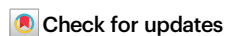


scMultiMap: Cell-type-specific mapping of enhancers and target genes from single-cell multimodal data

Received: 24 September 2024

Accepted: 15 April 2025

Published online: 26 April 2025

Chang Su^{1,2,4}✉, Dongsoo Lee¹, Peng Jin² & Jingfei Zhang^{1,3,4}✉

Mapping enhancers and target genes in disease-related cell types provides critical insights into the functional mechanisms of genome-wide association studies (GWAS) variants. Single-cell multimodal data, which measure gene expression and chromatin accessibility in the same cells, enable the cell-type-specific inference of enhancer-gene pairs. However, this task is challenged by high data sparsity, sequencing depth variation, and the computational burden of analyzing a large number of pairs. We introduce scMultiMap, a statistical method that infers enhancer-gene association from sparse multimodal counts using a joint latent-variable model. It adjusts for technical confounding, permits fast moment-based estimation and provides analytically derived *p*-values. In blood and brain data, scMultiMap shows appropriate type I error control, high statistical power, and computational efficiency (1% of existing methods). When applied to Alzheimer's disease (AD) data, scMultiMap gives the highest heritability enrichment in microglia and reveals insights into the regulatory mechanisms of AD GWAS variants.

The past two decades have seen significant advances in genome-wide association studies (GWAS), generating extensive catalogs of genetic variants linked to complex traits and diseases. However, over 90% of these identified variants are located in non-coding regions of the genome¹, and their disease-causing mechanisms remain largely unknown. Increasing evidence suggests that GWAS variants contribute to disease risk by modifying gene regulatory mechanisms in disease-relevant cell types^{1–3}. Mapping enhancers, a principal class of gene regulatory elements, and its target genes has shown great promise in uncovering the functions of GWAS variants in specific cellular contexts⁴. However, most existing analyses utilize data from bulk tissues, which may fail to capture the highly cell-type-specific nature of enhancers, or data from cell lines, which may not accurately represent the biology of primary cell types and cells from diseased subjects^{5,6}. Some other analyses detect cell-type-specific enhancers using epigenetic data from a single modality, but they usually cannot identify the associated target genes due to the lack of data modality with gene expression measurements⁵. Recent technologies such as 3D epigenetic

data and CRISPR screen data can be used to map enhancer-gene pairs in different cell types, but they remain laborious and costly to collect and may only be used to study cell lines^{5,7,8}.

The advent of single-cell multimodal technologies has unlocked unprecedented opportunities for mapping enhancers and target genes in specific cell types and contexts. Specifically, *paired* single cell assays for transposase-accessible chromatin using sequencing (scATAC-seq) and single cell RNA sequencing (scRNA-seq) allow for the profiling of both peak accessibility, a measurement of enhancer activity, and gene expression within the same cells. These data enable the identification of enhancer-gene pairs based on significant associations between peak accessibility and gene expression. As single-cell multimodal data are collected from both healthy individuals and those with various disease statuses across primary human cell types, it enables the computational inference of enhancer-gene pairs in a manner that is both cell-type-specific and context-dependent. Though this is also possible with unpaired scATAC-seq and scRNA-seq data⁹, paired data from multimodal technologies profile the coordinated variations of peaks and

¹Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA. ²Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, USA. ³Information Systems and Operations Management, Emory University, Atlanta, GA, USA. ⁴These authors jointly supervised this work: Chang Su, Jingfei Zhang. ✉e-mail: chang.su@emory.edu; emma.zhang@emory.edu

genes across individual cells, providing richer information and greater power for identifying peak-gene associations.

However, substantial computational and analytical challenges remain in the inference of peak-gene associations with single-cell multimodal data. From a computational perspective, the task is highly costly due to the computational burden of screening all candidate peak-gene pairs in the genome. For example, using a cis-region of 1Mb to define candidate pairs¹⁰ could result in $\sim 10^4$ pairs to be tested in a single cell type. Meanwhile, most existing methods use Monte Carlo methods for statistical inference, such as sampling peaks from different chromosomes to construct null distributions¹¹ and using bootstrapping for uncertainty quantification¹². For these methods, finding the p -value for a single peak-gene pair requires running the computational procedure $\sim 10^2$ to $\sim 10^4$ times. Combined with the large number of pairs and multiple cell types to consider, the overall computational costs of existing methods become prohibitively high, as the procedure needs to be run $\sim 10^6$ to $\sim 10^8$ times. From an analytical perspective, challenges arise due to the high sparsity, technical confounding and variations across biological samples in single-cell multimodal data. First, scATAC-seq data are highly sparse and often treated as binary¹³, leading to the common practice of using binarized counts in peak-gene association inference^{11,12}. Recent evidence, however, suggests that peak counts contain quantitative information and count-based modeling can improve downstream analysis^{13,14}. As a result, the heuristic treatment of peak counts in existing methods may result in a loss of information and detection power in peak-gene association inference. Second, the confounding effects of varying sequencing depths pose a challenge for peak-gene association inference. It has been shown that varying sequencing depths can lead to spurious associations when using most existing methods for inferring gene-gene associations from scRNA-seq data¹⁵. For peak-gene associations, similar spurious associations may arise as both peak and gene counts correlate with their respective sequencing depths, and the sequencing depths from these two data modalities tend to be correlated (Supplementary Fig. 1). Third, in the presence of multiple biological samples, inferring peak-gene associations are further challenged by coordinated variations in mean expression and accessibility across biological samples^{16–18}, leading to spurious associations. For example, consider a peak-gene pair that is not associated in sample A or sample B. If the mean expression and accessibility levels are both high in sample A and both low in sample B, then the peak-gene pair may appear associated if cells from samples A and B are pooled in the analysis without any careful adjustment.

To address these challenges, we present a new approach, called scMultiMap, that uses single-cell multi modal data to map cell-type-specific enhancer-gene pairs. scMultiMap is based on a multivariate latent-variable model that simultaneously models the gene counts and peak counts from single-cell multimodal data, and makes minimal parametric assumptions. It measures peak-gene association via the correlation between underlying gene expression and peak accessibility levels while accounting for variations in sequencing depths and across biological samples. Furthermore, we develop a highly computationally efficient moment-based estimation framework that provides both correlation estimates and theoretically justified analytical p -values for assessing statistical significance.

We evaluated scMultiMap by applying it to multiple paired scRNA-seq and scATAC-seq datasets, including datasets on peripheral blood mononuclear cells (PBMC) from healthy subjects and on postmortem brain samples from Alzheimer's disease (AD) patients and controls. Our results show that scMultiMap maintains appropriate type I error control and achieves higher statistical power when compared with existing methods. Additionally, results from scMultiMap are more reproducible across independent single-cell multimodal data and also more consistent with results from orthogonal data modalities on the same cell type, such as promoter capture Hi-C¹⁹, HiChIP²⁰ and

proximity ligation-assisted chromatin immunoprecipitation sequencing (PLAC-seq)³. We demonstrated the superior computational scalability of scMultiMap by benchmarking its computing time on real data. To illustrate its utility in studying functions of GWAS variants in disease-related cell types, we applied scMultiMap to data collected on microglia from AD patients and controls. This analysis revealed high enrichment for AD heritability in microglia enhancers and identified enhancer-gene pairs containing selective AD GWAS variants²¹, providing insights into the regulatory functions and disease-causing mechanisms of these AD variants in microglia.

There is another body of work that identifies peak-gene pairs based on variable importance scores in prediction models, such as SCENIC²². However, these methods do not incorporate statistical inference and lack p -values or false positive control. Hence, we evaluate their performance only on benchmark datasets and focus in this paper on methods that formally test peak-gene associations (Supplementary Discussion).

Results

Overview of scMultiMap

We propose a joint latent-variable model to simultaneously model gene expression and peak accessibility. Suppose there are p genes, q peaks, and n cells in the cell type of interest. Let x_{ij} and y_{ij} be the observed counts of gene j and peak j' in cell i , respectively. Furthermore, let z_{ij} be the underlying expression level for gene j , defined to be the number of mRNA molecules from each gene relative to the total number of mRNA molecules in a cell. Let $v_{ij'}$ be the underlying accessibility level for peak j' , defined to be the number of DNA fragments from each peak relative to the total number of DNA fragments in a cell. Use $s_i = \sum_{j=1}^p x_{ij}$ and $r_i = \sum_{j'=1}^q y_{ij'}$ to denote the sequencing depths for scRNA-seq and scATAC-seq in cell i , respectively. We propose the following model

$$(z_{i1}, \dots, z_{ip}, v_{i1}, \dots, v_{iq}) \sim F_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_2 \end{pmatrix},$$

$$x_{ij}|z_{ij} \sim \text{Poisson}(s_i z_{ij}), \quad y_{ij'}|v_{ij'} \sim \text{Poisson}(r_i v_{ij'}), \quad j \in [p], \quad j' \in [q], \quad (1)$$

where $F_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a non-negative $(p+q)$ -variate distribution with a mean vector $\boldsymbol{\mu}$ of length $p+q$ and a covariance matrix $\boldsymbol{\Sigma}$ of dimension $(p+q) \times (p+q)$. The covariance matrix $\boldsymbol{\Sigma}$ captures biological variations in the underlying gene expression and peak accessibility levels across cells, and is our main parameter of interest. Conditional on the latent expression level z_{ij} and accessibility level $v_{ij'}$, gene count x_{ij} and peak count $y_{ij'}$ are assumed to independently follow Poisson measurement models that depend on sequencing depths s_i and r_i , respectively. Our approach does not impose specific parametric assumptions on $F_{p+q}(\cdot)$, though (1) accommodates commonly considered distributions as special cases. For example, if z_{ij} ($v_{ij'}$) follows a Gamma distribution, then x_{ij} ($y_{ij'}$) follows a negative binomial distribution. In model (1), dispersion beyond the Poisson distribution is flexibly accommodated via $F_{p+q}(\cdot)$.

Under model (1), we measure peak-gene correlations via $\boldsymbol{\Sigma}_{12}$, which directly quantifies the correlation strength between the underlying expression $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})$ and accessibility $\mathbf{v}_i = (v_{i1}, \dots, v_{iq})$ and is not affected by variations in sequencing depths. When multiple biological samples (subjects) are present in the data set, we model the mean vector $\boldsymbol{\mu}$ as $\boldsymbol{\mu}_k$ for subject k . This consideration accounts for variations in means across biological samples and avoids spurious associations (see Methods section).

We model peaks using a count measurement model that accommodates sequencing depth variations and overdispersion. Existing methods either treat peak data as binary¹² or free from sequencing depth variations¹¹, leading to a loss of power and potentially confounded estimates. Supported by recent evidence that Poisson

modeling improves the downstream analysis of peak counts^{13,14} and the observation that peak counts display additional overdispersion compared to Poisson (Supplementary Fig. 2), model (1) is able to better leverage the quantitative information in peak counts, remove potential confounding due to sequencing depths, and improve power in detecting association.

Estimation and testing under model (1) are non-trivial, as we do not wish to make restrictive parametric assumptions on $F_{p+q}(\cdot)$. Moreover, there is usually a large number of peak-gene pairs to consider, making computational cost a practical and significant concern. To tackle these challenges, we propose a moment-based estimation framework that uses iteratively reweighted least squares (IRLS) with carefully specified weights to improve statistical efficiency. For genes $j \in [p]$, we estimate the mean parameters $\mu_{1,j}$'s and variance parameters $\sigma_{1,j}$'s using moment-based regressions for the first and second moments, iteratively until convergence. In this process, $\sigma_{1,j}$'s are estimated with $\hat{\mu}_{1,j}$'s and $\mu_{1,j}$'s are estimated with weights involving $\hat{\sigma}_{1,j}$'s. When multiple subjects are present in the data, we introduce a set of binary indicators for the subject-of-origin of each cell, allowing us to estimate subject-specific means and avoid confounding effects. A similar set of regressions is applied to estimate the mean $\mu_{2,j}$'s and variance $\sigma_{2,j}$'s of peaks $j' \in [q]$. Given the estimated mean and variance parameters for peaks and genes, we then estimate covariance with a moment-based regression that includes weighting for statistical efficiency. Computationally, we implement these regressions using matrix algebra to efficiently process multiple genes and peaks, eliminating explicit loops and enabling high parallelism and scalability. Under the proposed framework, we formulate a statistical test of hypothesis to evaluate the dependence between the underlying gene expression and peak accessibility levels from a given peak-gene pair, and analytically derive the null distribution. Correspondingly, the proposed test is theoretically justified and p -values can be analytically evaluated, without the need for time-consuming sampling-based inference. More details of the estimation and testing procedures can be found in the Methods section.

In summary, scMultiMap takes observed gene and peak counts as well as sequencing depths as input and generates analytical p -values for assessing the statistical significance of peak-gene associations. It properly models the distributions of gene and peak counts with a joint latent-variable model, accounting for variations in sequencing depths and across biological samples. The provided test has a controlled type I error rate, and enjoys a better statistical power. The computation of

correlation estimates and p -values is fast and can be quickly implemented for tens of thousands of peak-gene pairs.

scMultiMap has better detection accuracy and computational efficiency

To evaluate the performance of scMultiMap, we benchmarked its association detection accuracy and computational efficiency against existing methods using multiple single-cell multimodal datasets on PBMC from 10x Genomics^{23–26}. In the benchmark analysis, we considered two peak-gene association inference methods that provide p -values, including Signac^{10,11} and SCENT¹² (Methods). In particular, Signac estimates peak-gene associations using Pearson's correlations of gene counts normalized via *sctransform*²⁷ and raw peak counts. This method may also represent other tools that use Pearson's correlations to identify peak-gene pairs, such as ArchR²⁸ and scMEGA²⁹. For testing, it constructs a null distribution by randomly sampling peak counts from other chromosomes. This procedure does not account for sequencing depth variations in peak counts and has a computational cost of running $\sim 10^2$ procedures for a single pair. Additionally, correlations of marginally normalized data are known to be biased by mean and overdispersion^{15,30}, and random sampling is inadequate to adjust for such bias and may generate invalid p -values for inference. Specifically, correlations of marginally normalized data may be more attenuated for genes or peaks with either lower abundance levels or smaller overdispersion, as these are more affected by sequencing-induced measurement errors that attenuate associations. For valid statistical inference, the sampled genes and peaks used to estimate p -values should have the same mean and overdispersion parameters as the observed genes and peaks, which can hardly be guaranteed by random sampling. For SCENT, it employs a Poisson regression model that relates gene counts with binarized peak counts, and uses bootstrap methods for testing. This method may suffer from information loss due to the binarization of peak counts and confounding due to correlated sequencing depths. Additionally, it has a high computational cost ranging from $\sim 10^2$ to $\sim 10^4$ for a single pair¹².

To evaluate the type I error control of different methods, we construct null data by permuting peak counts randomly across cells. Correspondingly, the permuted peak counts are expected to be independent of gene counts (Methods). Fig. 1a shows that scMultiMap has an appropriate type I error control with empirical type I errors matching the nominal level of 0.05. In comparison, the empirical type I errors of SCENT are slightly conservative for most pairs and notably

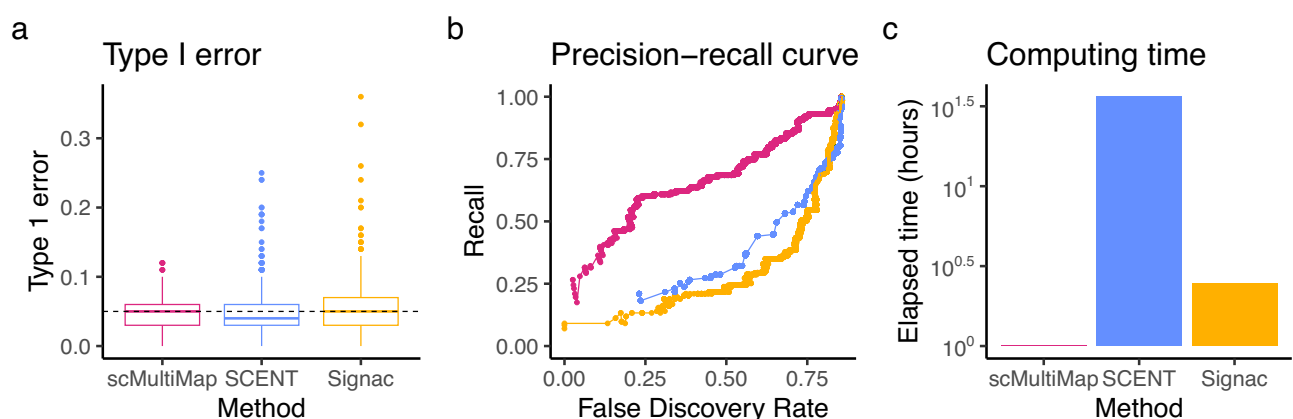


Fig. 1 | Performance evaluation of scMultiMap, SCENT and Signac using single-cell multimodal data on PBMC from 10x Genomics^{23–26}. **a** Empirical type I errors on null data with independent gene expression and peak accessibility levels for 1000 randomly selected peak-gene pairs. The dashed line marks the nominal level of 0.05. 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. **b** Precision-recall curves on simulated data, with the same color legend as in (a). **c** Computing time in hours (log scale) on a dataset with 729 cells and 31,132 candidate peak-gene pairs, evaluated on 1 core of an Intel(R) Xeon(R) Gold 6326 CPU @ 2.90GHz. Source data are provided as a Source Data file.

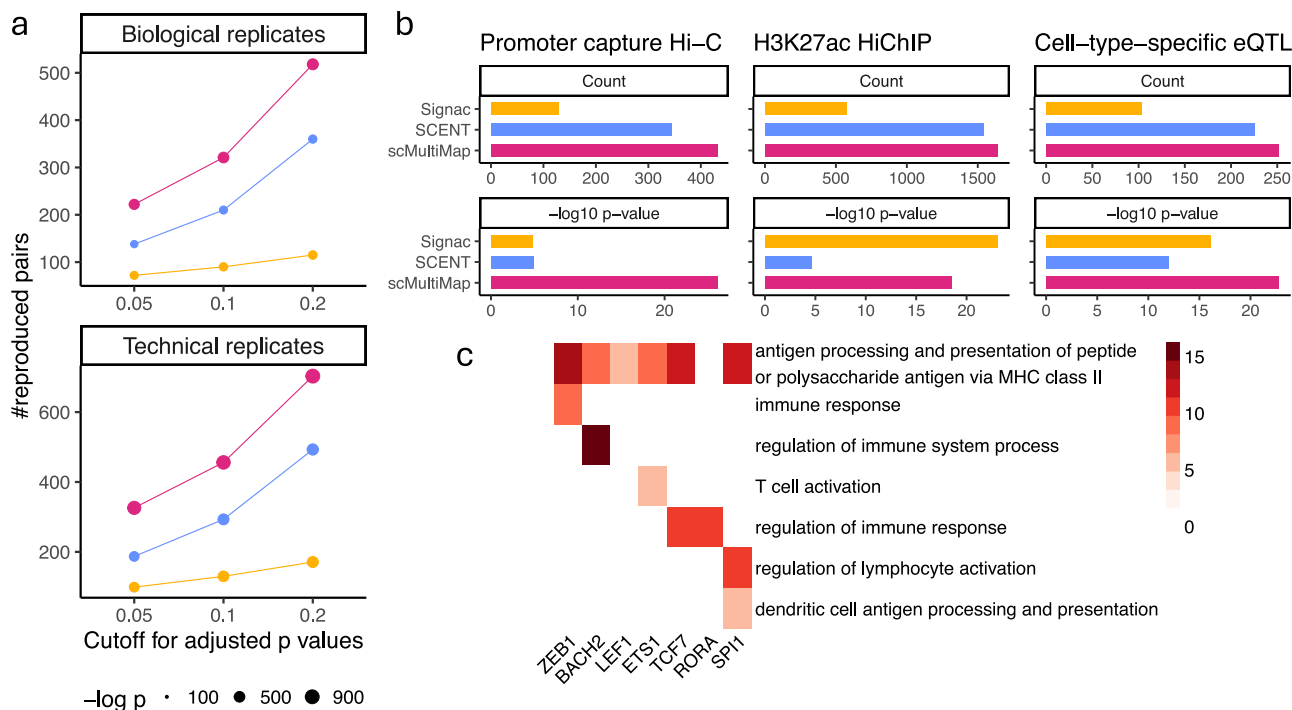


Fig. 2 | Comparison of reproducibility across methods and validation of regulatory trios inferred by scMultiMap. **a** Number of significant peak-gene pairs reproduced between biological replicates and between technical replicates of single-cell multimodal data on CD14 monocytes across varying BH-adjusted p -values cutoffs. See **b** for the color legend. **b** Consistency of scMultiMap findings with orthogonal data modalities. Significant peak-genes pairs on CD14 monocytes

(BH-adjusted p -value < 0.1) were compared against enhancer-gene pairs measured by promoter capture Hi-C¹⁹, H3K27ac HiChIP²⁰ and cell-type-specific eQTLs³¹ (Methods). In **a**, **b**, statistical significance of the overlapped counts is evaluated with one-sided Fisher exact tests ($-\log_{10} p$ values are shown). **c** Enrichment of GO biological processes among the target genes in trios identified for each TF in CD14 monocytes. Source data are provided as a Source Data file.

inflated for some pairs. The empirical type I errors of Signac are substantially inflated for a subset of pairs. Next, to evaluate the detection power of different methods, we simulate representative single-cell multimodal data using model parameters estimated from PBMC data (Methods). The precision-recall curves in Fig. 1b show that scMultiMap has higher power and a greater area under the curve. Both Signac and SCENT show reduced power, possibly because they are unable to fully extract the quantitative information in peak counts. When sparsity levels of simulated peak counts vary, the power of scMultiMap remains the highest, with greater improvement on less sparse data (Supplementary Fig. 3). The power of all methods increases when data are less sparse, with the exception of SCENT, which binarizes peak counts and benefits less from larger nonzero counts. Finally, to compare computational costs, we run all three methods on 729 CD14 monocytes and 31,132 peak-gene pairs using data from ref. 26. These pairs are derived from top 2000 highly expressed genes, top 20,000 highly accessible peaks, and a cis-region of width 1Mb¹⁰. Fig. 1c shows that the computational costs vary significantly across methods, with SCENT taking 1.47 days, Signac taking 1.46 hours, and scMultiMap taking only 8.35 seconds.

scMultiMap has higher reproducibility across independent datasets

The peak-gene association detection accuracy and power of scMultiMap are further evaluated through reproducibility analysis across multiple single-cell multimodal PBMC datasets. In particular, we evaluated reproducibility between biological replicates (different samples sequenced with the same instrument) and technical replicates (same samples sequenced with different instruments)^{23–25} (Methods). Fig. 2a shows that scMultiMap consistently yielded higher numbers of reproduced pairs in both reproducibility analyses across different BH-adjusted p -value cutoffs. Supplementary Fig. 4 shows that scMultiMap

also outperformed approaches in SCENIC²² with higher enrichment of reproduced pairs.

Furthermore, we compared the enhancer-gene pairs inferred from multimodal datasets with those inferred from orthogonal data types on the same cell type, including promoter capture Hi-C data, HiChIP data and cell-type-specific eQTL data (Methods). Fig. 2b shows that peak-gene pairs identified by scMultiMap are the most consistent with promoter capture Hi-C¹⁹, H3K27ac HiChIP²⁰, and cell-type-specific eQTL³¹ data on CD14 monocytes. These pairs show the largest numbers of overlaps, and the overlaps are statistically significant (Methods, see Supplementary Fig. 5 for percentages of overlaps). The enrichment by scMultiMap is also generally higher than approaches in SCENIC²² (Supplementary Fig. 6).

Finally, we illustrate the utility of scMultiMap in studying gene regulation in specific cell types. For this task, we constructed gene regulatory trios³², consisting of transcription factors (TFs), peaks and target genes, where a TF binds to the motif within the peak to regulate a target gene in close proximity to the peak. We quantified the association between TF and peak, peak and target gene with scMultiMap, and the association between TF and target gene with CS-CORE¹⁵. We considered the trios with simultaneous significant associations on TF-peak, peak-gene and TF-gene as cell-type-specific gene regulatory trios (Methods). Fig. 2c shows the Gene Ontology (GO) enrichment in biological processes among the target genes in identified trios for the seven top enriched TFs in CD14 monocytes. Consistent with the role of monocytes as antigen-presenting cells³³, the pathway for antigen processing and presentation via MHC class II was found to be strongly enriched, regulated by multiple TFs. Many other immune functions of CD14 monocytes were also identified, such as T cell and lymphocyte activation^{34,35}. Similar analysis of Signac and SCENT's results shows weaker enrichment among the identified trios (Supplementary Fig. 7).

scMultiMap identified biologically relevant gene regulatory mechanisms in brain cells

We further demonstrate the utility of scMultiMap with an analysis of the single-cell multimodal data collected from postmortem brain samples of Alzheimer's disease patients and controls³⁶. We focus on the cells from healthy controls. This part of the data include cells from eight subjects (biological samples), introducing potential variations across subjects and representing a more challenging scenario when compared to the PBMC data, which involves only one subject.

Using a bootstrap-based analysis (Supplementary Methods), we found significant variations in mean gene expression and mean peak accessibility across subjects (Supplementary Fig. 8). Interestingly, these variations tend to be correlated between peaks and genes (Supplementary Fig. 9), leading to spurious associations if left unaddressed. To understand the impact of coordinated variations across subjects on existing methods, we permuted real data to generate a null setting with independent peak accessibility and gene expression while preserving across-subject variations, and then evaluated the empirical type I errors (Methods). Figure 3a shows that Signac now exhibits much higher inflation in type I errors than in the PBMC data, as its correlation metric fails to account for variations across samples. SCENT also shows inflated type I errors, despite including subject id as a covariate in its regression model. In contrast, the empirical type I errors of scMultiMap align with the nominal level. We also evaluated the power of scMultiMap on simulated data with variations across subjects (Methods). Figure 3b shows that scMultiMap maintains the highest power. Notably, there is a larger gap between SCENT and Signac compared to the PBMC data (Fig. e 1b), as Signac is more confounded by the coordinated variations between peaks and genes across subjects.

We then evaluated the accuracy of scMultiMap on brain data through consistency and reproducibility analysis with other datasets measuring enhancers and target genes in brain cell types. Due to the large inflation (Fig. 3a) and systematic bias in Signac correlations by mean levels^{15,30}, we applied a permutation-based procedure to Signac to ensure a fair comparison with other methods (Methods). For consistency analysis, we used PLAC-seq data from³, which measures the interaction between promoter and distal regulatory regions in neuronal cells, oligodendrocytes and microglia using cell nuclei isolated from human brains. We compared the significant peak-gene pairs identified in single-cell multimodal data with PLAC-seq results (Methods). Fig. 3c shows that scMultiMap yields the highest enrichment for consistent enhancer-gene pairs in the three most abundant brain cell types, excitatory neurons, inhibitory neurons, and oligodendrocytes. The enrichment is below 1 and statistically insignificant for approaches in SCENIC+ applied to the same cell types (Supplementary Fig. 10), suggesting relatively poor overlap with PLAC-seq pairs. The numbers of consistent pairs in microglia are insignificant for all methods (raw p -value >0.05), possibly due to the smaller number of microglia in the single-cell multimodal dataset. The peaks from the significant peak-gene pairs from scMultiMap are also enriched for enhancer annotations on brain tissues from the same region and aged donors, as characterized by ChromHMM³⁷ in the ENCODE consortium³⁸ (Supplementary Fig. 11, Supplementary Discussion).

We further analyzed a second single-cell multimodal dataset from postmortem brain samples³⁹ and evaluated the reproducibility of findings with those from data in³⁶ (Methods). We focused on the three abundant brain cell types as in Fig. 3c and did not include astrocytes and microglia due to the limited number of cells from these two cell types in the second dataset³⁹. Fig. 3d, e show that scMultiMap generally has the highest enrichment and the largest number of overlapping pairs between the two independent datasets. While Signac has similar enrichment as scMultiMap, it has lower power and identified fewer pairs compared to scMultiMap. SCENT failed to generate reproducible discoveries when the significance cutoff is stringent, and the

enrichment of reproduced discoveries is much lower when the significance cutoff is lenient. This suggests that SCENT may suffer from lower power in identifying true pairs and inflated false positive discoveries that cannot be replicated across datasets. The enrichment of reproduced pairs by SCENIC+ is also lower than scMultiMap across three cell types (Supplementary Fig. 12).

We further applied scMultiMap to identify regulatory trios and studied the biological processes regulated by enriched TFs in five brain cell types (Fig. 3f, g, Supplementary Fig. 13) (Methods). The identified enriched GO biological processes are consistent with existing literature on the functions of these TFs in excitatory neurons (Fig. 3f). For example, FOXP2 is known for its regulatory role in neurite outgrowth⁴⁰ and neuronal differentiation⁴¹, MEF2C regulates synapse number and function to facilitate learning and memory⁴², and RORB contributes to the establishment of neocortical layers⁴³. Similarly, the trios yield biologically plausible enrichment in astrocytes (Fig. 3g). For instance, chemical synaptic transmission, a key function of astrocytes^{44,45}, is found to be enriched among the target genes of three TFs in Fig. 3g. Strong enrichment for multiple biological functions was identified for RORA, an important pluripotent transcription factor that supports the neuro-protective and anti-inflammatory role of astrocytes in the brain⁴⁶. The trios also yielded enrichment for cell-type-specific functions in oligodendrocytes (e.g. axon ensheathment, lipoprotein metabolic process), inhibitory neurons (e.g. chemical synaptic transmission, migration of Purkinje cell) and microglia (e.g. regulation of gliogenesis, inflammatory response) (Supplementary Fig. 13).

scMultiMap mapped GWAS variants of Alzheimer's disease to target genes in microglia

Previous AD GWAS have identified 75 loci associated with the risk of developing AD⁴⁷. However, the functional pathways and the cellular contexts in which these variants exert their effects remain unclear. Here, we use the same dataset as in the previous section³⁶, and leverage the candidate enhancer-gene pairs identified from scMultiMap in cells from healthy control and AD subjects to study cell-type and context-specific target genes of AD GWAS variants. We focus on microglia, the innate immune cell types in the brain whose cis-regulatory elements are the most enriched for AD genetic risk^{3,48}.

First, we demonstrate the power of scMultiMap in identifying candidate cis-regulatory elements in microglia via AD heritability enrichment analysis. Given the high microglia-specific enrichment reported in the literature^{3,48}, we hypothesize that a more powerful and accurate method for identifying cis-regulatory elements should yield higher heritability enrichment for AD in microglia. It is important to note that scMultiMap can capture general cis-regulatory elements that modulate the expression of target genes, even though it was initially motivated by detecting enhancer-gene pairs. Fig. 4a shows that using peaks from significant peak-gene pairs, scMultiMap generated higher and more significant AD heritability enrichment in a stratified linkage disequilibrium score (S-LDSC) analysis⁴⁹ based on three AD GWAS studies^{47,50,51} (Methods).

We then compared the inferred candidate peak-gene pairs in microglia from healthy controls and AD subjects. We identified significantly differentially associated peak-gene pairs (Fig. 4b) through a permutation analysis (Methods). Genes in differentially associated pairs (raw p -value <0.05) are significantly enriched for differentially expressed genes in microglia from⁵² (OR=1.58, p -value=2.55 $\cdot 10^{-6}$) and³⁶ (OR=1.51, p -value=0.01). We further intersected genes in significantly differentially associated peak-gene pairs and differentially expressed genes from³⁶ and evaluated the enriched GO biological processes. Fig. 4c shows the top enriched processes, three of which are closely related to lipid metabolism—a pathway previously implicated in AD pathogenesis and microglia^{50,53,54}. Similar analysis on astrocyte, another glial cell type critical to AD disease mechanisms^{55–57}, also revealed differentially associated peak-gene pairs and implicated pathways in lipid

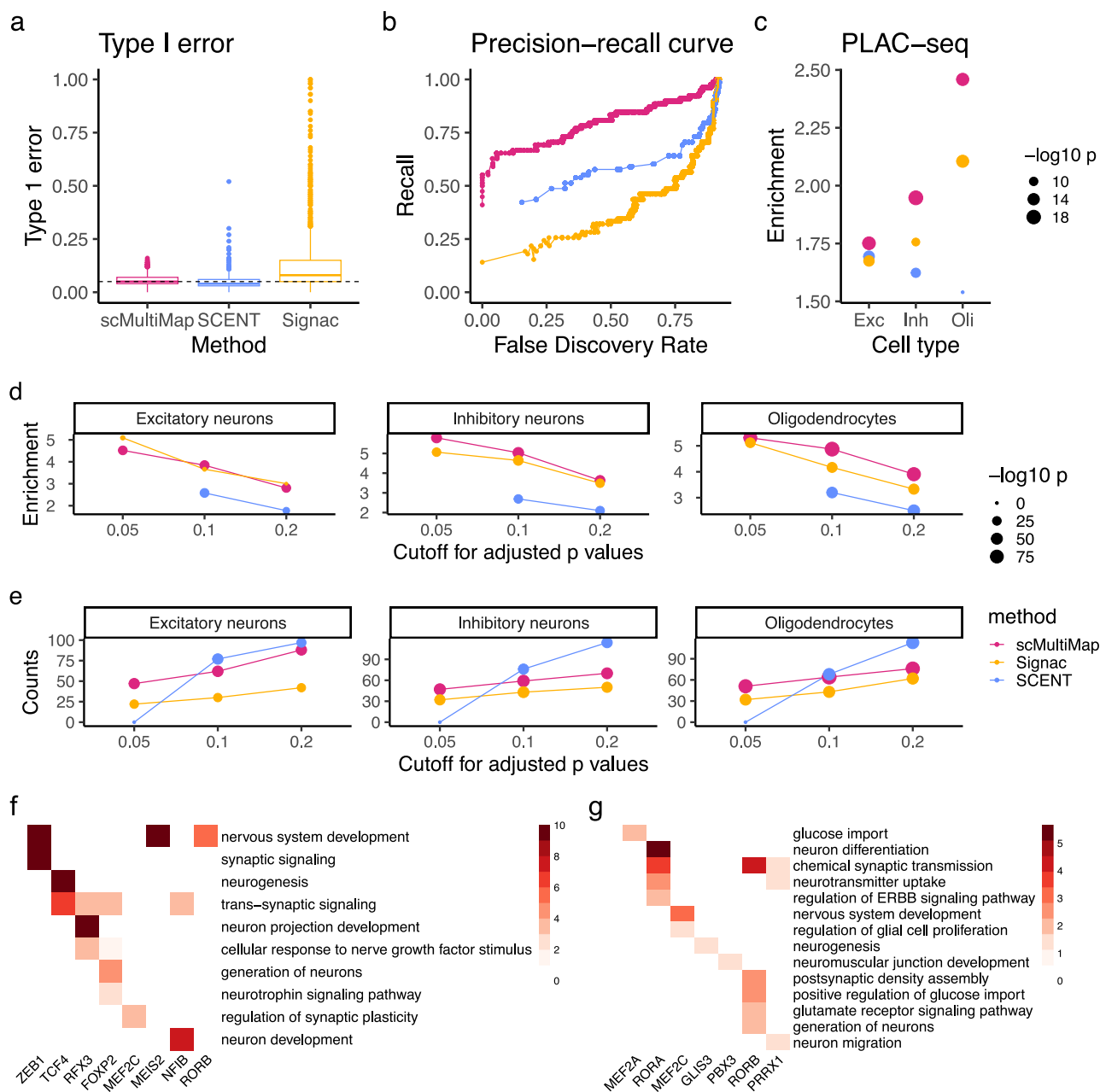


Fig. 3 | Results from scMultiMap on single-cell multimodal data from post-mortem brain samples in ref. 36. a Empirical type I errors on permuted data for 1000 randomly selected peak–gene pairs. Normalized peak values were randomly permuted within subject to break dependency with gene expression while preserving variations among subjects. The dashed line marks the nominal level of 0.05. 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. **b** Precision–recall curves on simulated data. See color legend in (a). **c** Consistency of significant pairs (BH-adjusted p -value < 0.2) with enhancer–gene pairs measured by PLAC-seq³ in excitatory neurons (Exc),

inhibitory neurons (Inh) and oligodendrocytes (Oli) (Methods). See color legend in (a). **d, e** Reproducibility of significant pairs with independent single-cell multimodal data on brain samples from³⁹ across cutoffs of BH-adjusted p -values, as evaluated by the enrichment (**d**) and the number (**e**) of reproduced counts. In **c–e** enrichment is quantified by odds ratio (OR) and log OR in (**c**) and (**d**) respectively (OR=0 not shown), and p -values are from one-sided Fisher exact tests. **f, g** Enrichment of GO biological processes among the target genes in trios identified for each TF in excitatory neurons (**f**) and astrocytes (**g**). Color intensity is given by BH-adjusted $-\log_{10} p$ -values from one-sided Fisher exact tests (values larger than 10 were set to 10). Source data are provided as a Source Data file.

metabolism and neurofibrillary tangle assembly, both of which have been linked to AD in astrocytes^{56,57} (Supplementary Fig. 14). We further used highly expressed genes in each cell type as the background and found that high-density lipoprotein particle assembly and positive regulation of lipid metabolic process remained significant for microglia and astrocytes, respectively, suggesting robustness to potential confounding effects from gene expression abundance.

We next integrated the scMultiMap findings with AD GWAS variants to study the functional roles of these variants in microglia. We

leveraged a set of AD GWAS variants that have been fine-mapped and prioritized based on microglia 3D epigenome annotations from a recent study²¹. scMultiMap found two significant peak–gene pairs that overlap with the selective GWAS variants and their putative target genes from²¹ (Fig. 4d, e). In particular, these pairs were validated by promoter capture Hi-C data on human pluripotent stem cell-derived microglia-like cells²¹, and they were not found by Signac or SCENT (raw p -value > 0.05). Figure 4d shows that scMultiMap identified a peak containing the AD GWAS variant rs10792831 as associated with *PICALM*

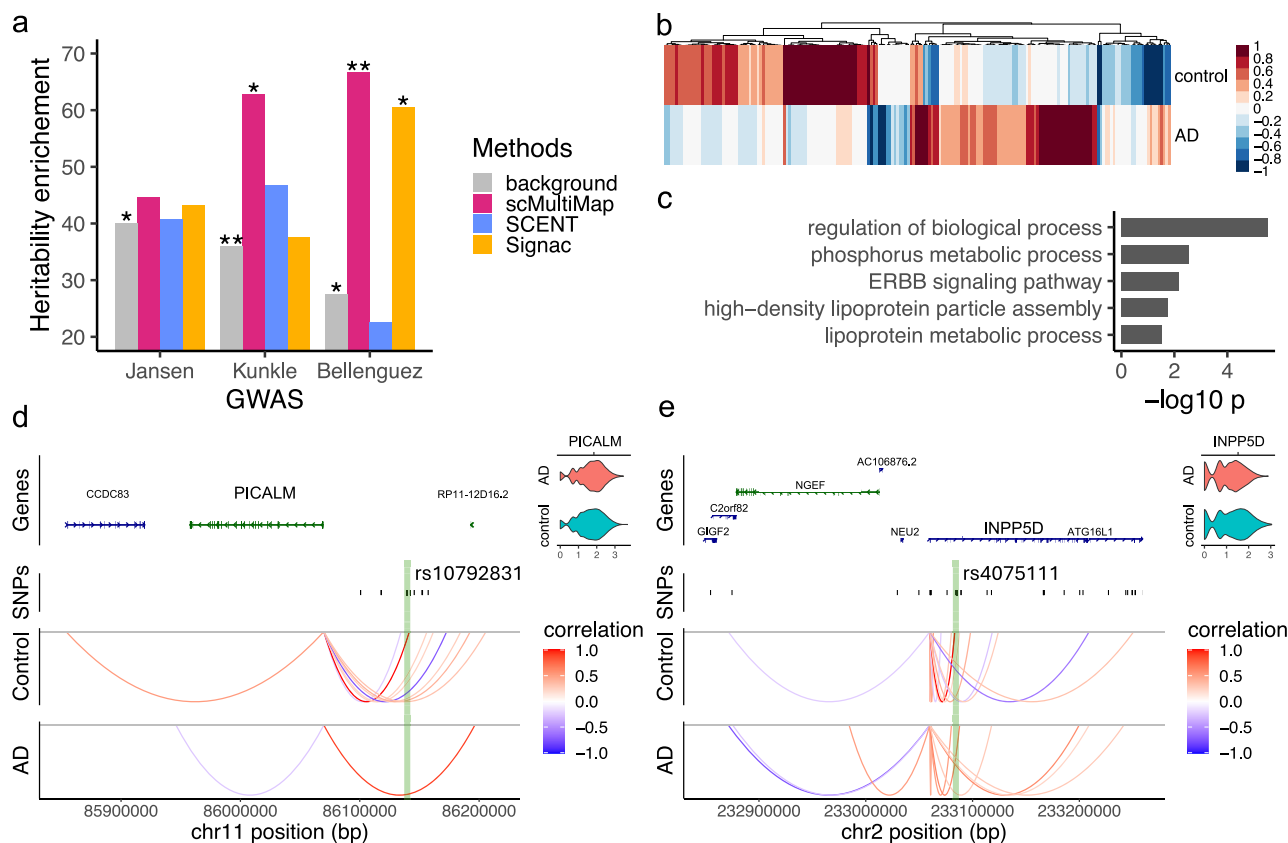


Fig. 4 | Studying the functional role of selective AD GWAS variants in microglia with scMultiMap. **a** AD heritability enrichment for peaks from significant peak-gene pairs (raw p -value < 0.05) in microglia. Summary statistics from three AD GWAS studies were used: Jansen⁵⁰, Kunkle⁵¹ and Bellenguez⁴⁷. * and ** denote p -value < 0.1 and 0.01, respectively for one-sided p -values of heritability enrichment from S-LDSC. **b** Differential peak-gene pairs in microglia from control and AD subjects. **c** Enrichment of GO biological processes among the genes from

significantly differential peak-gene pairs. BH-adjusted p -values from one-sided Fisher exact tests are shown. **d** scMultiMap mapped AD variant rs10792831 to *PICALM* in microglia from control subjects and the association is insignificant in microglia from AD subjects. **e** scMultiMap mapped AD variant rs4075111 to *INPP5D* in microglia from control subjects and the association is insignificant in microglia from AD subjects. In **(d, e)**, correlations with a magnitude greater than 0.2 are plotted. Source data are provided as a Source Data file.

in microglia. *PICALM* is a known AD locus⁵⁰ with upregulated expression in microglia from brain samples of AD patients³⁸. The variant rs10792831 is located 72 kb away from the transcription start site of *PICALM*. According to the FAVOR database⁵⁹, this variant has a high epigenetics active score (with elevated H3K27ac and H3K4me1 levels from ENCODE⁶⁰), a high TF score for overlap with TF binding sites, and a CADD score⁶¹ of 10.11 for deleteriousness. While these results suggest moderate regulatory potential without reference to cellular contexts, scMultiMap highlights its functional role in microglia and identifies the target gene regulated in this cell type. Interestingly, this pair (the peak with rs10792831 and *PICALM*) is also among the significantly differentially associated peak-gene pairs in microglia, where the association is no longer significant when microglia from AD samples are considered (Fig. 4d). This implies that the identified candidate enhancer is context-specific and that rs10792831 may contribute to AD risk by dysregulating the enhancer and, consequently, the expression of *PICALM* in microglia. Fig. 4e shows that scMultiMap identified another peak containing the AD GWAS variant rs4075111 as associated with *INPP5D* in microglia. *INPP5D* is a known AD locus with downregulated expression in microglia from AD samples³⁶. The associated variant, rs4075111, is located within the intronic region of *INPP5D*, 24 kb away from the transcription start site. It has been identified as being within the transcription factor binding site and enhancer for *INPP5D* by GeneHancer⁶², and it has a high epigenetics active score based on FAVOR⁵⁹. scMultiMap further highlights this enhancer-gene pair in the specific cellular context of microglia. Moreover, the association is not significant in microglia from AD patients, suggesting the context-

specificity of the enhancer and the potential regulatory role of the AD GWAS variant rs4075111 in microglia.

Discussion

We have introduced scMultiMap, a new statistical method for mapping cell-type-specific enhancer-gene pairs using single-cell multimodal data. By properly modeling peak counts and confounders in sequencing experiments, scMultiMap demonstrates higher statistical power to detect true enhancer-gene pairs and is robust against false positive associations due to varying sequencing depths and variations across biological samples. Utilizing a moment-based estimation framework and theoretically derived statistical tests, scMultiMap provides analytical p -values and has computational complexity that is less than 1% of existing methods. Systematic simulations and real data analyses show that scMultiMap better identifies reproducible and externally validated enhancer-gene pairs, making it a valuable tool for studying gene regulation in cell types. Integrative analysis with AD GWAS variants illustrates that scMultiMap can offer functional insights into the regulatory roles of GWAS variants in disease-related cell types, generating hypotheses for downstream validation and identifying potential targets for therapeutic intervention.

While scMultiMap is designed to quantify pairwise associations between peaks and genes, it can also serve as a tool for constructing gene regulatory networks^{22,63,64}. The gene regulatory trios illustrated here can be adapted to infer networks of target genes regulated by TFs. In this work, we have focused on paired scRNA-seq and scATAC-seq as an example of single-cell multimodal data, partly due to its availability.

However, scMultiMap can also be applied to other multimodal sequencing data, such as those that jointly profile transcriptome and histone modification in single cells^{65,66}, to study gene regulation in cell types.

We note that for the purpose of identifying enhancers, the results in this work are limited by the use of chromatin accessibility, which is necessary though not sufficient to define enhancers³. Additional data modalities, especially histone modification data in the cell types of interests, are needed to further validate the prioritized enhancers. As aforementioned, scMultiMap may also be applied to infer the association between gene expression and histone modification across single cells^{65,66}, which will offer complementary insights into enhancer-gene pairs compared to those inferred from paired scRNA-seq and scATAC-seq data.

In scMultiMap, we model the conditional distribution of observed counts given the underlying abundance levels using a Poisson distribution. This model is supported by existing literature, which indicates that a Poisson distribution adequately captures the variation introduced by sequencing experiments in scRNA-seq data^{27,67,68}, and similar observations on scATAC-seq data^{13,14}. If there exists higher dispersion from the sequencing step, it can be useful to adapt scMultiMap to model $x_{ij}|z_{ij}$ and $y_{ij'}|z_{ij'}$ using negative binomial distributions. This requires re-deriving the moment conditions (2)–(3) and updating the estimating and testing procedures accordingly. This is an important direction that requires a thorough and separate investigation, and we leave it to future work.

To apply scMultiMap, we would like to note two caveats with the datasets being used. First, scMultiMap cannot be applied to unpaired single-cell multi-omics data, since it needs the simultaneous observations of both modalities across single cells for estimation and inference. Secondly, when one of the modalities is scATAC-seq, one should use the fragment counts instead of read counts (or insertion counts) of DNA fragments, which enables the count modeling of scATAC-seq data and has been shown to achieve better performance in downstream analysis^{13,14}.

We have demonstrated the improved power of scMultiMap for identifying peak-gene pairs compared to existing methods. However, the power of this task is still limited by the sparsity and the small number of cells in single-cell multimodal data. It has been established that larger numbers of cells are needed for detecting over-dispersion⁶⁹, which is the variance of underlying abundance in the distribution F in (1). Similarly, for association inference, larger numbers of cells are also necessary to accurately estimate and test the covariance between underlying abundances. Often, there are fewer than a few hundred cells from the cell type of interest, providing limited information and yielding low to moderate power for association inference. Fortunately, larger consortium efforts are ongoing, collecting single-cell multimodal data with more cells across more biological samples. This will greatly improve the power of mapping enhancer-gene pairs in a cell-type-specific and context-specific manner. Given its better power and computational scalability, we expect scMultiMap to be a useful and practically appealing tool for analyzing these new consortium data, studying gene regulation, and elucidating the regulatory functions of GWAS variants in cell types. These results may further contribute to the broad collective efforts in the field, such as IGVF⁷⁰ and MorPhiC⁷¹, to map the functions of genetic variants and provide a more comprehensive understanding of their roles across cell types.

Methods

scMultiMap method

Based on model (1), the law of total expectation, and the law of total variance, we have:

$$\begin{aligned} E(x_{ij}) &= E(E(x_{ij}|z_{ij})) = E(s_i z_{ij}), \\ \text{Var}(x_{ij}) &= E(\text{Var}(x_{ij}|z_{ij})) + \text{Var}(E(x_{ij}|z_{ij})) = E(s_i^2 z_{ij}) + \text{Var}(s_i z_{ij}) \end{aligned}$$

Letting $\mu_{1,j}$ and $\sigma_{1,jj}$ denote the mean and variance of z_{ij} , respectively, it holds for the count of gene j that

$$E(x_{ij}) = s_i \mu_{1,j}, \quad \text{Var}(x_{ij}) = s_i \mu_{1,j} + s_i^2 \sigma_{1,jj}. \quad (2)$$

Similarly, it holds for the count of peak j' that

$$E(y_{ij'}) = r_i \mu_{2,j'} \quad \text{Var}(y_{ij'}) = r_i \mu_{2,j'} + r_i^2 \sigma_{2,j'j'}. \quad (3)$$

Furthermore, by model (1) and the law of total covariance, it holds that

$$\text{Cov}(x_{ij}, y_{ij'}) = E(\text{Cov}(x_{ij}, y_{ij'} | z_{ij}, v_{ij'})) + \text{Cov}(E(x_{ij} | z_{ij}), E(y_{ij'} | v_{ij'})) = s_i r_i \text{Cov}(z_{ij}, v_{ij'}).$$

Letting $\sigma_{12,jj'}$ denote the covariance between z_{ij} and $v_{ij'}$, it holds for counts from gene j and peak j' that

$$\text{Cov}(x_{ij}, y_{ij'}) = s_i r_i \sigma_{12,jj'}. \quad (4)$$

The estimation in scMultiMap includes two steps. The first step estimates the mean and variance parameters in (2)–(3) with IRLS. For example, for peak j' , we derive the following regression equations based on (3):

$$\begin{aligned} y_{ij'} &= r_i \mu_{2,j'} + \epsilon_{2,ij'}, \\ (y_{ij'} - r_i \mu_{2,j'})^2 &= r_i \mu_{2,j'} + r_i^2 \sigma_{2,j'j'} + \eta_{2,ij'}, \end{aligned}$$

where mean-zero error terms $\epsilon_{2,ij'}$'s are independent across i for peak j' , and the same for $\eta_{2,ij'}$'s. These moment-based regressions are consistent with (3) as taking expectation on both sides recovers (3). Due to the use of moment conditions, no restrictive assumptions need to be made on the distributions of $\epsilon_{2,ij'}$ and $\eta_{2,ij'}$, and the estimation procedure below holds for any peak counts from model (1). We propose to estimate $\mu_{2,j'}$ and $\sigma_{2,j'j'}$ with weighted least squares estimators $\hat{\mu}_{2,j'} = \arg \min_{\mu} \sum_{i=1}^n w_{2,ij'} (y_{ij'} - r_i \mu)^2$ and $\hat{\sigma}_{2,j'j'} = \arg \min_{\sigma} \sum_{i=1}^n h_{2,ij'} [(y_{ij'} - r_i \mu_{2,j'})^2 - r_i \mu_{2,j'} - r_i^2 \sigma^2]$, respectively. We set $w_{2,ij'} = 1/\text{Var}(\epsilon_{2,ij'})$ and $h_{2,ij'} = w_{2,ij'}^2$, which either equates or approximates the inverse variance of the response variable to improve the statistical efficiency of the estimators¹⁵. As $\hat{\mu}_{2,j'}$ depends on true $\mu_{2,j'}$ and $\sigma_{2,j'j'}$ through $w_{2,ij'}$ and $\hat{\sigma}_{2,j'j'}$ also depends on true $\mu_{2,j'}$, we propose an iterative procedure where we iterate between updating $\hat{\mu}_{2,j'}$ given $\hat{\sigma}_{2,j'j'}$ and vice versa. We propose to initiate the iteration with $w_{2,ij'} = 1$ (i.e. ordinary least squares estimator). The mean and variance for gene j can be estimated similarly from (2). A detailed algorithm of IRLS is included in Supplementary Algorithm 2.

Given these estimates, the second step of scMultiMap estimates the covariance between peaks and genes with weighted least squares (WLS). Based on (4), we drive the regression equation

$$(x_{ij} - s_i \mu_{1,j})(y_{ij'} - r_i \mu_{2,j'}) = s_i r_i \sigma_{12,jj'} + \xi_{ijj'},$$

where mean-zero error terms $\xi_{ijj'}$'s are independent across i for gene j and peak j' . These moment-based regressions are consistent with (4) as taking expectation on both sides recovers (4). We then propose to estimate $\sigma_{12,jj'}$ via $\arg \min_{\sigma} \sum_{i=1}^n g_{ijj'} [(x_{ij} - s_i \mu_{1,j})(y_{ij'} - r_i \mu_{2,j'}) - s_i r_i \sigma^2]$. We set the weights $g_{ijj'}$'s using $g_{ijj'} = 1/[\text{Var}(x_{ij})\text{Var}(y_{ij'})]$, which is the inverse variance of $\xi_{ijj'}$ under the null hypothesis of independence between gene and peak, and we use weights $w_{1,ij}$ and $w_{2,ij'}$ from step 1 to calculate $g_{ijj'}$. A detailed algorithm of WLS is included in Supplementary Algorithm 3.

For statistical inference on the association between peak and gene, we propose a test statistic based on the WLS estimator and analytically characterize its distribution under the null hypothesis. We define $T_{jj'} = \hat{\sigma}_{12,jj'} / \sqrt{\text{Var}(\hat{\sigma}_{12,jj'})}$, where $\hat{\sigma}_{12,jj'}$'s are estimated with true $\mu_{1,j}$, $\mu_{2,j'}$'s. Under the null hypothesis of independence between gene

expression and peak accessibility, we show that

$$T_{ij'} = \frac{\sum_i s_i r_i (x_{ij} - s_i \mu_{1,j}) (y_{ij'} - r_i \mu_{2,j'}) g_{ij'}}{\sqrt{\sum_i s_i^2 r_i^2 (s_i \mu_{1,j} + s_i^2 \sigma_{1,ij}) (r_i \mu_{2,j'} + r_i^2 \sigma_{2,j'}) g_{ij'}^2}}$$

and that $T_{ij'}$ asymptotically follows a standard normal distribution (Supplementary Methods). This result facilitates the analytical calculation of test statistics and p -values without the need for computationally intensive sampling, and also ensures the theoretical validity of the test. In practice, we compute this test statistics by plugging in the IRLS estimates of mean and variance parameters from the first step, which are all consistent estimators.

When multiple subjects or biological samples are present in the single-cell multimodal data, we extend the model in (1) to model the variations in mean expression and accessibility across subjects. Suppose there are K subjects. For cell $i = 1, \dots, n$ and subject $k = 1, \dots, K$, let $\mathbf{1}_i^k$ denote a binary indicator for subject k , i.e. $\mathbf{1}_i^k = 1$ if cell i is from subject k , otherwise 0. We assume that

$$E[(z_{i1}, \dots, z_{ip}, v_{i1}, \dots, v_{iq})] = \sum_{k=1}^K \mu_k \mathbf{1}_i^k, \quad (5)$$

where μ_k denotes the mean for all genes and peaks in subject k . This implies the following moment condition for peak j' : $E(y_{ij'}) = r_i (\sum_{k=1}^K \mu_{2,j'} \mathbf{1}_i^k)$, and similarly for gene j . We propose to estimate $\mu_{2,j'}$'s and $\mu_{1,j}$'s using a similar IRLS procedure and the estimation of variance and covariance follows analogously (Supplementary Methods). We note that when groups of subjects are present (e.g. control subjects and subjects with diseases), this will not remove the difference in covariance between groups and we model samples from each group separately (Supplementary Discussion).

Other methods for statistical inference of peak-gene associations

We compared scMultiMap with two other methods: Signac and SCENT. For Signac, we used the peak-gene association method originally developed by¹¹, and implemented in the software Signac¹⁰. In benchmarking computational costs (Fig. 1c), we used the *LinkPeaks* function in the R package Signac (v1.12.0) for the method Signac, and the R package SCENT (v1.0.0) for the method SCENT. SCENT was run without parallel computing to ensure a fair comparison with the other two methods.

In all other analyses, we used a re-implementation of *LinkPeaks* to evaluate Signac. This re-implementation corrects a coding error in the p -value calculation in its original implementation and speeds up computation (Supplementary Methods). We applied SCENT with fixed numbers of bootstrap samples in numerical analyses (Supplementary Methods).

We also evaluated the approaches in SCENIC²² for defining peak-gene pairs. However, since SCENIC+ is not based on statistical tests, lacks p -values and false positive control, it cannot be directly compared with methods that use statistical tests (see Supplementary Discussion). We implemented gradient boosting machine with the R package gbm (v2.2.2) to predict gene counts based on peaks. Following the methods in SCENIC²², we defined associated peak-gene pairs using three quantiles (85th, 90th, and 95th) of variable importance scores (see Supplementary Discussion).

Experiments for evaluating type I errors and power

To evaluate type I errors, we used a permutation-based procedure to generate null datasets where the gene expression and peak accessibility are independent. We used the gene counts as observed, and modified the observed peak counts following an approach that combines permutation and Poisson sampling as in¹⁵. This procedure

maximally preserves the characteristics of real data, such as the mean accessibility of peaks, dependency of peak counts on sequencing depths, and the correlation of sequencing depths between modalities, to faithfully evaluate type I errors in real data. When multiple subjects are present in the same dataset (Fig. 3), we applied the above procedure to cells from each subject separately, such that the variations in mean across subjects were preserved while any dependency between peaks and genes was removed. We obtained 100 independent replicates of null samples from this procedure and calculated the empirical p -values using the number of times the test statistic was rejected for all peak-gene pairs considered. To evaluate power, we simulated peak-gene pairs under model (1) with parameters estimated from real data, and incorporated model (5) for generating subject-specific means when multiple subjects are present in the dataset (Fig. 3). We then evaluated the precision-recall curve by varying the cutoff of p -values for calculating false discovery rate (FDR) and power. In all type I error and power evaluation, we used 1000 peak-gene pairs randomly sampled from the full set of candidate peak-gene pairs, with varying mean and variance parameters. More details can be found in Supplementary Methods.

Reproducibility analysis

We performed two sets of reproducibility analysis between independent single-cell multimodal data, including on PBMC in Fig. 2a and on brain samples in Fig. 3d, e. For PBMC analysis, we considered²³ and²⁴ as biological replicates, which were generated with the same 10x instrument but with different biological samples. We also included a pair of technical replicates^{24,25} that sequenced biological samples from the same individual but with different 10x instruments. We focused on CD14 monocytes as it is the most abundant cell types in these datasets. For brain analysis, we considered³⁶ and³⁹, which are two datasets independently generated by different labs, and evaluated the reproducibility on three most abundant brain cell types: excitatory neurons, inhibitory neurons, and oligodendrocytes. In both analyses, two peak-gene pairs are defined as overlapped if two genes are the same and two peaks overlap in genomic ranges.

We performed two sets of consistency analysis with orthogonal data types, including on PBMC in Fig. 2b and on brain samples in Fig. 3c. For PBMC analysis, we combined four datasets^{23–26} to maximize statistical power for detecting peak-gene pairs, and run scMultiMap and SCENT with indicators of datasets as covariates. In all analyses, we defined (candidate) cell-type-specific enhancer-gene pairs from orthogonal datasets following the original papers. In specific, for promoter capture Hi-C data¹⁹, we used interactions with a CHiCAGO score ≥ 5 ²²; for H3K27ac HiChIP data²⁰, we used the significant interactions based on FitHiChIP⁷³ (FDR < 0.1); for cell-type-specific eQTL, we used cis-SNPs significant in Naive CD14 Monocytes (FDR < 0.05)³¹; for PLAC-seq³, profiles chromatin loops between promoters and distal regulatory regions in NEUN⁺ neuronal, OLIG2⁺ oligodendrocyte, PU.1⁺ microglia nuclei isolated from human brain tissues, and we used the provided data on enhancers, promoters, and interactomes on these cell types to define enhancer-gene pairs (Supplementary Methods). For datasets whose genomic locations are in hg19, we use *liftOver* in R package rtracklayer (v1.62.0) to map the locations to hg38.

In both reproducibility and consistency analysis, when multiple subjects are present in the data, we further adjusted Signac by its estimates on permuted null data that preserve across-subject variations. This is because correlations in Signac are known to be biased by mean^{15,30}, and such systematic bias might cause artificially high reproducibility and consistency that are unfair to other methods with no systematic mean bias. Specifically, we computed Signac p -values using a background null distribution computed on permuted null data replicates, which corrects for the mean bias in Signac statistics.

Analysis of cell-type-specific gene regulatory trios

Following the conceptual model in ref. 32, we inferred cell-type-specific regulatory trios based on the associations between TF and target gene, TF and peak, peak and target gene using cells from the cell type of interest. We considered the highly expressed genes (mean expression ranks top 2000) and highly accessible peaks (mean accessibility ranks top 20,000). We identified enriched motifs in highly accessible peaks using motifs from JASPAR 2020 database⁷⁴ and *FindMotifs* from R package Signac (v1.12.0). For each TF whose motifs are enriched and all highly expressed genes in the cell type, we constructed candidate trios using this TF, a highly expressed target gene, and highly accessible peaks located within the cis-region of the target gene (width 1Mb) and harbor the binding motifs of the TF. For all candidate trios, we then inferred the associations between TF and peak and peak and target gene with scMultiMap, and associations between TF and target gene with CS-CORE¹⁵. Based on the *p*-values, we prioritized trios whose all three edges are significant as cell-type-specific regulatory trios and evaluated the enrichment for biological processes among the target genes from all trios for each TF. More details can be found in Supplementary Information.

GO enrichment analysis

We used *gost* in R package gprofiler2 (v0.2.2) to perform enrichment analysis for GO biological processes with one-sided Fisher exact tests and selected driver GO terms with FDR <0.05. We combined the selected terms manually based on the function of each cell type.

Differential peak-gene associations

In differential analysis, we considered top 5000 highly expressed genes and top 50,000 highly accessible peaks and the resulting candidate peak-gene pairs within the cis-region of width 1Mb in microglia and astrocytes. We applied scMultiMap while adjusting for the variations across subjects to cells from the control subjects and subjects with AD, respectively. To test the changes between two groups, we conducted a permutation analysis where we randomly permuted the disease label for 100 times and calculated *p*-values as the proportion of permutation replicates with the difference of covariance greater than that in observed data. In Fig. 4b, we focus on gene-peak pairs that are significantly associated in either control or AD cells (raw *p*-value < 0.05) with significant difference between groups (raw *p*-value < 0.05) and the difference in correlation is greater than a magnitude of 0.2. We further intersected this set of genes with cell-type-specific differentially expressed genes from ref. 36.

LDSC analysis for heritability enrichment

AD heritability enrichment analysis was conducted using stratified linkage disequilibrium score (S-LDSC) regression⁴⁹ (v1.0.1) to determine if peaks from significant peak-gene pairs (raw *p*-value < 0.05) obtained by scMultiMap have higher heritability enrichment for AD in microglia compared to Signac and SCENT. S-LDSC calculates heritability enrichment in a stratum using the ratio of the proportion of heritability explained versus the proportion of SNPs in the stratum, based on GWAS summary statistics and an ancestry-specific reference panel. We used three AD GWAS summary statistics^{47,50,51} and the European samples from 1000 Genomes data⁷⁵ as a reference panel. We estimated and tested the heritability enrichment of significant peaks from each method among all peaks considered in peak-gene association analysis, while adjusting for 97 functional annotations in baseline-LD v2.2, to best detect cell-type-specific heritability enrichment^{76,77}.

Preprocessing of fragment counts

We used fragment counts to quantify chromatin accessibility, which are more appropriate for count-based modeling and can yield improved performance in downstream analysis compared to read counts^{13,14}. For four single-cell multimodal datasets on PBMC^{23–26} and

the dataset on postmortem brain tissues³⁶, we called peak in each cell types, respectively with MACS2 (bioconda v2.2.9.1) to obtain fragment counts and to maximize the discovery of cell-type-specific peaks. For the validation brain data³⁹, we calculated fragment counts based on read counts following the procedure in¹⁴.

Reporting summary

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

Data availability

The PBMC data from 10x Genomics^{23–26} used in this study are available from the websites of 10x Genomics (see citations^{23–26} for the URLs). The brain datasets^{36,39} used in this study are available in the GEO database under accession code GSE214979 and available from URL https://personal.broadinstitute.org/bjames/AD_snATAC/MFC/. The ENCODE annotations of brain enhancers were based on subjects ENCDO980BZD and ENCDO871IWW are available on the ENCODE portal [<https://www.encodeproject.org>]. The detailed summary of single-cell multimodal data sets is in Supplementary Table 1. Source data are provided with this paper.

Code availability

The R package that implements scMultiMap⁷⁸ is publicly available and has been deposited in GitHub at <https://github.com/ChangSuBiostats/scMultiMap> under GPL license. The vignettes for using this R package are hosted at <https://changsubiostats.github.io/scMultiMap/>. The code used to perform the analyses and generate results in this study is publicly available and has been deposited in GitHub at https://github.com/ChangSuBiostats/scMultiMap_analysis under GPL license. The specific versions of the code associated with this publication are archived in Zenodo and are accessible via <https://doi.org/10.5281/zenodo.1494845778> and <https://doi.org/10.5281/zenodo.1496185079>.

References

- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory dna. *Science* **337**, 1190–1195 (2012).
- Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
- Nott, A. et al. Brain cell type-specific enhancer–promoter inter-actome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
- Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–295 (2013).
- Gasparini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
- Feingold, E. et al. The encode (encyclopedia of dna elements) project. *Science* **306**, 636–640 (2004).
- Xie, S., Armendariz, D., Zhou, P., Duan, J. & Hon, G. C. Global analysis of enhancer targets reveals convergent enhancer-driven regulatory modules. *Cell Rep.* **29**, 2570–2578 (2019).
- Gasparini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390 (2019).
- Trevino, A. E. et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053–5069 (2021).
- Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with signac. *Nat. methods* **18**, 1333–1341 (2021).
- Ma, S. et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell* **183**, 1103–1116 (2020).

12. Sakaue, S. et al. Tissue-specific enhancer–gene maps from multi-modal single-cell data identify causal disease alleles. *Nat. Genet.* **56**, 615–626 (2024).
13. Miao, Z. & Kim, J. Uniform quantification of single-nucleus atac-seq data with paired-insertion counting (pic) and a model-based insertion rate estimator. *Nat. Methods* **21**, 32–36 (2024).
14. Martens, L. D., Fischer, D. S., Yépez, V. A., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell atac-seq analysis. *Nat. Methods* **21**, 28–31 (2024).
15. Su, C. et al. Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nat. Commun.* **14**, 4846 (2023).
16. Tung, P.-Y. et al. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).
17. Squair, J. W. et al. Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
18. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
19. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).
20. Chandra, V. et al. Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. *Nat. Genet.* **53**, 110–119 (2021).
21. Yang, X. et al. Functional characterization of alzheimer’s disease genetic variants in microglia. *Nat. Genet.* **55**, 1735–1744 (2023).
22. Bravo González-Blas, C. et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. methods* **20**, 1355–1367 (2023).
23. Pbmcs from a healthy donor - granulocytes removed through cell sorting (10k). <https://www.10xgenomics.com/datasets/pbmcs-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-1-0-0>. Date Published: 2020-09-09.
24. 10k human pbmcs, multiome v1.0, chromium controller. <https://www.10xgenomics.com/datasets/10-k-human-pbmcs-multiome-v1-0-chromium-controller-1-standard-2-0-0>. Date Published: 2021-08-09.
25. 10k human pbmcs, multiome v1.0, chromium x. <https://www.10xgenomics.com/datasets/10-k-human-pbmcs-multiome-v1-0-chromium-x-1-standard-2-0-0>. Date Published: 2021-08-09.
26. Pbmcs from a healthy donor - granulocytes removed through cell sorting (3k). <https://www.10xgenomics.com/datasets/pbmcs-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-3-k-1-standard-1-0-0>. Date Published: 2020-09-09.
27. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
28. Granja, J. M. et al. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
29. Li, Z., Nagai, J. S., Kuppe, C., Kramann, R. & Costa, I. G. scmega: single-cell multi-omic enhancer-based gene regulatory network inference. *Bioinforma. Adv.* **3**, vbad003 (2023).
30. Wang, Y., Hicks, S. C. & Hansen, K. D. Addressing the mean–correlation relationship in co-expression analysis. *PLoS Comput. Biol.* **18**, e1009954 (2022).
31. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
32. Jiang, Y. et al. Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and cis-regulatory regions. *Cell Syst.* **13**, 737–751 (2022).
33. Jakubzick, C. V., Randolph, G. J. & Henson, P. M. Monocyte differentiation and antigen-presenting functions. *Nat. Rev. Immunol.* **17**, 349–362 (2017).
34. Banchereau, J. & Steinman, R. M. Dendritic cells and the control of immunity. *Nature* **392**, 245–252 (1998).
35. Ginhoux, F. & Jung, S. Monocytes and macrophages: developmental pathways and tissue homeostasis. *Nat. Rev. Immunol.* **14**, 392–404 (2014).
36. Anderson, A. G. et al. Single nucleus multiomics identifies zeb1 and mafk as candidate regulators of alzheimer’s disease-specific cis-regulatory elements. *Cell Genomics* **3**, (2023).
37. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. methods* **9**, 215–216 (2012).
38. Moore, J. E. et al. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
39. Xiong, X. et al. Epigenomic dissection of alzheimer’s disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422–4437 (2023).
40. Vernes, S. C. et al. Foxp2 regulates gene networks implicated in neurite outgrowth in the developing brain. *PLoS Genet.* **7**, e1002145 (2011).
41. Chiu, Y.-C. et al. Foxp2 regulates neuronal differentiation and neuronal subtype specification. *Developmental Neurobiol.* **74**, 723–738 (2014).
42. Barbosa, A. C. et al. Mef2c, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function. *Proc. Natl Acad. Sci. USA* **105**, 9391–9396 (2008).
43. Oishi, K., Aramaki, M. & Nakajima, K. Mutually repressive interaction between brn1/2 and rorb contributes to the establishment of neocortical layer 2/3 and layer 4. *Proc. Natl Acad. Sci. USA* **113**, 3371–3376 (2016).
44. Newman, E. A. New roles for astrocytes: regulation of synaptic transmission. *Trends Neurosci.* **26**, 536–542 (2003).
45. Haydon, P. G. & Carmignoto, G. Astrocyte control of synaptic transmission and neurovascular coupling. *Physiological Rev.* **86**, 1009–1031 (2006).
46. Journiac, N. et al. The nuclear receptor ror α exerts a bi-directional regulation of il-6 in resting and reactive astrocytes. *Proc. Natl Acad. Sci.* **106**, 21365–21370 (2009).
47. Bellenguez, C. et al. New insights into the genetic etiology of alzheimer’s disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
48. Gjonneska, E. et al. Conserved epigenomic signals in mice and humans reveal immune basis of alzheimer’s disease. *Nature* **518**, 365–369 (2015).
49. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
50. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer’s disease risk. *Nat. Genet.* **51**, 404–413 (2019).
51. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed alzheimer’s disease identifies new risk loci and implicates a β , tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
52. Mathys, H. et al. Single-cell multi-region dissection of alzheimer’s disease. *Nature* **632**, 858–868 (2024).
53. Chausse, B., Kakimoto, P. A. & Kann, O. Microglia and lipids: how metabolism controls brain innate immunity. In *Seminars in cell & developmental biology*, vol. 112, 137–144 (Elsevier, 2021).
54. Haney, M. S. et al. Apoe4/4 is linked to damaging lipid droplets in alzheimer’s disease microglia. *Nature* **628**, 154–161 (2024).
55. Bellaver, B. et al. Astrocyte reactivity influences amyloid- β effects on tau pathology in preclinical alzheimer’s disease. *Nat. Med.* **29**, 1775–1781 (2023).
56. Qi, G. et al. Apoe4 impairs neuron-astrocyte coupling of fatty acid metabolism. *Cell Rep.* **34**, 108572 (2021).

57. Serrano-Pozo, A. et al. Reactive glia not only associates with plaques but also parallels tangles in alzheimer's disease. *Am. J. Pathol.* **179**, 1373–1384 (2011).
58. Ando, K. et al. Clathrin adaptor calm/picalm is associated with neurofibrillary tangles and is cleaved in alzheimer's brains. *Acta neuropathologica* **125**, 861–878 (2013).
59. Zhou, H. et al. Favor: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res.* **51**, D1300–D1311 (2023).
60. Consortium, E. P. et al. An integrated encyclopedia of dna elements in the human genome. *Nature* **489**, 57 (2012).
61. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids Res.* **47**, D886–D894 (2019).
62. Fishilevich, S. et al. Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database* **2017**, bax028 (2017).
63. Wang, L. et al. Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nat. Methods* **20**, 1368–1378 (2023).
64. Yuan, Q. & Duren, Z. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nat. Biotech.* 1–11 (2024).
65. Zhu, C. et al. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. methods* **18**, 283–292 (2021).
66. Xie, Y. et al. Droplet-based single-cell joint profiling of histone modifications and transcriptomes. *Nat. Struct. Mol. Biol.* 1–6 (2023).
67. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).
68. Lause, J., Berens, P. & Kobak, D. Analytic pearson residuals for normalization of single-cell rna-seq umi data. *Genome Biol.* **22**, 1–20 (2021).
69. Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scrna-seq. *Genome Biol.* **23**, 27 (2022).
70. Consortium, I. et al. The impact of genomic variation on function (igvf) consortium. *ArXiv* (2023).
71. Molecular phenotypes of null alleles in cells (morphic). <https://www.genome.gov/research-funding/Funded-Programs-Projects/Molecular-Phenotypes-of-Null-Alleles-in-Cells>.
72. Cairns, J. et al. Chicago: robust detection of dna looping interactions in capture hi-c data. *Genome Biol.* **17**, 1–17 (2016).
73. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant chromatin contacts from hichip data by fithichip. *Nat. Commun.* **10**, 4221 (2019).
74. Fornes, O. et al. Jasp2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids Res.* **48**, D87–D92 (2020).
75. Consortium, G. P. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (2012).
76. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
77. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling s-ldsc and ldak functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).
78. Su, C. R package for scMultiMap: Cell-type-specific mapping of enhancers and target genes from single-cell multimodal data. GitHub <https://doi.org/10.5281/zenodo.14948456> (2025).
79. Su, C. & Lee, D. Analysis code for scMultiMap: Cell-type-specific mapping of enhancers and target genes from single-cell multimodal data. GitHub <https://doi.org/10.5281/zenodo.14961850> (2025).

Acknowledgements

We thank the High Performance Computing cluster at the Rollins School of Public Health, Emory University for providing computing support. Su was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award number UL1TR002378. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Jin was supported, in part, by the National Institutes of Health (NS111602 and HD104458). Zhang was supported by National Science Foundation grants DMS 2210469 and 2329296.

Author contributions

C.S. and J.Z. designed research; C.S. and D.L. performed research and analyzed data; C.S. contributed analytic tools; P.J. provided feedback on real data analysis; C.S., D.L., and J.Z. wrote the paper; C.S. and J.Z. jointly supervised the work.

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/s41467-025-59306-z>.

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

Peer review information *Nature Communications* thanks Wei Sun 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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025