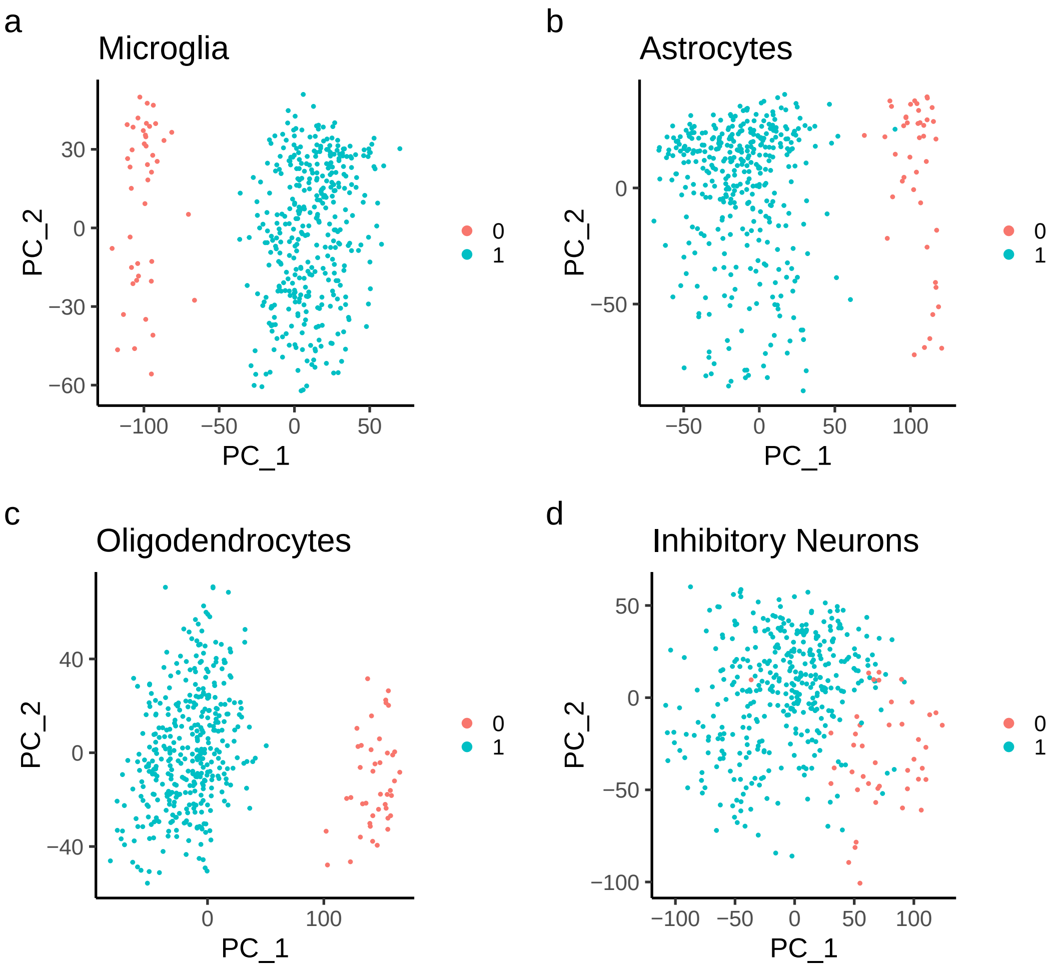

Figure S9. PCA of pseudo-bulk gene expression colored by the estimated hidden batch effect in four cell types.

References

[1] Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology* **20**, 296 (2019).

[2] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* **15**, 1–21 (2014).

[3] Abe, H., Lin, P., Zhou, D., Ruderfer, D. M. & Gamazon, E. R. Mapping dynamic regulation of gene expression using single-cell transcriptomics and application to complex disease genetics. *Human Genetics and Genomics Advances* **6**, 100397 (2025).

[4] Zeng, L. *et al.* A single-nucleus transcriptome-wide association study implicates novel genes in depression pathogenesis. *Biological Psychiatry* **96**, 34–43 (2024).

[5] Fujita, M. *et al.* Cell subtype-specific effects of genetic variation in the alzheimer's disease brain. *Nature Genetics* **56**, 605–614 (2024).