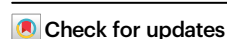


scTWAS: a powerful statistical framework for single-cell transcriptome-wide association studies

Received: 1 May 2025

Accepted: 25 February 2026

Published online: 12 March 2026

Zhaotong Lin¹ & Chang Su^{2,3} 

Transcriptome-wide association studies (TWAS) have successfully identified genes associated with complex traits and diseases, but most have been performed using bulk gene expression data, which aggregate signals across heterogeneous cell types. Population-scale single-cell RNA sequencing data now make it possible to perform TWAS at the cell-type resolution, but present unique challenges due to strong noises, technical variations, and high sparsity. Here, we propose scTWAS, a statistical method to conduct cell-type-specific TWAS using single-cell data. Leveraging a latent-variable model and moment-based estimation to address the challenges of single-cell data, scTWAS consistently improves the prediction of genetically regulated gene expression across cell types in both blood and brain tissues. Compared to existing methods, scTWAS identifies substantially more gene-trait associations across 29 hematological traits and three immune-related diseases in immune cell types. An application to Alzheimer's disease also reveals cell-subtype-specific associations, including *MS4A6A* in the disease-associated microglial subtype and *PPP1R37* in the inflammatory microglial subtype.

Genome-wide association studies (GWAS) have successfully identified numerous genetic variants associated with complex traits and diseases. However, the underlying biological mechanisms driving these associations remain unclear, as the majority of risk variants lie in non-coding regions and may exert their effects through gene regulation^{1–3}. Transcriptome-wide association studies (TWAS) address this challenge by integrating GWAS data with gene expression to identify genes whose genetically regulated expression (GReX) is associated with the trait of interest^{4–6}. It consists of two stages, which first trains a prediction model on the genetic regulation of gene expression with a reference panel of paired genotype and transcriptome data. In Stage 2, these prediction models are applied to GWAS data (either individual-level genotype and phenotype data or GWAS summary statistics) to test for associations between predicted gene expression and the trait of interest. Applications of TWAS have yielded numerous findings on the biological functions underlying disease pathogenesis and have further prioritized concrete targets for intervention and treatment^{7–13}.

Most TWAS to date use reference transcriptome data collected from disease-related tissues, such as peripheral blood mononuclear cells (PBMC) for immune diseases^{13,14} and brain tissues for Alzheimer's disease (AD)^{7,8}, and prioritize genes whose tissue-level GReX is associated with complex diseases. However, tissues consist of biologically heterogeneous cell types, among which the genetic regulation of gene expression differ. For example, expression quantitative trait loci (eQTL) studies have demonstrated distinct genetic regulation across cell types in PBMC^{15–17} and in brain¹⁸. Furthermore, specific cell types may contribute more significantly than others in disease pathogenesis. For example, heritability enrichment analysis shows that microglia carries a sevenfold higher genetic risk for AD compared to other brain cell types¹⁹. Beyond cell types, recent studies also highlight the importance of investigating cell subtypes to better understand disease mechanisms, where cell subtypes represent more refined groups within a cell type defined by distinct gene expression signatures. For instance, a microglia subtype, known as disease-associated microglia,

¹Department of Statistics, Florida State University, Tallahassee, FL, USA. ²Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA.

³Department of Human Genetics, Emory University, Atlanta, GA, USA. ✉ e-mail: chang.su@emory.edu

contributes uniquely to restricting neurodegeneration and has a higher expression of AD risk genes compared to other homeostatic microglia^{20,21} with potentially subtype-specific genetic regulation¹⁸. In contrast, most existing TWAS use reference gene expression data collected on bulk tissues, where mRNA molecules from heterogeneous cell types have been aggregated during data collection. As a result, they may fail to capture genetic regulation and disease associations specific to cell types and subtypes that play critical roles in diseases.

Conducting TWAS at the level of cell types is essential for elucidating the molecular and cellular mechanisms of diseases. While this could be performed with bulk data collected on sorted cell populations, the heterogeneity across cell subtypes may be obscured, as cell sorting primarily relies on well-established markers for major cell types, and the sample size may be insufficient for well-powered statistical analysis^{16,22}. The recently available population-scale single-cell RNA-sequencing data (scRNA-seq) provide an unprecedented opportunity to address these limitations. These data profile hundreds of individuals, each with thousands of cells per individual in disease-related tissues, identifying a dozen of major cell types and tens of refined cell subtypes [e.g., refs. 17,18,23,24]. This enables the modeling of genetic regulation and the identification of disease-associated genes across diverse cellular contexts.

However, the unique characteristics of scRNA-seq data, including strong noises, technical variations, and high sparsity, challenge existing statistical methodologies proposed for TWAS. Count data from scRNA-seq measure underlying gene expression levels in single cells with strong noises and technical variations from the sequencing experiments²⁵. If left unaddressed, these measurement noises would attenuate the estimates of genetic effects, while technical variations across cells introduce artificial differences in mean expression and heteroskedastic variance, even among cells with the same underlying gene expression levels. The counts are also highly sparse due to the limited starting materials in single cells and the low capture efficiency of the technology²⁶. In contrast, most existing TWAS methods were developed for bulk RNA-seq data and are therefore better suited to normalized and continuous expression traits with relatively homoskedastic variance [e.g., refs. 4,5]. This has motivated recent cell-type-specific TWAS analysis to apply heuristic normalization to single-cell data before applying existing TWAS methods^{18,27–30}. However, these normalization strategies may lack theoretical justification and fail to account for the unique characteristics of single-cell data^{25,26,31}, resulting in reduced accuracy in learning the genetic regulation of gene expression and diminished power to detect associations between GREX and diseases.

To address these challenges, we propose scTWAS, a statistical method for conducting cell-type-specific TWAS with scRNA-seq data. scTWAS introduces a latent-variable model that simultaneously models the effects of genetic regulation on the underlying gene expression and the measurement noises and technical variations introduced by sequencing experiments. By disentangling biological gene expression from technical noises, this model enables accurate quantification and estimation of GREX. To improve statistical efficiency under this model, we further introduce a variance-weighted moment-based regression that explicitly accounts for the heteroskedastic noise in estimation. These designs lead to improved accuracy of GREX prediction and higher statistical power of identifying TWAS genes in cellular contexts of interests with scRNA-seq data.

We applied scTWAS on multiple population-scale scRNA-seq datasets to construct GREX prediction models and test the association between predicted GREX and traits in relevant cellular contexts. Our results demonstrate that scTWAS consistently outperformed existing approaches in GREX prediction across cell types and independent datasets. The improvement is particularly pronounced in settings with fewer cells, where GREX is more difficult to predict. Using GREX models for immune cell types, scTWAS systematically identified more

associations with hematological traits and uncovered cell-type-specific regulatory mechanisms underlying genes linked to immune-related diseases, providing new findings that have not been previously reported by GWAS or TWAS with bulk gene expression data. Using GREX models for brain cell types, scTWAS revealed the cellular contexts in which both established and previously unreported genes may contribute to AD pathogenesis. Notably, it uncovered regulatory mechanisms of *MS4A6A* specific to a subtype of disease-associated microglia that are not shared with surveilling or reactive microglia, underscoring the importance of refined cellular resolution and single-cell-based TWAS for understanding the functional impact of disease-associated variants. The GREX prediction models for immune and brain cell types are made publicly available (“Data availability”).

Results

Method overview

scTWAS consists of two stages: Stage 1 constructs cell-type-specific GREX prediction models for each gene using the genotype data and single-cell gene counts from a specific cell type; and Stage 2 tests for the association of predicted GREX and a trait of interest using the trained Stage 1 models and GWAS summary statistics or individual-level data.

In Stage 1, we model pseudo-bulk counts, defined as the total gene counts across all single cells per cell type and individual, in order to mitigate the high sparsity in single cells and to represent individual-level gene expression. For a given gene-cell type pair, let z_i denote the underlying expression level, defined to be the relative abundance of this gene (i.e., the number of mRNA molecules from this gene relative to the total number of molecules) from this cell type in individual i . Let $\mathbf{g}_i$ be a vector of *cis*-genotype (*cis*-single nucleotide polymorphisms, SNPs) in individual i and $\mathbf{c}_i$ be a vector of covariates (e.g., intercept, sex, age, genotype principal components). We assume that

$$z_i \sim F(\mu_i, \sigma_i^2), \mu_i = \mathbf{g}_i' \boldsymbol{\beta} + \mathbf{c}_i' \boldsymbol{\gamma}, \quad (1)$$

where $F(\mu_i, \sigma_i^2)$ denotes a nonnegative distribution with mean μ_i and variance σ_i^2 , and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ denote the coefficients for genotypes and covariates, respectively. This assumes that the underlying gene expression is a random variable whose expectation depends on genotypes and covariates, and whose variance reflects the biological variations across individuals. It makes no parametric assumption on the distribution of z_i , accommodating common assumptions such as Gamma distribution^{26,32}. We call this the expression model, which models the genetic regulation of gene expression in the cell type. In practice, z_i is a latent-variable to be measured by sequencing experiments. To model the measurement process, we use x_i to denote the observed pseudo-bulk counts from individual i and s_i to denote the sequencing depth for individual i , defined as the total counts across all genes from individual i in the cell type of interest. We propose the following measurement model:

$$x_i | z_i \sim \text{Poisson}(s_i z_i). \quad (2)$$

This assumes that the observed gene counts are measured with Poisson sampling noises^{25,26}, whose magnitude further depends on sequencing depths, a technical factor known to substantially vary across samples²⁶. Together, (1) and (2) give the expression-measurement model^{25,31}, which accurately quantifies the genetic regulation of underlying gene expression with $\mathbf{g}_i' \boldsymbol{\beta}$ while accounting for noises and technical variations in measured gene counts. To build GREX prediction models (i.e., $\mathbf{g}_i' \boldsymbol{\beta}$), we propose a penalized moment-based regression to estimate $\boldsymbol{\beta}$, where an iteratively re-weighted least squares (IRLS) method is employed to downweight noisy gene counts with large variance and improve statistical efficiency. Details are presented in Section “scTWAS method”.

In Stage 2, we test associations between predicted GReX and the trait of interest in the cell type of interest. When individual-level GWAS data (both genotype and phenotype) are available, cell-type-level gene expression can be directly predicted using SNP coefficients estimated from Stage 1 (i.e., β), and its association with the phenotype can be tested using regression methods⁴. In most cases where individual-level data are not available, we can perform the test using GWAS summary statistics and a linkage disequilibrium (LD) reference panel⁵ (Section “scTWAS method”).

In summary, scTWAS integrates single-cell data with genetic information to identify cell-type-specific gene-trait associations. By leveraging an expression-measurement model and a moment-based estimation framework, scTWAS effectively accounts for noises and variations from the measurement process and models the genetic regulation of underlying expression levels, eliminating the need for heuristic normalization. It provides more accurate GReX predictions, which then improves the statistical power for association testing with complex traits. This method enhances our ability to uncover biologically meaningful insights at the cell-type-level, advancing our understanding of molecular and cellular mechanisms of complex diseases.

scTWAS achieves higher prediction accuracy and power on simulated data

To evaluate the Stage 1 prediction performance and Stage 2 power of scTWAS, we simulated realistic single-cell gene expression data based on the characteristics of OneK1K study¹⁷, which is one of the first population-scale scRNA-seq datasets, comprising 1.27 million cells from 14 immune cell types across 982 individuals. We evaluated the performance of scTWAS against two representative approaches used in recent cell-type-specific TWAS analyses with scRNA-seq data: NA-TWAS²⁹ and AN-TWAS²⁸. Both approaches used an elastic net model for Stage 1 prediction, but applied different normalization steps to single-cell data. Briefly, NA-TWAS first normalized the single-cell count data before aggregating it based on the suggestion of a recent benchmarking paper³³, while AN-TWAS aggregated the single-cell count data first and applied normalization on the pseudo-bulk count (Supplementary Fig. 1, see Section “Other TWAS methods under comparison”). We simulated GReX with parameters estimated on real data, and the observed single-cell counts based on an expression-measurement model^{25,26,31} (see Section “Simulation settings”), which captured the sparse and heteroskedastic nature of scRNA-seq data, allowing for a realistic evaluation of method performance. We selected six cell types (CD4_{NC}, CD8_{ET}, B_{IN}, Mono_C, NK_R, and Plasma), each representing different levels of cell type abundance in real data, with a mean of 472, 209, 84, 40, 10, and 5 cells per individual, respectively. See Section “Genotype and scRNA-seq data” for the definition of cell types.

We first performed one-sample individual-level simulations where both single-cell gene expression and trait phenotypes were generated at the individual-level using the OneK1K cohort (Section “One-sample individual-level TWAS”). Our results demonstrate that scTWAS consistently outperformed NA-TWAS and AN-TWAS in both prediction accuracy and power across cell types with varying abundance. We evaluate Stage 1 GReX prediction models by the prediction *p*-value from regressing observed expression on the estimated GReX in a fivefold cross-validation (see Section “Prediction model training and evaluation”). Figure 1a shows that scTWAS yielded a greater number of significant Stage 1 GReX prediction models compared to AN-TWAS and NA-TWAS (nominal *p*-value < 0.05). In Fig. 1b, scTWAS also exhibited higher power to detect gene-trait associations than the other two approaches, with the extent of the improvement varying by cell type abundance. Notably, the gain in power was more pronounced in less abundant cell types such as B_{IN} and Mono_C (0.23 and 0.30 higher than AN-TWAS, respectively). This highlights scTWAS’s ability to effectively leverage sparse data and enhance power in challenging scenarios, which are common among immune cell types in the OneK1K dataset,

where 9 out of 13 have abundances less than or equal to B_{IN} (Supplementary Data 1). Even in highly sparse cell types such as NK_R and Plasma, where all methods exhibited low power, scTWAS still maintained an advantage. These results suggest that scTWAS is more powerful in uncovering gene-disease association across cellular contexts of varying abundance, with particularly strong gains in disease-relevant but less abundant cell types and subtypes. This includes microglia (subtypes) for AD as an example, which will be further demonstrated in Section “Cell-subtype-level scTWAS analysis identifies AD-gene associations in microglia subtypes”.

To mimic typical applications, we also performed two-sample simulations in which Stage 2 uses GWAS summary statistics independent of the gene expression cohort (Section “Two-sample TWAS with GWAS summary data”). As shown in Fig. 1c, we considered two expression reference panel sample sizes, $N_{\text{eQTL}} \in \{500, 982\}$, where N_{eQTL} denotes the number of individuals in the expression reference panel used to train the stage 1 GReX model, and three different GWAS sample sizes, $N_{\text{GWAS}} \in \{3 \times 10^4, 5 \times 10^4, 10^5\}$. As expected, power increased with larger expression reference panel sample size and with larger GWAS sample size. Across all settings, scTWAS remained the most powerful method. For the rare NK_R cell type, however, all methods exhibited low power when the gene expression sample size is 500, reflecting the limited predictive performance of the Stage 1 models (Supplementary Fig. 2).

We evaluated the methods’ type-I errors in two scenarios: one where gene expression was independent of the *cis*-SNPs (i.e., a null Stage 1), and another where the phenotype was independent of the gene expression (i.e., a null Stage 2). All methods, including scTWAS, successfully controlled the type-I error rates across all cell types (Supplementary Table 1, Supplementary Fig. 3). Under a null Stage 1, all methods yielded out-of-sample R^2 close to 0 as expected (Supplementary Fig. 4).

We also performed additional simulations to evaluate the three methods in diverse scenarios. First, we varied per-gene trait heritability (Supplementary Fig. 5). As expected, power increased as a gene explained a larger proportion of trait heritability. Second, we considered a sparser genetic architecture with three causal *cis*-SNPs (Supplementary Fig. 6), which reduced Stage 1 predictability and, in turn, Stage 2 power. Third, we introduced LD reference mismatch at Stage 2 by using an out-of-sample LD matrix (Supplementary Section 1.3.1), where results were similar to the matched-LD setting (Supplementary Fig. 7). Across these settings, scTWAS consistently achieved the highest power. Finally, when we introduced horizontal pleiotropy (Supplementary Section 1.3.2), all three methods exhibited inflation (Supplementary Figs. 8 and 9), consistent with the well-known sensitivity of standard TWAS to pleiotropic effects. We note this limitation and discuss potential extensions in the “Discussion”.

scTWAS enhances GReX prediction across cell types in blood and brain tissues

To benchmark GReX prediction models on real data, we applied scTWAS, NA-TWAS, and AN-TWAS to build cell-type-specific models using two population-scale single-cell datasets on PBMC¹⁷ and brain tissues¹⁸, and evaluated their performance based on both within-study and cross-study prediction accuracy (“Methods”). In the first application, we used OneK1K genotype and scRNA-seq data to build prediction models for 13 immune cell types. The average number of cells per individual varied widely across cell types, from 472 in CD4_{NC} to 5 in Plasma, with a median of 49 (Supplementary Data 1). We built prediction models for genes whose total counts in the cell type are greater than 1000.

We assessed the accuracy of scTWAS, NA-TWAS, and AN-TWAS GReX prediction models using the prediction *p*-value in a fivefold cross-validation (see Section “Prediction model training and evaluation”). As shown in Fig. 2a, scTWAS identified more significantly

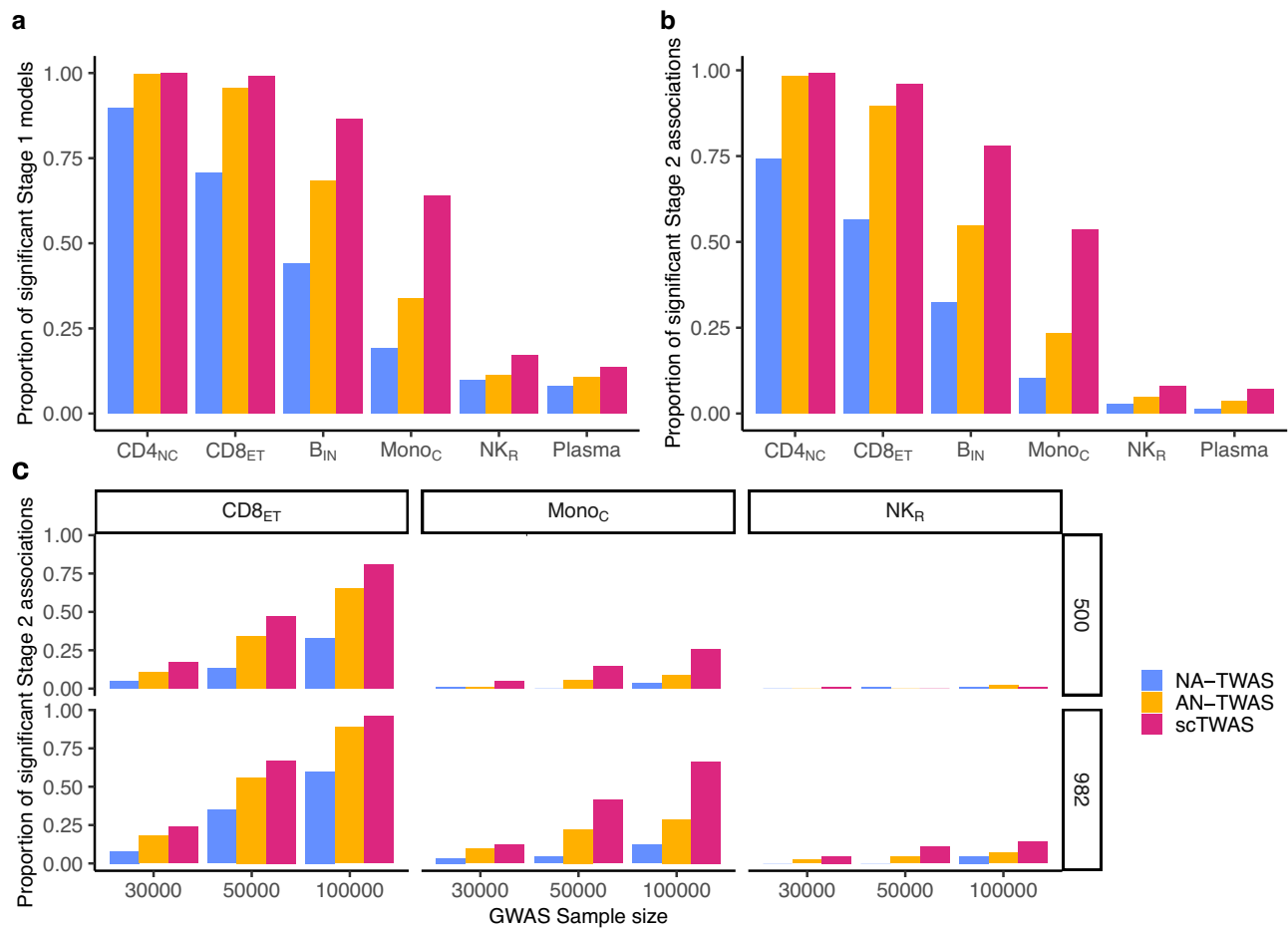


Fig. 1 | Simulation results. Performance evaluation of scTWAS, AN-TWAS, and NA-TWAS in one-sample individual-level simulations (**a**, **b**) and two-sample simulations with GWAS summary statistics (**c**) based on OneK1K. **a** Proportion of predictive GREx models (nominal p -value < 0.05). The nominal p -values are from a two-sided Wald test of the slope coefficient in a linear regression of observed gene expression on predicted gene expression, evaluated on validation folds within a cross-validation framework. **b** Empirical power to detect gene-trait associations across cell types of varying abundances when the gene's *cis*-SNPs explained 0.05 of the trait variance. **c** Empirical power across $N_{\text{eQTL}} \in \{500, 982\}$ and

$N_{\text{GWAS}} \in \{3 \times 10^4, 5 \times 10^4, 10^5\}$ when the gene's *cis*-SNPs explained 0.1/180 of the trait variance. In (**b**, **c**), two-sided p -values are from the Stage 2 association test and are adjusted for multiple testing within each cell type. The numbers of tests are: CD4_{NC} (2189), CD8_{ET} (1234), B_{IN} (580), Mono_C (476), NK_R (122), and Plasma (91), corresponding to the numbers of significant Stage 1 genes within each cell type based on the real data. CD4_{NC}, CD8_{ET}, B_{IN}, Mono_C, NK_R, and Plasma have a mean of 472, 209, 84, 40, 10, and 5 cells per individual, respectively. Colors represent different methods being compared.

predicted genes across cell types, followed by AN-TWAS and NA-TWAS (Fig. 2b, Bonferroni corrected p -values < 0.05). To further demonstrate the cross-study prediction performance of scTWAS, we applied the prediction models trained on OneK1K to predict cell-type-specific bulk gene expression data from the database of immune cell expression (DICE, eQTL, and epigenomics) project¹⁶ for six immune cell types that match between two independent studies ("Methods"). Figure 2b shows that scTWAS generated a median of 16% and 32% more significantly predicted genes compared to AN-TWAS and NA-TWAS, respectively (nominal p -values < 0.05 for Pearson correlation, Supplementary Fig. 10).

In addition to immune cell types, we further benchmarked scTWAS on GREx prediction models for brain cell types. We leveraged data from the Religious Orders Study and Memory and Aging Project (ROSMAP) study, which collected multi-omics data on post-mortem brain samples, including paired whole genome sequencing data³⁴ and single-nucleus RNA-seq (snRNA-seq) data on dorsolateral prefrontal cortex across 424 individuals¹⁸. The nuclei have been annotated at two nested levels: major brain cell types and their subtypes^{18,21}. Accordingly, we built prediction models for 6 brain cell types and 75 subtypes ("Methods"). The abundances of major brain

cell types range from 1403 cells per individual in excitatory neurons to 135 cells per individual in oligodendrocyte precursor cells (OPC) with a median of 525 cells per individual. For subtypes, the median abundance is 29 cells per individual (SD = 58) (Supplementary Data 1).

Figure 3a compares the prediction p -values in a fivefold cross-validation for microglia, demonstrating that scTWAS outperformed the other two methods across genes in generating more significant GREx models. To further validate prediction performance across studies, we applied the trained models to an independently collected snRNA-seq dataset on frontal cortex with 2.3 million nuclei from 427 individuals from the ROSMAP study, where 396 individuals have paired genotype data available³⁵. As shown in Fig. 3b, scTWAS again achieved more significant prediction p -values compared to the existing methods across the six major cell types. Furthermore, at the subtype level (Fig. 3c and Supplementary Fig. 18), scTWAS identified markedly more predictable genes than the other methods across the 75 subtypes, with the largest gains observed in microglia subtypes. This highlights the advantages of scTWAS in predicting GREx for cellular contexts with sparser data and fewer cells. Taken together, these results on OneK1K and ROSMAP studies (Figs. 2 and 3) demonstrate that scTWAS

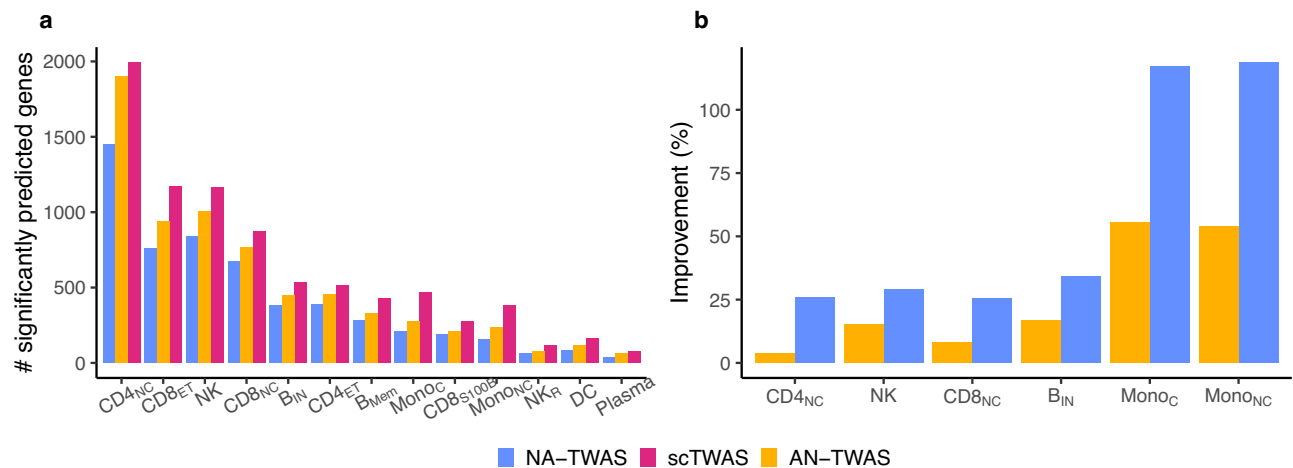


Fig. 2 | Improvement of scTWAS over existing methods in Stage 1 GREX prediction models for immune cell types. **a** Number of significantly predicted genes by three methods on the OneKIK dataset. **b** Percentage increase in the number of

significantly predicted genes by scTWAS compared to NA-TWAS and AN-TWAS, evaluated on the independent DICE validation dataset. Colors represent different methods being compared.

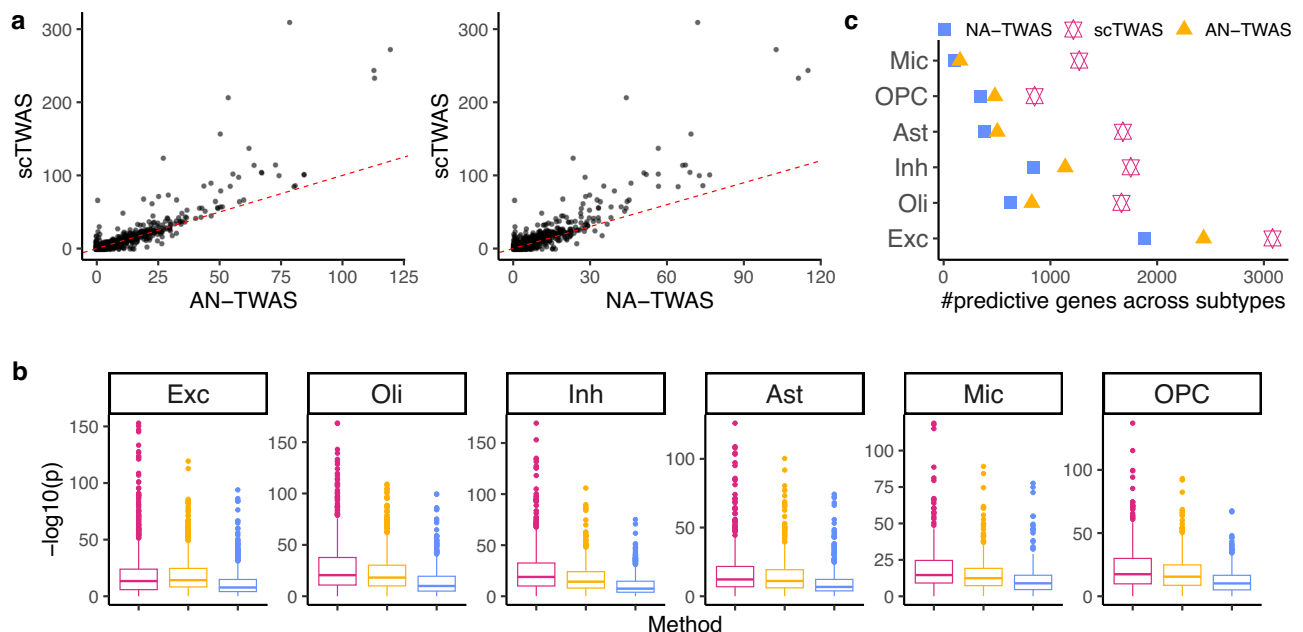


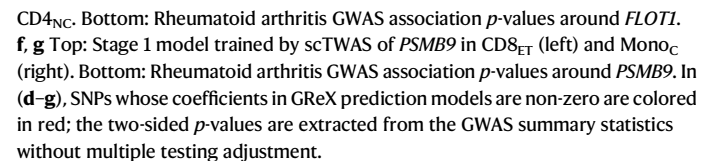
Fig. 3 | Improvement of scTWAS over existing methods in GREX prediction models trained on the ROSMAP datasets. **a** Comparison of $-\log_{10} p$ -values of GREX prediction between scTWAS and AN-TWAS (left), NA-TWAS (right) on microglia from 424 individuals, evaluated within the same study¹⁸ using a cross-validation framework. **b** $-\log_{10} p$ -values of GREX prediction across studies, using models trained with ref. 18 and tested on an independently generated snRNA-seq dataset from frontal cortex with 396 individuals³⁵. Colors represent different methods being compared, see legend in (c). The unadjusted p -values in (a, b) are from a two-sided Wald test of the slope coefficient in a linear regression of observed gene expression on predicted gene expression. Boxplots display the

median (center), the first to the third quartiles (box), and whiskers extending to values within $1.5\times$ the distance between the quartiles. Points indicate outliers beyond this range. The number of observations in each box corresponds to the number of genes evaluated: Exc (1159), Oli (683), Inh (609), Ast (569), Mic (301), and OPC (386). **c** Number of significantly predicted genes by Stage 1 prediction models for brain cell subtypes, aggregated across subtypes to their corresponding major cell types. Excitatory neuron (Exc), oligodendrocyte (Oli), inhibitory neuron (Inh), astrocyte (Ast), microglia (Mic), OPC have a mean of 1404, 748, 556, 495, 181, 135 cells per individual, respectively. Details of the number of cells and individuals are provided in Supplementary Data 1.

consistently generates better GREX prediction across tissue types and across datasets, which lays the foundation for increased power in testing gene-trait associations.

Additionally, to evaluate the gain by cell-type-specific TWAS compared to analyses that do not resolve cell types, we consider bulk-like TWAS analyses in which single cells are aggregated across all cell types to approximate bulk tissue samples, thereby enabling a

controlled comparison using the same single-cell data (Section “Prediction model training and evaluation”). On average, the bulk-like TWAS achieved higher prediction accuracy, which is expected given the larger sequencing depth and the reduced sparsity of aggregated expression data. However, cell-type-specific TWAS improved prediction accuracy for a subset of genes and uncovered regulatory patterns that were obscured in the bulk analysis



types from Section “scTWAS enhances GREX prediction across cell types in blood and brain tissues”, we performed cell-type-level TWAS of 29 hematological traits with GWAS summary statistics (“Methods”). Transcriptome-wide significance was determined using the Bonferroni corrected p -value threshold to account for the total number of genes tested in each cell type. Figure 4a compares the total number of significant gene-trait associations identified by three methods across cell types (also see Supplementary Fig. 11 for a breakdown by individual GWAS). Consistent with the Stage 1 results, scTWAS identified the most

associations across cell types, followed by AN-TWAS and NA-TWAS, and more significant gene-trait associations were identified in more abundant cell types. We further classified the significant genes into loci (“Methods”). If a locus harbors a genome-wide significant GWAS variant, it is considered known; otherwise, it is considered novel. scTWAS increased the total number of loci by an average of 17% and 45% across cell types compared to AN-TWAS and NA-TWAS, respectively, including 18% and 56% more novel loci (Fig. 4b). The most substantial increases were observed in monocytes (both classical and non-classical), with the relative increment in the number of loci being up to 41% compared to AN-TWAS and 95% compared to NA-TWAS.

Additionally, we compared our cell-type-level TWAS results with bulk TWAS on the same traits³⁶, which used bulk RNA-seq data from whole blood on 922 European individuals to train the GReX models. As shown in Fig. 4c, cell-type-level TWAS uniquely identified novel loci that were missed by tissue-level TWAS, with scTWAS identifying the greatest number of novel loci overall. For example, *HHEX* is a member of the homeobox gene family, which plays an important role in embryogenesis and hematopoietic progenitor development³⁷. While both tissue-level and cell-type-level TWAS identified its association with mean platelet volume, eosinophil count, eosinophil percentage, and lymphocyte percentage, the cell-type-level TWAS revealed that these associations were specific to CD8_{ET} and NK cells. This cell-type specificity underscores the unique value of cell-type-level TWAS in pinpointing the relevant cellular contexts where gene expression influences blood cell traits, providing more detailed and biologically meaningful insights than tissue-level analyses. Furthermore, scTWAS uniquely identified an association between *HHEX* expression and monocyte percentage in CD8_{ET} cells, which was not identified by NA-TWAS and AN-TWAS. These findings suggested a potential broader role for *HHEX* in hematopoiesis and immune cell regulation, consistent with its previously reported functions^{38,39}.

Another example is the *DGCR6* locus, where scTWAS uniquely identified its association with mean platelet volume and platelet count in Mono_C, an association that was not discovered with AN-TWAS, NA-TWAS, or bulk level TWAS. scTWAS revealed that the GReX of *DGCR6* was positively associated with platelet count and negatively associated with mean platelet volume, corroborating recent findings that increased *DGCR6* expression was implicated in reducing mean platelet volume⁴⁰. Furthermore, *DGCR6*, located in the 22q11.2 region, is a candidate for involvement in DiGeorge syndrome (22q11.2 deletion syndrome) pathogenesis⁴¹. Individuals with 22q11.2 deletion syndrome have been shown to exhibit lower platelet counts and increased mean platelet volume⁴². Our cell-type-specific findings may offer new insights into the regulatory roles of *DGCR6*, particularly in how its expression in Mono_C might influence platelet dynamics, potentially informing the mechanisms underlying 22q11.2 deletion syndrome. We also compared with the bulk-like TWAS constructed from the same OneK1K dataset (Section “Prediction model training and evaluation”), which confirmed the complementary role of cell-type-specific TWAS relative to bulk TWAS by identifying associations that may be missed in bulk analyses, with scTWAS yielding the largest number of new findings (Supplementary Fig. 15).

scTWAS reveals mechanistic insights into immune-mediated diseases in immune cell types

Rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and asthma are complex immune-mediated diseases that collectively affect millions of individuals worldwide and pose a significant public health burden^{43–45}. GWAS have identified hundreds of risk loci associated with these conditions^{46–48}, underscoring their highly polygenic architectures, but the functional interpretation of these loci remains limited due to the complexity of regulatory mechanisms and cellular heterogeneity in disease-relevant tissues. Immune cell types play a central role in the etiology of these diseases, as T cells, B cells, plasma cells,

and monocytes have all been implicated through functional and transcriptomic studies, highlighting the importance of cell-type-specific analysis in understanding disease mechanisms^{23,24,49}.

Here, we performed cell-type-level TWAS on these three immune-related diseases using cell-type-specific GReX models trained on OneK1K immune cell data from Section “scTWAS enhances GReX prediction across cell types in blood and brain tissues”. In this analysis, scTWAS identified 64, 41, and 111 genes associated with RA, SLE, and asthma, respectively, in at least one of the 13 immune cell types. This corresponded to 11, 15, and 34 more genes than those identified by NA-TWAS, and 6, 6, and 16 more than those identified by AN-TWAS. Moreover, scTWAS uniquely identified 8 genes for RA, 11 for SLE, and 29 for asthma that were not identified by either NA-TWAS or AN-TWAS. We highlight a few examples here, while a complete list of significant genes is provided in Supplementary Tables 3–5. scTWAS uniquely identified an association between *IRF5* and RA and SLE in Mono_C and NK cells, uncovering its cell-type-specific roles in disease mechanisms. *IRF5* is a well-documented transcription factor critical in the regulation of immune responses, particularly in the production of type I interferons, and has been proposed as a therapeutic target for autoimmune diseases⁵⁰. Previous studies have shown that *IRF5* mediates joint inflammation by integrating Toll-like receptor signals and regulating proinflammatory cytokine and chemokine production in a mouse model^{51,52}, and its inhibition suppresses SLE onset and progression in mice^{53,54}. Tissue-based TWAS also identified its association with RA and SLE^{12,55}. Our results corroborate and extend prior evidence by identifying specific immune cell types, Mono_C and NK cells, through which *IRF5* may contribute to SLE pathogenesis. These results support further investigation of *IRF5*, particularly its inhibition in Mono_C and NK cells, as a promising direction for SLE drug discovery. scTWAS also uniquely identified *SULF2* as significantly associated with asthma in plasma cells. Known for its role in modifying heparan sulfate proteoglycans, *SULF2* can impact key biological processes such as cell signaling, inflammation, and tissue remodeling, including airway remodeling, a hallmark of chronic asthma⁵⁶. Notably, while genetic variants in *SULF2* gene have been associated with several diseases, there is currently no reported association with asthma or other lung diseases. Our findings provide evidence linking *SULF2* to asthma pathogenesis, specifically through its potential role in plasma cells, and highlight it as a candidate for future functional studies and therapeutic exploration. Furthermore, a comparison with the bulk-like TWAS showed that scTWAS identified dozens of associations that were not observed in the bulk-like analysis (Section “Prediction model training and evaluation”, Supplementary Fig. 15).

Next, we investigated the similarity and specificity of scTWAS findings across immune cell types. 21 out of 64 RA genes identified by scTWAS were specific to a single-cell type, while the majority were shared across two or more cell types (also see Supplementary Fig. 16 for a detailed UpSet plot). Several possibilities could lead to the cell-type specificity observed in scTWAS results. First, a gene may be expressed exclusively in one cell type. For example, *ZFP57* was highly expressed (total counts > 1000) only in CD4_{NC} cells and was associated with RA in that cell type. Second, a gene may be expressed in multiple cell types, but a predictive GReX model could be obtained only in a specific cell type, potentially due to cell-type-specific genetic regulation. For example, *FLOT1* was expressed in all 13 immune cell types, but only the scTWAS GReX model in Mono_{NC} was predictive, and significant associations with RA, SLE, and asthma were observed only in this cell type. Figure 4d shows the estimated genotypes’ effects by scTWAS for *FLOT1* in Mono_{NC} with a prediction *p*-value of 4.3×10^{-12} . In contrast, the prediction *p*-value is only 0.36 in CD4_{NC} (Fig. 4e), even though it is the most abundant cell type, where stronger predictive performance would be expected if genetic regulation were present. This cell-type-specific signal in Mono_{NC} is further corroborated by the independent DICE study, where the eQTLs associated with *FLOT1* were

found to be more significant in non-classical monocytes¹⁶. Third, the association between GReX and disease may be inherently cell-type-specific. While *PSMB9* was significantly predictable in several cell types, its association with RA was observed only in CD8_{ET} cells. For example, although *PSMB9* was more significantly predicted in Mono_C with a p -value 6.9×10^{-14} (Fig. 4g), as compared to 3.3×10^{-9} in CD8_{ET} (Fig. 4f), there was no corresponding signal in the RA GWAS, resulting in a non-significant TWAS association in Mono_C. Previous research has reported its cell-type-specific involvement in idiopathic inflammatory myopathies in peripheral blood⁵⁷, and further investigation into its function in RA is warranted. Lastly, differences in statistical power due to cell type abundance or sample sizes can also lead to apparent cell-type specificity in detecting gene-disease associations. A similar trend of cell-type similarity and specificity was also observed in SLE and asthma (Supplementary Fig. 16), where among the genes identified by scTWAS, 20 of 41 were cell-type-specific for SLE, and 57 of 111 were cell-type-specific for asthma.

To further explore potential cell-type-specific biological pathways revealed by our cell-type-level TWAS, we performed pathway enrichment analysis with g:Profiler⁵⁸ on significant genes for each disease to test for enrichment of gene ontology (GO) terms. We uncovered several interesting cell-type-specific pathways. In RA, we identified GO terms related to the production of molecular mediator of immune response exclusively in plasma cells. In an earlier study, plasma cells have been found to express receptor activator of NF- κ B ligand, a key mediator in osteoclastogenesis⁵⁹. This expression promotes bone resorption and contributes to periarticular bone loss, highlighting the role of plasma cells in both immune response and bone destruction in RA. In SLE, we found GO terms related to extracellular vesicles (EVs) and exosomes in Mono_{NC}, which play key roles in intercellular communication, carrying inflammatory signals and autoantigens, and thus a crucial physiological and pathological role in SLE⁶⁰. A recent study observed that the Mono_{NC} cells were prominently affected by the exosome, suggesting potential major cell types responsible for exosome treatment⁶¹. In asthma, GO terms related to IgE binding and IgE receptor activity were identified in Mono_C. IgE is a key mediator in allergic responses, and these findings suggest that Mono_C may play an important role in mediating the allergic inflammatory responses through IgE binding^{62,63}.

Cell-subtype-level scTWAS analysis identifies AD-gene associations in microglia subtypes

AD is estimated to affect over 6 million elderly individuals in the United States, and this number is projected to double by 2060⁶⁴. Genetics play an important role in the pathogenesis of AD, with a high heritability of 60%–80%⁶⁵. This motivates large-scale GWAS, which have identified over 70 independent risk loci significantly associated with AD⁶⁶. However, the functional consequences of these AD-associated variants remain to be characterized, especially in the cellular contexts relevant to the pathophysiological processes of AD. Recent advances in highly multiplexed snRNA-seq have facilitated the characterization and identification of these AD-associated cell subtypes in the brain^{21,67}, such as subpopulations of microglia^{21,68–70}, astrocytes^{21,71,72}, and oligodendrocytes^{21,72,73}. Gene expression has been found to be heterogeneous across cell subtypes, and evidence from paired genotype and snRNA-seq data further suggests unique patterns of genetic regulation across cell subtypes¹⁸.

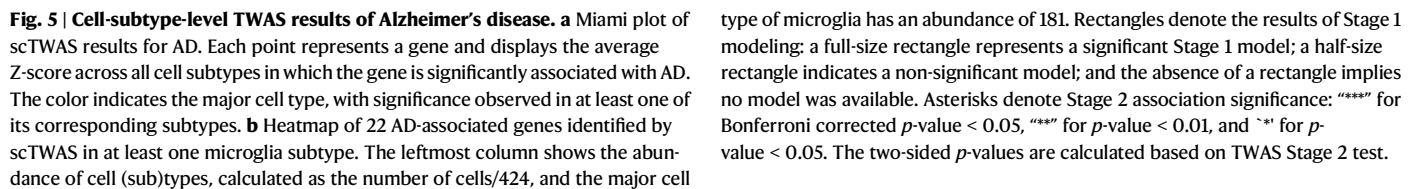
Here, we leveraged scTWAS to investigate the cellular and functional mechanisms of AD-associated variants in brain cell subtypes. Using GReX models trained on ROSMAP dorsolateral prefrontal cortex data (Section “scTWAS enhances GReX prediction across cell types in blood and brain tissues”), scTWAS identified a total of 78 genes associated with AD risk across all cell subtypes, with 33, 18, 20, 21, 22, and 5 genes in subtypes of excitatory neuron, oligodendrocyte, inhibitory neuron, astrocyte, microglia, and OPC, respectively (Supplementary

Data 5). Figure 5 shows that scTWAS reveals the cellular contexts through which AD-associated genes may contribute to disease risk. First, 19 genes were found across multiple cell types (Supplementary Data 5). Notably, *ARL17B* and *KANS1* were identified in all major cell types, suggesting that the mechanisms involving these genes may operate broadly across cellular contexts. For example, *KANS1* was previously suggested to be involved in transcription regulation, a fundamental process likely relevant across diverse cell types⁷⁴. In contrast, many scTWAS signals were cell-type-specific (Supplementary Fig. 17). For example, *CRI*, a gene from an established AD locus⁷⁵, showed strong associations with AD exclusively in oligodendrocyte subtypes (Fig. 5a and Supplementary Data 5). Prior studies suggest that *CRI* may contribute to the clearance of β -amyloid (A β) peptides and modulate immune-related pathways^{75,76}, with additional evidence for its function in neuronal contexts⁷⁷. These scTWAS findings suggest that *CRI* may play a cell-type-specific role in AD pathogenesis through its activity in oligodendrocytes, motivating additional studies of its regulatory mechanisms in this context. Pathway enrichment analysis further reveals the enrichment of immune-related biological processes across cell types, the involvement of lipid pathways in astrocyte (regulation of lipid metabolic process) and microglia (protein-lipid complex assembly), and pathways identified in only one cell type, such as regulation of catalytic activity and negative regulation of receptor-mediated endocytosis in microglia. We further compared with the bulk-like TWAS evaluated on the same ROSMAP dataset (Section “Prediction model training and evaluation”), which shows that scTWAS identified dozens of associations that were not observed in the bulk-like analysis (Section “Prediction model training and evaluation”, Supplementary Fig. 14b).

We next focused on microglia, the brain cell type most enriched for AD genetic risks^{19,78}, to investigate the potential heterogeneity of AD pathogenesis within this cell population. The ROSMAP dataset annotates 16 subtypes of microglia with distinct biological states, including Mic.13, which is associated with lipid and enriched for signatures of disease-associated microglia, and Mic.15, which is enriched for inflammation-related signatures in human²¹. First, several genes were identified as significantly associated with AD across multiple microglia subtypes (Fig. 5b). These include genes located in established AD GWAS loci, whose functional roles in immune pathways or in microglia are supported by prior literature (e.g., *PICALM*⁷⁹, *HLA-DRB1*⁸⁰, *EPHA1-AS1*⁷⁸, *PTK2B*⁸¹, *BIN1*⁸²). In addition, these results included genes such as *DLEU1*, *RASGEF1C* that have not been previously reported in AD GWAS loci¹⁸, warranting future functional studies to investigate their potential roles in AD pathogenesis within microglia. We note that although *APOE* has previously been identified as a AD TWAS gene in microglia⁸³, we did not highlight it in our scTWAS results because it did not pass Stage 1 screening, despite nominally significant Stage 2 associations in proliferate, surveilling, and reacting microglia (nominal p -values < 0.05). Similar results are observed for AN-TWAS and NA-TWAS. This likely reflects the limited sample size, sparsity, and noise of the single-cell dataset, and we discuss plans to further improve GReX prediction in the “Discussion”.

Next, we highlight two examples that demonstrate subtype-specific signals captured by scTWAS. First, scTWAS identified *MS4A6A* as significantly associated with AD specifically in Mic.13, whereas its Stage 1 prediction and Stage 2 association tests were generally insignificant across other microglial subtypes (Fig. 5b). *MS4A6A* has been previously implicated in AD TWAS^{18,84} and is known to play an important role in microglia^{85,86}. Notably, despite the relatively low abundance of Mic.13, the association signal was strong (p -value = 1.4×10^{-10}), suggesting that AD-associated variants may exert strong and subtype-specific regulatory effects on *MS4A6A* in disease-associated microglia, which were not observed in surveilling or reacting microglia.

As a second example, scTWAS identified *PPP1R37* as significantly associated with AD exclusively in Mic.15. *PPP1R37* is located near the



We applied scTWAS to two population-scale single-cell datasets to train GrEx models and perform gene-based association tests across various traits and diseases. Using data from the OneKIK study, we built cell-type-specific GrEx models across 13 immune cell types and conducted TWAS for 29 hematological traits and three autoimmune diseases: RA, SLE, and asthma. We also applied scTWAS

on data from the ROSMAP study to build GREX models across 6 major cell types and 75 cell subtypes in brain tissues, and conducted TWAS for AD. The cell-(sub)type-level TWAS analysis replicated disease-associated genes that were previously identified in tissue-level studies, while providing more granular insights into the cellular contexts in which these associations arise. For example, *IRF5*, a well-established gene associated with RA and SLE, was specifically identified in monocytes and natural killer cells. Importantly, because both the genetic regulation of gene expression and the effect of GREX on disease might differ across cell types, bulk tissue-level analyses may overlook these genes. scTWAS identified novel gene-trait associations that were not identified in previous tissue-level studies, highlighting the necessity of conducting cell-type-level TWAS. Moreover, we also identified cell-type-specific disease-associated genes that appeared uniquely in one specific cell (sub) type. For example, *MS4A6A*, a well-established AD-gene, was identified only in Mic.13, a microglial subtype associated with lipid metabolism and enriched for disease-associated microglia signatures, highlighting its cell-type-specific role and therapeutic relevance. These results show that scTWAS successfully captured known signals and pointed to previously underappreciated cellular mechanisms in disease pathogenesis, which are potentially masked in bulk-level analyses.

A recent study by ref. 89 also introduced a new cell-type-specific TWAS framework called scPrediXcan, which leverages Enformer⁹⁰, a deep learning-based sequence-to-function model, to generate in-silico cell-type-specific gene expression for TWAS analysis. Unlike traditional TWAS methods, scPrediXcan does not use paired genotype and gene expression data across individuals to estimate genetic effects. Instead, it leverages Enformer-derived epigenomic features to generate in-silico cell-type-specific gene expression from DNA sequence alone, and then performs TWAS using these in-silico expression profiles. While scPrediXcan offers an exciting opportunity to extend TWAS to cell types lacking population-level single-cell data, performing TWAS using in-silico expression introduces important considerations. Because the predicted in-silico expression reflects regulatory relationships imputed by a deep learning model, rather than the observed genotype-driven variation in gene expression across individuals, the resulting associations may capture model-driven regulatory patterns. Recent studies have shown that these patterns may have limited consistency with observed genotype-expression associations across individuals^{91,92}. In contrast, scTWAS directly uses paired genotype and scRNA-seq count data and a statistically principled framework to learn genotype-driven variation in gene expression across individuals. Thus, the two methods are designed for different applications. When population-scale single-cell data are not available for the cell type of interest, scPrediXcan provides sequence-based regulatory predictions, though its interpretation depends on the deep learning model's accuracy in predicting epigenomic features across individuals. scTWAS directly leverages population-scale single-cell data, which are increasingly available for diverse tissues and diseases^{24,93,94}, to learn the observed association between genotype and gene expression in cell types.

While scTWAS provides a powerful statistical framework for cell-type-specific TWAS, several limitations should be noted. First, although the method can be conceptually viewed as a two-stage least squares approach, the resulting associations are not necessarily causal. Due to LD, multiple genes in the same region may appear associated, and additional challenges such as horizontal pleiotropy or shared regulatory variants across genes and cell types can complicate interpretation⁹⁵. A natural extension of scTWAS is to incorporate fine-mapping strategies^{96–98} that more precisely distinguish true causal genes from LD-linked neighbors. In addition, recent approaches that fine-map gene-tissue pairs^{99,100} can be extended to

the single-cell setting to provide evidence about whether a gene's effect is specific to a particular cell type or shared across multiple cell types. Integrating such cell-type-aware fine-mapping with scTWAS would further enhance the resolution and interpretability of scTWAS signals and represents an important direction for future methodological development. Second, in its current form, scTWAS trains a separate GREX prediction model for each cell type independently. While this design captures cell-type-specific genetic effects, it does not borrow strength across cell types, which could be valuable given the shared genetic regulation of gene expression in related cell populations²⁷. A promising future direction is to extend the framework to jointly model multiple cell types, with the goal of improving prediction accuracy by sharing information across cell types while still preserving cell-type-specific signals. Third, scTWAS treats all SNPs equally when building the prediction model. However, incorporating functional annotations to prioritize SNPs with regulatory potential, such as chromatin accessibility or transcription factor binding, could further enhance prediction performance as well as model interpretability. This strategy has shown promise in related methods⁶⁹, and we leave this as an important direction for future work. Finally, scTWAS currently operates within a single-ancestry framework. Extending the method to multi-ancestry settings^{101,102} would increase applicability and improve fine-mapping resolution, as differences in LD structure and ancestry-specific genetic effects can provide additional power and resolution beyond single-population analyses. Incorporating these considerations represents a promising avenue for future work.

In summary, we introduce scTWAS, a powerful and biologically grounded statistical framework for performing TWAS using population-scale single-cell data. As these data become increasingly available^{24,93,94}, scTWAS provides a promising tool for building cell-type-specific GREX models and identifying cell-type-specific gene-trait associations for more diseases in disease-relevant cell types. We anticipate that scTWAS will facilitate novel biological insights, refine the mapping from genotype to phenotype, and ultimately contribute to a deeper understanding of disease mechanisms at the cellular level.

Methods

scTWAS method

scTWAS adopts a moment-based estimation framework to estimate β for GREX prediction. For a given pair of gene and cell type, let x_i denote the observed pseudo-bulk counts for the gene in the cell type of interest, s_i the sequencing depth. We use $\mathbf{g}_i$ to denote the *cis*-genotypes for individual i , and $\mathbf{c}_i$ the vector of covariates. The expected underlying expression level for individual i is given by $\mu_i = \mathbf{g}_i'\beta + \mathbf{c}_i'\gamma$, and σ_i^2 denotes the biological variance.

Under the expression-measurement model Eqs. (1) and (2), the observed pseudo-bulk counts x_i satisfy the following moment conditions:

$$E(x_i) = s_i\mu_i = s_i(\mathbf{g}_i'\beta + \mathbf{c}_i'\gamma), \quad (3)$$

$$\text{Var}(x_i) = s_i\mu_i + s_i^2\sigma_i^2, \quad (4)$$

This motivates our scTWAS Stage 1 model:

$$x_i = s_i \cdot (\mathbf{g}_i'\beta + \mathbf{c}_i'\gamma) + \epsilon_i = s_i\mathbf{g}_i'\beta + s_i\mathbf{c}_i'\gamma + \epsilon_i, \quad (5)$$

where ϵ_i 's are independent mean-zero random errors. The variance of ϵ_i is given by Eq. (4), which may vary greatly across samples due to technical variations in sequencing depths. This leads to heteroskedasticity and statistically inefficient estimates if left unaddressed. To mitigate this, we propose to estimate β by minimizing the following

weighted least squares loss function with an elastic net penalty on β :

$$L(\beta, \gamma) = \sum_{i=1}^n \omega_i (x_i - s_i \mathbf{g}_i' \beta - s_i \mathbf{c}_i' \gamma)^2 + \lambda (0.5 \|\beta\|_1 + 0.5 \|\gamma\|_2^2), \quad (6)$$

where we use $\omega_i = 1/\text{Var}(x_i)$ as weights to downweight observations with higher variance and upweight those with lower variance, thereby enhancing statistical efficiency. We implement the optimization of Eq. (6) with R package `glmnet`¹⁰³, where the response is defined as $x_i \sqrt{\omega_i}$, the design matrix consists of genotypes $\mathbf{g}_i$ and covariates $\mathbf{c}_i$ scaled by sequencing depth s_i and $\sqrt{\omega_i}$, and no penalty is imposed on γ . We further use cross-validation to select the hyperparameter λ which minimizes the out-of-sample loss function.

As shown in Eqs. (4) and (3), the weights ω_i 's used in Eq. (6) depend on unknown parameters β (and γ). To address this, we propose an IRLS algorithm that iterates between updating ω_i 's and updating β , γ . In specific, we update β , γ given ω_i 's by optimizing Eq. (6), and update ω_i 's given β , γ by setting $\sigma_i^2 = \mu_i^2 / \theta$ and leveraging Eqs. (1) and (4). Here θ denotes the over-dispersion parameter, encoding a mean-variance dependency as commonly observed in bulk and single-cell RNA-seq data^{26,32}. We estimate θ and the initial weights using R package `scTransform`²⁶ under a null model with no genetic effects. The IRLS algorithm is presented in Box 1, which typically converges within ten iterations. See Supplementary Methods for more details on the algorithm.

For genes that are well-predicted (as evaluated by p -values of prediction in fivefold cross-validation^{4,5}), Stage 2 of scTWAS tests for the association between predicted gene expression and the trait of interest. When only the GWAS summary statistics are available, the association between predicted expression and the trait can be expressed as⁵:

$$\mathbf{z}_{\text{TWAS}} = \frac{\hat{\beta} \mathbf{z}_{\text{GWAS}}}{\sqrt{\hat{\beta}' \mathbf{V} \hat{\beta}}}, \quad (7)$$

where $\hat{\beta}$ is the estimated SNP weights in Stage 1, $\mathbf{z}_{\text{GWAS}}$ is a vector of GWAS summary Z-scores of the corresponding *cis*-SNPs, and $\mathbf{V}$ is the LD correlation matrix of the *cis*-SNPs, which can be estimated by a reference panel. Under the null hypothesis of no association, $\mathbf{z}_{\text{TWAS}}$ follows a standard normal distribution and the p -value can be calculated analytically.

Box 1: IRLS for building Stage 1 models in scTWAS.

Require: Pseudo-bulk counts of a gene from n individuals $X = (x_i)_{n \times 1}$ in the cell type of interest, standardized genotype matrix of *cis*-SNPs around the gene $\mathbf{G} = [\mathbf{g}_1', \dots, \mathbf{g}_n']'$, covariate matrix (including intercept) $\mathbf{C} = [\mathbf{c}_1', \dots, \mathbf{c}_n']'$, sequencing depths $\mathbf{s} = [s_1, \dots, s_n]'$, initial weights $\omega^{(0)} = [\omega_1^{(0)}, \dots, \omega_n^{(0)}]'$, overdispersion parameter θ , number of iterations $m > 0$

- 1: for $t = 1, \dots, m$ do
- 2: $\mathbf{U}^{(t)} \leftarrow [\mathbf{G} \odot \mathbf{s} \odot \sqrt{\omega^{(t-1)}}, \mathbf{C} \odot \mathbf{s} \odot \sqrt{\omega^{(t-1)}}], \mathbf{y}^{(t)} \leftarrow \mathbf{x} \odot \sqrt{\omega^{(t-1)}}$
- 3: $\triangleright \odot$ denotes the element-wise product (by row).
- 4: $\beta^{(t)}, \gamma^{(t)} \leftarrow \text{cv.glmnet}(\mathbf{y} = \mathbf{y}^{(t)}, \mathbf{x} = \mathbf{U}^{(t)}, \dots)$
- 5: $\triangleright \text{cv.glmnet}$ is a function from R package `glmnet`. See Supplementary Methods for details on how to set other arguments.
- 6: $\mu_i^{(t)} \leftarrow \mathbf{g}_i' \beta^{(t)} + \mathbf{c}_i' \gamma^{(t)}, i = 1, \dots, n$
- 7: $\omega_i^{(t)} \leftarrow 1 / (s_i \mu_i^{(t)} + s_i^2 (\mu_i^{(t)})^2 / \theta), i = 1, \dots, n$
- 8: $\hat{\beta} = \beta^{(m)}$
- 9: return $\hat{\beta}$

Other TWAS methods under comparison

In this work, we compare scTWAS with two existing approaches that use single-cell data for TWAS analysis, referred to as NA-TWAS²⁹, which applies Normalization followed by Aggregation, and AN-TWAS²⁸, which applies Aggregation followed by Normalization. Both

approaches rely on the traditional TWAS framework^{4,5} developed for bulk gene expression data, which assumes the following Stage 1 model on normalized gene expression data $\tilde{x}_i$:

$$\tilde{x}_i = \mathbf{g}_i' \beta + \mathbf{c}_i' \gamma + \tilde{\epsilon}_i, \quad (8)$$

where $\tilde{\epsilon}_i$ denotes random noise. Throughout, we used elastic net penalty to estimate β , γ by minimizing the following loss function

$$\sum_{i=1}^n (\tilde{x}_i - \mathbf{g}_i' \beta - \mathbf{c}_i' \gamma)^2 + \lambda (0.5 \|\beta\|_1 + 0.5 \|\gamma\|_2^2), \quad (9)$$

while other penalty functions on β can also be used. This framework has been widely applied to normalized and continuous bulk gene expression traits with relatively homoskedastic variance^{4,5}.

To be compatible with the above traditional TWAS framework, NA-TWAS and AN-TWAS propose to apply various normalization strategies prior to applying the framework (Supplementary Fig. 1). In specific, NA-TWAS begins by applying counts per million normalization to the cell-by-gene count matrix for a specific cell type, followed by log transformation and mean aggregation across cells for each individual; AN-TWAS first aggregates single-cell counts by individual to create a pseudo-bulk count matrix, then applies TMM normalization (via the R package `edgeR`) and log transformation. Finally, both approaches perform inverse normal transformation.

Data sets

Genotype and scRNA-seq data

The OneK1K study. The OneK1K study comprises scRNA-seq data from 1.27 million PMBCs collected from 982 donors¹⁷. Cells have been classified into 14 immune cell types by the authors, including plasma cells (Plasma); dendritic cells (DC); two B cell types—immature and naïve B (B_{IN}) cells and memory B (B_{Mem}) cells; two natural killer cell types—natural killer (NK) cells and NK recruiting (NK_R) cells; two monocyte cell types—classical (Mono_C) and non-classical (Mono_{NC}) monocytes; three CD4⁺ T cell types—CD4 naïve and central memory T cells (CD4_{NC}), CD4⁺ T cells with an effector memory or central memory phenotype (CD4_{ET}), and CD4⁺ T cells expressing SOX4 (CD4_{SOX4}); and three CD8⁺ T cell types—CD8 naïve and central memory T cells (CD8_{NC}), CD8⁺ T cells with expression of SI00B (CD8_{SI00B}), CD8⁺ T cells with an effector memory phenotype (CD8_{ET})¹⁷. For each cell type, we constructed a pseudo-bulk count matrix by extracting unique molecular identifier (UMI) counts across all cells and aggregating them for each individual. Genes with pseudo-bulk UMI counts less than 1000 were filtered out. Finally, CD4_{SOX4} was excluded from the analysis due to the limited number of genes remained. The number of samples, number of cells, and number of genes retained in the analysis for each cell type is provided in Supplementary Data 1. The genotype data were obtained from the authors, where 483,255 autosomal SNPs were genotyped and imputed to the Haplotype Reference Consortium panel, with only SNPs having an imputation $R^2 > 0.8$ being retained. We further performed quality control using PLINK 2.0¹⁰⁴ to retain variants with minor allele frequency > 0.01 and in Hardy–Weinberg equilibrium ($-\text{maf } 0.01 -\text{hwe } 1e-6$).

The ROSMAP study. The ROSMAP study is designed to study aging and dementia with multi-omics profiling of human brain, including whole genome sequencing (WGS) data³⁴ and population-scale single-nucleus RNA-sequencing data on postmortem brain samples^{18,35}. We performed similar quality control of genotype data using PLINK 2.0. To train Stage 1 prediction models, we used the snRNA-seq data on dorsolateral prefrontal cortex from¹⁸, which profiled 1.65 million nuclei and identified a total of 16 major cell types and 95 cell subtypes. Among 437 individuals with snRNA-seq data, 424 have both genotype and snRNA-seq data available. We excluded major cell types with less

than 10,000 cells and focused on the remaining six major cell types (excitatory neurons, oligodendrocytes, inhibitory neurons, astrocytes, microglia, and OPCs) and their 75 subtypes. For example, there are 16 subtypes of microglia under various biological states in this dataset (Fig. 5b). For each of these cell (sub)types, we constructed pseudo-bulk count matrices similar to the procedure above and filtered out genes with less than 1000 total counts. The number of samples, number of cells, and number of genes retained in the analysis for each cell (sub) type are provided in Supplementary Data 1. The only exception is the analysis of microglia subtypes in Fig. 5, where we focused on the genes with at least 1000 total counts in the microglia major cell type to expand the set of genes studied within the cellular context critical to AD mechanisms. For cross-study validation, we further used an independently generated snRNA-seq dataset on prefrontal cortex³⁵, that profiled 2.3 million nuclei and identified seven major cell types. Among 427 individuals with snRNA-seq data, 396 have both genotype and snRNA-seq data available. We focused on excitatory neurons, oligodendrocytes, inhibitory neurons, astrocytes, microglia, and OPCs to validate the GREX models trained on data from dorsolateral prefrontal cortex¹⁸.

The DICE project. The DICE project collects genotype data and bulk RNA-sequencing data on FACS-sorted immune cells from 91 subjects¹⁶. We used the genotype data from the original publication, which have been imputed with the 1000 Genomes Project phase 3¹⁰⁵. We used the bulk gene expression data from the original publication, which have been normalized with Transcript Per Million (TPM)¹⁰⁶. Among the 15 immune cell types profiled in DICE, we matched six of them to the most similar cell types in OneKIK, including classical Monocytes, non-classical Monocytes, CD16 NK cells, Naive B cells, Naive CD4 T cells (matched to CD4_{NC} in OneKIK), and Naive CD8 T cells (matched to CD8_{NC} in OneKIK).

GWAS summary data. GWAS summary statistics for hematological traits were generated by Neale Lab (<https://www.nealelab.is/uk-biobank>) using UK Biobank (UKB) White ancestry data ($N = 349,856$). GWAS summary statistics for the three immune-related diseases (RA, SLE, and Asthma) were obtained from⁴⁶ ($N_{\text{cases}} = 14,361$, $N_{\text{controls}} = 43,923$)⁴⁷, ($N_{\text{cases}} = 5,201$, $N_{\text{controls}} = 9,066$)⁴⁸, ($N_{\text{cases}} = 88,486$, $N_{\text{controls}} = 447,859$), respectively. These GWAS summary statistics were then imputed using ImpG¹⁰⁷ with OneKIK samples as the reference panel. GWAS summary statistics for AD were obtained from⁶⁶, which were imputed using ImpG with ROSMAP WGS data as the reference panel. We further conducted quality control on the imputed GWAS data, including filtering out imputed SNPs with imputation accuracy $R^2 < 0.6$, and any SNPs with ambiguous alternative alleles.

Simulation settings

One-sample individual-level TWAS. We performed simulation studies to evaluate scTWAS with genotype and gene expression data of 982 individuals from the OneKIK study. We simulated the raw count x_{ij} for the j -th cell of the i -th individual with a Poisson-Gamma model, and the trait of interest y_i of the i -th individual as follows:

$$\mu_i = \mathbf{g}_i \boldsymbol{\beta} + \beta_0, z_{ij} \sim \text{Gamma}(\text{shape} = \theta, \text{scale} = \mu_i / \theta); \quad (10)$$

$$x_{ij} | z_{ij} \sim \text{Poisson}(s_{ij}(z_{ij} + a_i)), \text{ where } \alpha_i \sim \mathcal{N}(0, \beta_0 / 100); \quad (11)$$

$$y_i = \mu_i + e_i, \text{ where } e_i \sim \mathcal{N}(0, \frac{1 - h_g^2}{h_g^2} \text{Var}(\boldsymbol{\mu})), h_g^2 = 0.05. \quad (12)$$

In Eq. (10), $\mathbf{g}_i$ were the standardized genotypes of the *cis*-SNPs around a gene, and 10 causal *cis*-SNPs were randomly selected with the corresponding weights generated from Uniform $((-3 \times 10^{-6}, -5 \times 10^{-9}) \cup (5 \times 10^{-9}, 3 \times 10^{-6}))$. The intercept β_0 was set to 1×10^{-4} , and θ was set to 10. In Eq. (11), the sequencing depth of the j -th cell s_{ij} in a specific cell type was directly extracted from the OneKIK real data. A random intercept α_i was simulated to introduce correlations among cells within the i -th individual¹⁰⁸. In Eq. (12), the GWAS trait y was simulated such that the genetically regulated gene expression accounted for around 5% of its variance ($h_g^2 = 0.05$). We used this relatively large per-gene trait heritability h_g^2 to ensure nontrivial Stage 2 association power given the small GWAS sample size ($N_{\text{GWAS}} = 982$) in one-sample simulations.

We also considered two scenarios to evaluate the type I-error rate of the methods. In the first setting, we simulated the expected underlying gene expression level $\mu_i = \beta_0$ independent of the genotype data (*cis*-SNPs). In the second setting, we simulated the GWAS trait from a standard normal distribution independent of the gene expression.

We selected six cell types (CD4_{NC}, CD8_{ET}, B_{IN}, Mono_C, NK_R, and Plasma) for the simulation, representing different levels of cell type abundance in the real data. Twenty genes were randomly selected and 50 replicates were performed for each gene-cell type pair, with a total of 1000 simulation replicates in a specific cell type. In each replicate, we only replaced the count of the specific gene with the simulated single-cell count, while keeping the counts of other genes the same as the OneKIK real data.

With the cell by gene count matrix, we applied the proposed scTWAS, and NA-TWAS and AN-TWAS described in Section “Other TWAS methods under comparison”. To evaluate the power of TWAS analysis, we defined a significant gene-trait association as one where the prediction p -value < 0.05 in fivefold cross-validation in Stage 1 (see Section “Prediction model training and evaluation”), and the Stage 2 association was also significant with a p -value $< 0.05/M$, where M corresponded to the number of tests in the real data application in each cell type. To evaluate the type-I error, we defined a significant gene-trait association as one where both the Stage 1 GREX prediction in cross-validation and the Stage 2 association were significant (nominal p -value < 0.05).

Two-sample TWAS with GWAS summary data. We extended our simulation design to a two-sample framework. Stage 1 single-cell gene expression data were generated the same way as in Section “One-sample individual-level TWAS” using OneKIK individual-level data, and for Stage 2, GWAS summary data for the complex trait were generated directly.

For a given pair of gene and cell type, we computed the per-SNP genotype standard deviations $D_{jj} = \sqrt{2p_j(1 - p_j)}$, where p_j is the minor allele frequency, and the LD matrix $\mathbf{V}$ from the OneKIK dataset. Given the true (joint) SNP effect on the GWAS trait $\mathbf{b}_{\text{GWAS}}$, the observed marginal GWAS summary statistics were generated as follows¹⁰⁹:

$$\hat{\mathbf{b}}_{\text{GWAS}} \sim \mathcal{N}(\mathbf{D}^{-1} \mathbf{V} \mathbf{D} \mathbf{b}_{\text{GWAS}}, (\sigma^2 / N_{\text{GWAS}}) \mathbf{D}^{-1} \mathbf{V} \mathbf{D}^{-1}), \text{ SE}_j = \frac{\sigma}{\sqrt{N_{\text{GWAS}} 2p_j(1 - p_j)}}, \quad (13)$$

where $\mathbf{D} = \text{diag}(D_{jj})$, N_{GWAS} is the sample size of GWAS, and σ is the residual standard deviation of the GWAS trait. To generate the joint SNP effect on the GWAS trait $\mathbf{b}_{\text{GWAS}}$, we set $\mathbf{b}_{\text{GWAS}} = \gamma \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the SNP effect on the gene expression generated in Stage 1. We then chose $\gamma = \sqrt{h_g^2 / \text{Var}(\mathbf{G}\boldsymbol{\beta})}$, where $\text{Var}(\mathbf{G}\boldsymbol{\beta}) = \boldsymbol{\beta}' \mathbf{D} \mathbf{V} \mathbf{D} \boldsymbol{\beta}$, so that the trait heritability explained by the genetically regulated gene expression is h_g^2 , and $\sigma^2 = 1 - h_g^2$.

In this simulation, we focused on gene *AAK1* on chromosome 2, and evaluated three immune cell types, CD8_{ET} (abundant), Mono_C (less abundant), and NK_R (rare). Stage 1 GReX prediction weights were trained from individual-level data using either a subsample of $N_{\text{eQTL}} = 500$ or all 982 OneK1K individuals. Stage 2 GWAS summary statistics were generated at sample sizes $N_{\text{GWAS}} \in \{3 \times 10^4, 5 \times 10^4, 10^5\}$ under the LD-aware model described above. We considered *cis* architectures in which either 10 or 3 causal *cis*-SNPs had nonzero effects in β . We scaled the effect size to achieve per-gene trait heritability $h_g^2 \in \{0, 0.05/180, 0.10/180, 0.20/180\}$, where $h_g^2 = 0$ represents type-I error (null) settings. For each configuration, 100 simulation replicates were performed.

To assess robustness, we examined an out-of-sample LD setting by resampling an independent LD matrix around the OneK1K LD. We also considered horizontal pleiotropy by adding an independent direct effect of *cis*-SNPs. Full details are provided in the Supplementary Section 1.3.

Real data applications

Prediction model training and evaluation. With the aforementioned genotype and scRNA-seq data in Section “Genotype and scRNA-seq data”, we built Stage 1 prediction models for each gene-cell type pair. For each gene, *cis*-SNPs within 500 kb upstream of the gene transcription start site and 500 kb downstream of the transcription end site were used as predictors. Additionally, to ensure a high overlap between the SNPs in the prediction models and those in the GWAS summary statistics, we only considered SNPs present in the UKB GWAS summary statistics for the OneK1K models, and those present in the AD GWAS⁶⁶ for the ROSMAP models. For OneK1K models, we adjusted for covariates including sex, age, and six genotype principal components provided by the authors. For ROSMAP models, we adjusted for covariates including the first three genotype PCs, age at death, sex, post-mortem interval, study (ROS or MAP), and a hidden batch effect covariate inferred by principal component analysis (Supplementary Methods, Supplementary Figs. 19 and 20). R package *glmnet* was used to train the elastic net model for all three methods with the mixing value $\alpha = 0.5$ and the tuning parameter λ was selected using fivefold cross-validation.

We evaluated the within-study prediction accuracy by fivefold cross-validation. For NA-TWAS and AN-TWAS, we regressed the observed normalized expression on the predicted expression across all predicted folds as implemented in FUSION-TWAS. For scTWAS, we used weighted linear regression to regress the observed pseudo-bulk counts on the predicted $s \odot \hat{G}\hat{\beta}$ with estimated weights across all predicted folds. Only genes significantly predicted in Stage 1, defined as those whose estimated GReX is predictive of observed gene expression in cross-validation (Bonferroni corrected p -value < 0.05), were included in Stage 2 analysis to test for the association with the trait of interest.

We also evaluated the cross-study prediction performance of three cell-type-specific TWAS methods. For the GReX models of immune cell types trained with OneK1K data, we used the external validation dataset from the DICE project¹⁶. We focused on 6 cell types that match between OneK1K and DICE, including classical monocytes, non-classical monocytes, NK cells, naive B cells, naive CD4⁺ T cells, and naive CD8⁺ T cells. For each cell type and method considered, we used the prediction model trained with the corresponding cell type from OneK1K data and the DICE genotype data to predict cell-type-specific gene expression for DICE subjects. We calculated the Pearson correlation between predicted gene expression and log(TPM+1) normalized bulk gene expression data from DICE, and compared the number of significantly predicted genes (nominal p -value < 0.05 , Pearson's correlation) across three methods. For the GReX models of brain cell types trained with ROSMAP data on dorsolateral prefrontal cortex¹⁸, we used an independently generated snRNA-seq dataset on prefrontal

cortex from ref. 35 for external validation. We focused on six cell types that overlap between two studies (excitatory neuron, oligodendrocyte, inhibitory neuron, astrocyte, microglia, and OPC) and used the prediction models trained with the corresponding cell type in ref. 18 and the genotype data to predict cell-type-specific gene expression in ref. 35. The prediction performance of each method is evaluated by prediction p -values, as defined above, following the approach used in cross-validation.

In addition to the three cell-type-specific methods, we also constructed a bulk-like TWAS model for comparison. Specifically, we aggregated cells across all cell types in the single-cell dataset to mimic bulk expression in tissue samples, which enables a controlled comparison using the same single-cell dataset. We then applied the same normalization procedure used in AN-TWAS (see Supplementary Fig. 1). Using the synthetic bulk expression data, we trained bulk GReX prediction models and evaluated their cross-validation prediction accuracy.

TWAS analysis using GWAS summary statistics. We applied cell-type-level TWAS to GWAS summary data using FUSION-TWAS^{5,110}. FUSION-TWAS leverages GWAS summary statistics, a gene expression predictive model, and an LD reference panel to test the association between genetically predicted gene expression and a phenotype of interest, analyzing one gene at a time. Specifically, we conducted TWAS analysis on 29 hematological traits and 3 immune-related diseases using Stage 1 models trained on OneK1K immune cell types with scTWAS, NA-TWAS, and AN-TWAS, respectively, and the OneK1K study as the LD reference panel. We also conducted TWAS analysis on AD using Stage 1 models trained on ROSMAP brain cell (sub)types with scTWAS, NA-TWAS, and AN-TWAS, respectively, and the ROSMAP study as the LD reference panel. To account for the large number of hypotheses tested, we used Bonferroni corrected p -value threshold of $0.05/M$ to identify significant gene-trait associations in a specific cell type of interest, where M was the number of genes significant in at least one prediction model. For comparison with the bulk-like TWAS analysis, we also conducted TWAS using the GReX prediction models trained on aggregated single-cell data.

To define a locus, we started by ranking significant genes according to their TWAS p -values, with higher ranks assigned to smaller p -values. For the top-ranked gene, we defined a locus as a 1 million base pair window around it and included all genes within this region as part of the locus. This process was repeated for the remaining significant genes until all significant genes were associated with loci.

GO enrichment analysis

In Section “Real data applications”, we used *gost* in R package *gprofiler2* (v0.2.2) to perform enrichment analysis for GO with one-sided Fisher exact tests and selected GO terms with adjusted p -value < 0.05 using the default “*g_scs*” method.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The GReX prediction models for immune and brain cell types generated in this study have been deposited in Figshare at [<https://doi.org/10.6084/m9.figshare.31329184>] and [<https://doi.org/10.6084/m9.figshare.31329139>]^{111,112}. The availability for datasets used to train or validate GReX models is as follows: OneK1K genotype and scRNA-seq data are publicly available from the OneK1K study¹⁷. Genotype and gene expression data of immune cell types from the DICE project¹⁶ are available under controlled access in dbGap under accession code phs001703.v5.p1 [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001703.v5.p1], access can be obtained by

requesting Authorized Access on dbGap. WGS and snRNA-seq data from the ROSMAP study are available on Synapse under controlled access with accession codes syn11724057 [<https://www.synapse.org/Synapse:syn11724057>], syn53366818 [<https://www.synapse.org/Synapse:syn53366818>], and syn52293417 [<https://www.synapse.org/Synapse:syn52293417>], access can be obtained by submitting a Data Use Certificate to Synapse.

Code availability

R package for scTWAS is available at <https://github.com/ZhaotongL/scTWAS>¹³. Code for real data analyses is available at https://github.com/ZhaotongL/scTWAS_paper¹⁴.

References

- Emilsson, V. et al. Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
- Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
- Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
- Mai, J., Lu, M., Gao, Q., Zeng, J. & Xiao, J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. *Commun. Biol.* **6**, 899 (2023).
- Sun, Y. et al. A transcriptome-wide association study of Alzheimer's disease using prediction models of relevant tissues identifies novel candidate susceptibility genes. *Genome Med.* **13**, 1–11 (2021).
- Guo, S. & Yang, J. Bayesian genome-wide TWAS with reference transcriptomic data of brain and blood tissues identified 141 risk genes for Alzheimer's disease dementia. *Alzheimer's Res. Ther.* **16**, 120 (2024).
- Reay, W. R. & Cairns, M. J. Advancing the use of genome-wide association studies for drug repurposing. *Nat. Rev. Genet.* **22**, 658–671 (2021).
- Chen, F. et al. Multi-ancestry transcriptome-wide association analyses yield insights into tobacco use biology and drug repurposing. *Nat. Genet.* **55**, 291–300 (2023).
- Hu, S.-y. et al. Synovial transcriptome-wide association study implicates novel genes underlying rheumatoid arthritis risk. *Rheumatology* **64**, 2515–2524 (2025).
- Khunsriraksakul, C. et al. Multi-ancestry and multi-trait genome-wide association meta-analyses inform clinical risk prediction for systemic lupus erythematosus. *Nat. Commun.* **14**, 668 (2023).
- Valette, K. et al. Prioritization of candidate causal genes for asthma in susceptibility loci derived from UK Biobank. *Commun. Biol.* **4**, 700 (2021).
- Wu, C. et al. Transcriptome-wide association study identifies susceptibility genes for rheumatoid arthritis. *Arthritis Res. Ther.* **23**, 1–10 (2021).
- Dimas, A. S. et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325**, 1246–1250 (2009).
- Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715 (2018).
- Yazar, S. et al. Single-cell eqtl mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
- Fujita, M. et al. Cell subtype-specific effects of genetic variation in the Alzheimer's disease brain. *Nat. Genet.* **56**, 605–614 (2024).
- Yang, X. et al. Functional characterization of Alzheimer's disease genetic variants in microglia. *Nat. Genet.* **55**, 1735–1744 (2023).
- Keren-Shaul, H. et al. A unique microglia type associated with restricting development of Alzheimer's disease. *Cell* **169**, 1276–1290 (2017).
- Green, G. S. et al. Cellular communities reveal trajectories of brain ageing and Alzheimer's disease. *Nature* **633**, 634–645 (2024).
- Lopes, K. D. P. et al. Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. *Nat. Genet.* **54**, 4–17 (2022).
- Perez, R. K. et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science* **376**, eabf1970 (2022).
- Natri, H. M. et al. Cell-type-specific and disease-associated expression quantitative trait loci in the human lung. *Nat. Genet.* **56**, 595–604 (2024).
- Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).
- Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
- Thompson, M. et al. Multi-context genetic modeling of transcriptional regulation resolves novel disease loci. *Nat. Commun.* **13**, 5704 (2022).
- Zeng, L. et al. A single-nucleus transcriptome-wide association study implicates novel genes in depression pathogenesis. *Biol. Psychiatry* **96**, 34–43 (2024).
- 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. *Hum. Genet. Genomics Adv.* **6**, 100397 (2025).
- Venkatesh, S. et al. Single-nucleus transcriptome-wide association study of human brain disorders. Preprint at <https://www.medrxiv.org/content/10.1101/2024.11.04.24316495v1> (2024).
- Su, C. et al. Cell-type-specific co-expression inference from single cell rna-sequencing data. *Nat. Commun.* **14**, 4846 (2023).
- Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
- Cuomo, A. S. et al. Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biol.* **22**, 188 (2021).
- De Jager, P. L. et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Sci. Data* **5**, 1–13 (2018).
- Mathys, H. et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. *Cell* **186**, 4365–4385 (2023).
- Rowland, B. et al. Transcriptome-wide association study in UK Biobank Europeans identifies associations with blood cell traits. *Hum. Mol. Genet.* **31**, 2333–2347 (2022).
- Paz, H., Lynch, M. R., Bogue, C. W. & Gasson, J. C. The homeobox gene *hhx* regulates the earliest stages of definitive hematopoiesis. *Blood, J. Am. Soc. Hematol.* **116**, 1254–1262 (2010).
- Jang, S. W. et al. Homeobox protein Hhex negatively regulates Treg cells by inhibiting Foxp3 expression and function. *Proc. Natl. Acad. Sci. USA* **116**, 25790–25799 (2019).
- Watanabe, H. et al. Transcription factor hematopoietically expressed homeobox protein (Hhex) negatively regulates osteoclast differentiation by controlling cyclin-dependent kinase inhibitors. *J. Bone Miner. Res.* **6**, e10608 (2022).

40. Zamariolli, M. et al. The impact of 22q11. 2 copy-number variants on human traits in the general population. *Am. J. Hum. Genet.* **110**, 300–313 (2023).
41. Demczuk, S., Thomas, G. & Aurias, A. Isolation of a novel gene from the DiGeorge syndrome critical region with homology to *Drosophila* *gdl* and to human *LAMC1* genes. *Hum. Mol. Genet.* **5**, 633–638 (1996).
42. Campbell, I. M. et al. Platelet findings in 22q11. 2 deletion syndrome correlate with disease manifestations but do not correlate with GPIb surface expression. *Clin. Genet.* **103**, 109–113 (2023).
43. Smolen, J. S. et al. Rheumatoid arthritis. *Nat. Rev. Dis. Prim.* **4**, 18001 (2018).
44. Siegel, C. H. & Sammaritano, L. R. Systemic lupus erythematosus: a review. *JAMA* **331**, 1480–1491 (2024).
45. Holgate, S. T. et al. Asthma. *Nat. Rev. Dis. Prim.* **1**, 15025 (2015).
46. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
47. Bentham, J. et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).
48. Han, Y. et al. Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. *Nat. Commun.* **11**, 1776 (2020).
49. Zhang, F. et al. Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. *Nature* **623**, 616–624 (2023).
50. Ebrahimiyan, H., Bagheri-Hosseinabadi, Z. & Abbasifard, M. Interferon regulatory factor 5 in rheumatoid arthritis and systemic lupus erythematosus. *Rheumatol. Res.* **6**, 1–13 (2021).
51. Duarte, J. H. IRF5 mediates joint inflammation. *Nat. Rev. Rheumatol.* **11**, 562–562 (2015).
52. Duffau, P. et al. Interferon regulatory factor 5 promotes inflammatory arthritis. *Arthritis Rheumatol.* **67**, 3146 (2015).
53. Ban, T. et al. Genetic and chemical inhibition of IRF5 suppresses pre-existing mouse lupus-like disease. *Nat. Commun.* **12**, 4379 (2021).
54. Song, S. et al. Inhibition of IRF5 hyperactivation protects from lupus onset and severity. *J. Clin. Investig.* **130**, 6700–6717 (2020).
55. Ni, J. et al. Novel insight into the aetiology of rheumatoid arthritis gained by a cross-tissue transcriptome-wide association study. *RMD Open* **8**, e002529 (2022).
56. Ntenti, C., Papakonstantinou, E., Fidani, L., Stolz, D. & Goulas, A. The genetics behind sulfation: impact on airway remodeling. *J. Pers. Med.* **14**, 248 (2024).
57. Ghannam, K. et al. Upregulation of immunoproteasome subunits in myositis indicates active inflammation with involvement of antigen presenting cells, CD8 T-cells and IFN γ . *PLoS ONE* **9**, e104048 (2014).
58. Raudvere, U. et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
59. Komatsu, N. et al. Plasma cells promote osteoclastogenesis and periarticular bone loss in autoimmune arthritis. *J. Clin. Investig.* **131**, e143060 (2021).
60. Zheng, C. et al. The role of extracellular vesicles in systemic lupus erythematosus. *Front. Cell Dev. Biol.* **10**, 835566 (2022).
61. Kim, M. et al. Exploring the clinical transition of engineered exosomes designed for intracellular delivery of therapeutic proteins. *Stem Cells Transl. Med.* **13**, 637–647 (2024).
62. Pyle, D. M., Yang, V. S., Gruchalla, R. S., Farrar, J. D. & Gill, M. A. IgE cross-linking critically impairs human monocyte function by blocking phagocytosis. *J. Allergy Clin. Immunol.* **131**, 491–500 (2013).
63. Simonin, E. M. & Wagner, B. IgE-binding monocytes upregulate the coagulation cascade in allergic horses. *Genes Immun.* **24**, 130–138 (2023).
64. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement.* **19**, 1598–1695 <https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13016> (2023).
65. Gatz, M. et al. Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* **63**, 168–174 (2006).
66. Bellenguez, C. et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
67. Cain, A. et al. Multicellular communities are perturbed in the aging human brain and Alzheimer's disease. *Nat. Neurosci.* **26**, 1267–1280 (2023).
68. Prater, K. E. et al. Human microglia show unique transcriptional changes in Alzheimer's disease. *Nat. Aging* **3**, 894–907 (2023).
69. Wang, Q. et al. Single cell transcriptomes and multiscale networks from persons with and without Alzheimer's disease. *Nat. Commun.* **15**, 5815 (2024).
70. Haney, M. S. et al. Apoe4/4 is linked to damaging lipid droplets in Alzheimer's disease microglia. *Nature* **628**, 154–161 (2024).
71. Habib, N. et al. Disease-associated astrocytes in Alzheimer's disease and aging. *Nat. Neurosci.* **23**, 701–706 (2020).
72. Sadick, J. S. et al. Astrocytes and oligodendrocytes undergo subtype-specific transcriptional changes in Alzheimer's disease. *Neuron* **110**, 1788–1805 (2022).
73. Pandey, S. et al. Disease-associated oligodendrocyte responses across neurodegenerative diseases. *Cell Rep.* **40**, 111189 (2022).
74. Sims, R., Hill, M. & Williams, J. The multiplex model of the genetics of Alzheimer's disease. *Nat. Neurosci.* **23**, 311–322 (2020).
75. Harold, D. et al. Genome-wide association study identifies variants at *CLU* and *PICALM* associated with Alzheimer's disease. *Nat. Genet.* **41**, 1088–1093 (2009).
76. Zhu, X.-C. et al. CR1 in Alzheimer's disease. *Mol. Neurobiol.* **51**, 753–765 (2015).
77. Hazrati, L.-N. et al. Genetic association of CR1 with Alzheimer's disease: a tentative disease mechanism. *Neurobiol. Aging* **33**, 2949–e5 (2012).
78. Kosoy, R. et al. Genetics of the human microglia regulome refines Alzheimer's disease risk loci. *Nat. Genet.* **54**, 1145–1154 (2022).
79. Kozlova, A. et al. *PICALM* Alzheimer's risk allele causes aberrant lipid droplets in microglia. *Nature* **646**, 1178–1186 (2025).
80. Le Guen, Y. et al. Multiancestry analysis of the HLA locus in Alzheimer's and Parkinson's diseases uncovers a shared adaptive immune response mediated by HLA-DRB1* 04 subtypes. *Proc. Natl. Acad. Sci. USA* **120**, e2302720120 (2023).
81. Bello, E. et al. An Alzheimer's disease-associated common regulatory variant in a *ptk2b* intron alters microglial function. Preprint at <https://www.biorxiv.org/content/10.1101/2023.11.04.565613v4> (2023).
82. Sudwarts, A. et al. BIN1 is a key regulator of proinflammatory and neurodegeneration-related activation in microglia. *Mol. Neurodegener.* **17**, 33 (2022).
83. Bryois, J. et al. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* **25**, 1104–1112 (2022).
84. Sun, Y. et al. A transcriptome-wide association study identifies novel blood-based gene biomarker candidates for Alzheimer's disease risk. *Hum. Mol. Genet.* **31**, 289–299 (2022).

85. You, S.-F. et al. MS4A4A modifies the risk of alzheimer disease by regulating lipid metabolism and immune response in a unique microglia state. Preprint at <https://www.medrxiv.org/content/10.1101/2023.02.06.23285545v1> (2023).
86. Deming, Y. et al. The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. *Sci. Transl. Med.* **11**, eaau2291 (2019).
87. Miyashita, A. et al. Sorl1 is genetically associated with late-onset Alzheimer's disease in Japanese, Koreans, and caucasians. *PLoS ONE* **8**, e58618 (2013).
88. Liu, N. et al. Hippocampal transcriptome-wide association study and neurobiological pathway analysis for Alzheimer's disease. *PLoS Genet.* **17**, e1009363 (2021).
89. Zhou, Y. et al. scpredixcan integrates deep learning methods and single-cell data into a cell-type-specific transcriptome-wide association study framework. *Cell Genomics* **5**, 100875 (2025).
90. Avsec, Ž et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).
91. Huang, C. et al. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat. Genet.* **55**, 2056–2059 (2023).
92. Sasse, A. et al. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.* **55**, 2060–2064 (2023).
93. Cuomo, A. S. et al. Impact of rare and common genetic variation on cell type-specific gene expression. Preprint at <https://www.medrxiv.org/content/10.1101/2025.03.20.25324352v1> (2025).
94. Emani, P. S. et al. Single-cell genomics and regulatory networks for 388 human brains. *Science* **384**, eadi5199 (2024).
95. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
96. Mancuso, N. et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675–682 (2019).
97. Zhao, S. et al. Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. *Nat. Genet.* **56**, 336–347 (2024).
98. Liu, L. et al. Conditional transcriptome-wide association study for fine-mapping candidate causal genes. *Nat. Genet.* **56**, 348–356 (2024).
99. Strober, B. J., Zhang, M. J., Amariuta, T., Rossen, J. & Price, A. L. Fine-mapping causal tissues and genes at disease-associated loci. *Nat. Genet.* **57**, 42–52 (2025).
100. Yang, Y., Lorincz-Comi, N. & Zhu, X. Uncovering causal gene-tissue pairs and variants through a multivariate TWAS controlling for infinitesimal effects. *Nat. Commun.* **16**, 6098 (2025).
101. Lu, Z. et al. Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide association studies. *Am. J. Hum. Genet.* **109**, 1388–1404 (2022).
102. Knutson, K. A. & Pan, W. Mats: a novel multi-ancestry transcriptome-wide association study to account for heterogeneity in the effects of cis-regulated gene expression on complex traits. *Hum. Mol. Genet.* **32**, 1237–1251 (2023).
103. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1 (2010).
104. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742–015 (2015).
105. Consortium, G. P. et al. A global reference for human genetic variation. *Nature* **526**, 68 (2015).
106. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
107. Pasaniuc, B. et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906–2914 (2014).
108. Squair, J. W. et al. Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
109. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.* **11**, 1561 (2017).
110. Mancuso, N. et al. Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat. Commun.* **9**, 4079 (2018).
111. Lin, Z. & Su, C. *Figshare Dataset: Cell-Type-Specific GReX Models Trained on OneK1K by scTWAS* https://figshare.com/articles/dataset/Cell-type-specific_GReX_models_trained_on_OneK1K_by_scTWAS/31329184 (2026).
112. Lin, Z. & Su, C. *Figshare Dataset: Cell-Type-Specific GReX Models Trained on ROSMAP by scTWAS* https://figshare.com/articles/dataset/Cell-type-specific_GReX_models_trained_on_ROSMAP_by_scTWAS/31329139 (2026).
113. Lin, Z. *Zhaotongl/sctwas: First Release* <https://doi.org/10.5281/zenodo.18434920> (2026).
114. Lin, Z. & Su, C. *Zhaotongl/sctwas_paper: First Release* <https://doi.org/10.5281/zenodo.18434979> (2026).

Acknowledgements

Z.L. was supported in part by the CTSA Provost Award at Florida State University. C.S. was supported in part by the URC Research Award at Emory University. We thank Joseph Powell and Angli Xue for help with obtaining the OneK1K data. This work was supported in part by the Rollins School of Public Health High Performance Computing Cluster at Emory University. The results published here are in part based on data obtained from the AD Knowledge Portal.

Author contributions

C.S. and Z.L. designed the research and wrote the paper; Z.L. implemented the method, performed simulation studies, and analyzed the OneK1K data; C.S. analyzed the ROSMAP data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-70374-7>.

Correspondence and requests for materials should be addressed to Chang Su.

Peer review information *Nature Communications* thanks Jiebiao Wang, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026