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

Chang Su*, Dongsoo Lee, Peng Jin, Jingfei Zhang*,

* Co-corresponding authors: Chang Su, Jingfei Zhang

Email: chang.su@emory.edu, emma.zhang@emory.edu

Contents

Supplementary Methods

Details of scMultiMap method

Here, we derive the formula for $T_{jj'}$ in the main text based on its definition $T_{jj'} = \hat{\sigma}_{12,jj'}/\sqrt{\text{Var}(\hat{\sigma}_{12,jj'})}$. Let $\hat{\sigma}_{12,jj'}$ be the WLS estimate with true $\mu_{1j}, \mu_{2j'}$, we have

$$
\begin{aligned}
T_{jj'} &= \frac{\hat{\sigma}_{12,jj'}}{\sqrt{\text{Var}(\hat{\sigma}_{12,jj'})}} \\
&= \frac{\sum_i s_i r_i (x_{ij} - s_i\mu_{1,j})(y_{ij'} - r_i\mu_{2,j'}) g_{ijj'} / \sum_i s_i^2 r_i^2 g_{ijj'}}{\sqrt{\text{Var}(\hat{\sigma}_{12,jj'})}} && \text{WLS} \\
&= \frac{\sum_i s_i r_i (x_{ij} - s_i\mu_{1,j})(y_{ij'} - r_i\mu_{2,j'}) g_{ijj'} / \sum_i s_i^2 r_i^2 g_{ijj'}}{\sqrt{\sum_i s_i^2 r_i^2 \text{Var}[(x_{ij} - s_i\mu_{1,j})(y_{ij'} - r_i\mu_{2,j'})] g_{ijj'}^2 / (\sum_i s_i^2 r_i^2 g_{ijj'})^2}} && \text{Independence across } i \\
&= \frac{\sum_i s_i r_i (x_{ij} - s_i\mu_{1,j})(y_{ij'} - r_i\mu_{2,j'}) g_{ijj'}}{\sqrt{\sum_i s_i^2 r_i^2 \text{Var}[(x_{ij} - s_i\mu_{1,j})(y_{ij'} - r_i\mu_{2,j'})] g_{ijj'}^2}} \\
&= \frac{\sum_i s_i r_i (x_{ij} - s_i\mu_{1,j})(y_{ij'} - r_i\mu_{2,j'}) g_{ijj'}}{\sqrt{\sum_i s_i^2 r_i^2 \text{Var}(x_{ij} - s_i\mu_{1,j}) \text{Var}(y_{ij'} - r_i\mu_{2,j'}) g_{ijj'}^2}} && \text{under } \mathrm{H}_0 \\
&= \frac{\sum_i s_i r_i (x_{ij} - s_i\mu_{1,j})(y_{ij'} - r_i\mu_{2,j'}) g_{ijj'}}{\sqrt{\sum_i s_i^2 r_i^2 (s_i\mu_{1,j} + s_i^2\sigma_{1,jj})(r_i\mu_{2,j'} + r_i^2\sigma_{2,j'j'}) g_{ijj'}^2}} && \text{Equation (2) and (3)},
\end{aligned}
$$

which is exactly the formula shown in the main text. As $\hat{\sigma}_{12,jj'}$ is a M-estimator, $T_{jj'}$ converges in distribution to a standard normal distribution[1].

Next, We describe the algorithm of scMultiMap. Let $\odot$ denote the Hadamard product, and $\mathbf{1}_m$ denote a vector of 1's of length m. Algorithms 1-3 are for scMultiMap, and the IRLS and WLS step in scMultiMap, respectively. We write the algorithms in the general case where there could be K biological samples in the dataset, with $K = 1$ (i.e. no multiple biological sample present) as a special case. Steps are written in matrix algebra whenever possible to resemble our efficient software implementation.

Algorithm 1 scMultiMap

Input: Gene count matrix $X = (x_{ij})_{n\times p}$, peak count matrix $Y = (y_{ij})_{n\times q}$, binary matrix $E = (e_{ik})_{n\times K}$, where $e_{ik} = 1$ if cell i is from biological sample k.

$\mathbf{s} \leftarrow X\mathbf{1}_p, \mathbf{r} \leftarrow Y\mathbf{1}_q$.

// Step 1:

// Estimate mean and variance parameters for genes.

$\{\hat{\boldsymbol{\mu}}_{1,j}, \hat{\sigma}_{1,jj}\}_{j=1}^{p}, \{\mathbf{w}_{1,j}\}_{j=1}^{p} \leftarrow$scMultiMap_IRLS$(X, \mathbf{s}, E)$

// Estimate mean and variance parameters for peaks

$\{\hat{\boldsymbol{\mu}}_{2,j'}, \hat{\sigma}_{2,j'j'}\}_{j'=1}^{q}, \{\mathbf{w}_{2,j'}\}_{j'=1}^{q} \leftarrow$ scMultiMap_ IRLS$(Y, \mathbf{r}, E)$

// Step 2:

// Estimate peak-gene association and compute test statistics

$\{\hat{\sigma}_{12,jj'}, \hat{T}_{jj'}\} \leftarrow$ scMultiMap_WLS$(X, \mathbf{s}, \{\hat{\boldsymbol{\mu}}_{1,j}, \hat{\sigma}_{1,jj}\}_{j=1}^{p}, \{\mathbf{w}_{1,j}\}_{j=1}^{p}, Y, \mathbf{r}, \{\hat{\boldsymbol{\mu}}_{2,j'}, \hat{\sigma}_{2,j'j'}\}_{j'=1}^{q}, \{\mathbf{w}_{2,j'}\}_{j'=1}^{q})$

Output: $\hat{\sigma}_{12,jj'}, \hat{T}_{jj'}$ for $j = 1, \ldots, p, j' = 1, \ldots, q$

Algorithm 2 scMultiMap_IRLS

Input: Count matrix $X = (x_{ij})_{n\times p}$ for n cells and p features (genes or peaks), sequencing depths $\mathbf{s} = (s_1, \ldots, s_n)^\top$, binary matrix $E = (e_{ik})_{n\times K}$, where $e_{ik} = 1$ if cell i is from biological sample k.
Set $\Delta^{(0)} = 1$, $t = 0$, $D = (\mathbf{s}\mathbf{1}_K^\top) \odot E$. Let $\boldsymbol{x}_j$ be the j-th column of matrix X.
// Define weighted least suqares operators for estimating $\boldsymbol{\mu}_j$ and σ_{jj}, respectively.
$f_{\boldsymbol{\mu}_j}(\mathbf{w}) \coloneqq (D^\top \mathrm{diag}(\mathbf{w}) D)^{-1}(D^\top \mathrm{diag}(\mathbf{w})\boldsymbol{x}_j)$
$f_{\sigma_{jj}}(\boldsymbol{\mu}, \mathbf{h}) \coloneqq [(\mathbf{s}^2)^\top \mathrm{diag}(\mathbf{h})\mathbf{s}^2]^{-1} \left\{(\mathbf{s}^2)^\top \mathrm{diag}(\mathbf{h})[(\boldsymbol{x}_j - D\boldsymbol{\mu}\right)^2 - D\boldsymbol{\mu}]\}$
// Initialize with ordinary least squares
for $j = 1, \ldots, p$ **do**
  $\hat{\boldsymbol{\mu}}_j^{(t)} \leftarrow f_{\boldsymbol{\mu}_j}(\mathbf{1})$, $(\hat{\sigma}_{jj})^{(t)} \leftarrow f_{\sigma_{jj}}(\hat{\boldsymbol{\mu}}_j^{(t)}, \mathbf{1})$.
end for
// Iteratively reweighted least squares
while $\Delta^{(t)} \geq 0.05$ **do**
  $t \leftarrow t + 1$
  for $k = 1, \ldots, K$ **do**
    // Regularize θ_j estimates for weighting
    $\hat{\theta}_k^{(t)} \leftarrow \mathrm{median}_j\{(\hat{\sigma}_{jj})^{(t-1)}/(\hat{\mu}_{jk}^{(t-1)})^2\}$
  end for
  // Update $\boldsymbol{\mu}_j, \sigma_{jj}$ estimates
  for $j = 1, \ldots, p$ **do**
    $\mathbf{w}_j^{(t)} \leftarrow 1/[D\hat{\boldsymbol{\mu}}_j^{(t-1)} + (D\hat{\boldsymbol{\mu}}_j^{(t-1)})^2 \times \hat{\boldsymbol{\theta}}^{(t)}]$
    $\hat{\boldsymbol{\mu}}_j^{(t)} \leftarrow f_{\boldsymbol{\mu}_j}(\mathbf{w}_j^{(t)})$
    $\mathbf{h}_j^{(t)} \leftarrow 1/[D\hat{\boldsymbol{\mu}}_j^{(t)} + (D\hat{\boldsymbol{\mu}}_j^{(t)})^2 \times \hat{\boldsymbol{\theta}}^{(t)}]^2$
    $(\hat{\sigma}_{jj})^{(t)} \leftarrow f_{\sigma_{jj}}(\hat{\boldsymbol{\mu}}_j^{(t)}, \mathbf{h}_j^{(t)})$
  end for
  // Assess convergence
  $\Delta^{(t)} \leftarrow \max_j |\log \hat{\sigma}_{jj}^{(t)} - \log \hat{\sigma}_{jj}^{(t-1)}|$
end while
$\hat{\boldsymbol{\mu}}_j \leftarrow \hat{\boldsymbol{\mu}}_j^{(t)}, \hat{\sigma}_{jj} \leftarrow \hat{\sigma}_{jj}^{(t)}, \mathbf{w}_j \leftarrow \mathbf{w}_j^{(t)}$, for $j = 1, \ldots, p$
Output: $\{\hat{\boldsymbol{\mu}}_j, \hat{\sigma}_{jj}\}_{j=1}^p, \{\mathbf{w}_j\}_{j=1}^p$

Algorithm 3 scMultiMap_WLS

Input: $X, \mathbf{s}, \{\boldsymbol{\mu}_{1,j}, \sigma_{1,jj}\}_{j=1}^{p}, \{\mathbf{w}_{1,j}\}_{j=1}^{p}, Y, \mathbf{r}, \{\boldsymbol{\mu}_{2,j'}, \sigma_{2,j'j'}\}_{j'=1}^{q}, \{\mathbf{w}_{2,j'}\}_{j'=1}^{q}$

Set $D = (\mathbf{s}\mathbf{1}_K^\top) \odot E$. Let $\boldsymbol{x}_j$ be the j-th column of matrix X.

for $j = 1, \ldots, p,\ j' = 1, \ldots, q$ **do**

  // Compute regression weights

  $\mathbf{g}_{jj'} = \mathbf{w}_{1,j} \odot \mathbf{w}_{2,j'}$, let $G = \mathrm{diag}(\mathbf{g}_{jj'})$

  // WLS estimator for covariance

  $\hat{\sigma}_{12,jj'} = \left[(\mathbf{s} \odot \mathbf{r})^\top G(\mathbf{s} \odot \mathbf{r})\right]^{-1} \left\{(\mathbf{s} \odot \mathbf{r})^\top G\left[(\boldsymbol{x}_j - D\boldsymbol{\mu}_{1,j}) \odot (\boldsymbol{y}_{j'} - D\boldsymbol{\mu}_{2,j'})\right]\right\}$

  // Compute test statistics

  $$\hat{T}_{jj'} = \frac{(\mathbf{s} \odot \mathbf{r})^\top G\left[(\boldsymbol{x}_j - D\boldsymbol{\mu}_{1,j}) \odot (\boldsymbol{y}_{j'} - D\boldsymbol{\mu}_{2,j'})\right]}{\sqrt{(\mathbf{s} \odot \mathbf{r})^\top \mathrm{diag}\left[(D\boldsymbol{\mu}_{1,j} + \mathbf{s}^2\sigma_{1,jj}) \odot (D\boldsymbol{\mu}_{2,j'} + \mathbf{r}^2\sigma_{2,j'j'}) \odot \mathbf{g}_{jj'}^2\right](\mathbf{s} \odot \mathbf{r})}}$$

end for

Output: $\hat{\sigma}_{12,jj'}, \hat{T}_{jj'}$ for $j = 1, \ldots, p,\ j' = 1, \ldots, q$

Evidence for overdispersion and variations across biological samples

To demonstrate overdispersion in peak counts in addition to Poisson, we performed goodness of fit test against Poisson for peak counts following the analysis for gene counts in [2]. We first randomly sampled 2,000 highly accessible peaks with probability inverse proportional to the density of mean accessibility, such that the sampled peaks evenly cover the mean accessibility levels [3]. For each of the sampled peak, we fitted a Poisson GLM model with sequencing depths as an offset, calculated the randomized quantile residuals with *statmod::qresid* in R [4], and performed a goodness of fit test using the residuals. We then used p-values from the goodness of fit test as evidence that Poisson models are insufficient to describe peak counts, and variance of quantile residuals as descriptive statistics for quantifying overdispersion in peak counts. We performed this analysis for three single-cell multimodal datasets on PBMC from 10x Genomics: PBMC_10k [5], PBMC_10k_nextgem [6], PBMC_10k_nextgem_X [7] and four abundant cell types in each of the datasets. The results are shown in Supplementary Figure 2.

To demonstrate the variations across biological samples, we quantified the variance in underlying abundance explained by the set of sample indicators $\{\mathbf{1}\{\text{cell } i \in \text{sample } k\}\}_{k=1}^{K}$ for genes and peaks respectively using F-test statistics. We then tested if the variance explained is equal to 0 using bootstrap hypothesis testing. We calculated p-values with 1,000 bootstrap samples for the genes and peaks and 3 abundant brain cell types analyzed in Figure 3, and the results are shown in Supplementary Figure 8.

To demonstrate that the mean variations tend to be correlated between peaks and genes, we evaluated the p-values for the Pearson correlations of variations between modalities (i.e.

$\mathrm{Cor}(\hat{\boldsymbol{\mu}}_{1,j}, \hat{\boldsymbol{\mu}}_{2,j'})$ for $\hat{\boldsymbol{\mu}}_{1,j}$, $\hat{\boldsymbol{\mu}}_{2,j'}$ as defined in Algorithm 1) using a two-sided test implemented by *cor.test* in R. Supplementary Figure 9 shows the distribution of these p-values for the peak-gene pairs in 3 abundant brain cell types analyzed in Figure 3.

Implementations of Signac and SCENT

At the time of the study, the *LinkPeaks* function implemented by Signac mistakenly computes the two-sided p-value to be half of the correct value (see `https://github.com/stuart-lab/signac/blob/8ecdde2/R/links.R#L481C29-L481C30`). As a result, the p-value generated by Signac is bounded between 0 and 0.5, yielding artificially inflated type I errors in peak-gene association inference. We re-implemented *LinkPeaks* to correct for this coding error. We also separated the computation of background null distribution from the calculation of p-values, such that the null distribution can be fixed for all data replicates in permutation and simulation experiments (Figures 1a-b), speeding up the computation. For SCENT, we fixed the number of bootstrap samples to be 500 in permutation and simulation analysis to avoid extreme computational costs, and 5,000 in real data analysis to obtain refined p-values for multiple testing adjustment.

Details of evaluating type I errors and power

We used data from CD14 monocytes in PBMC data [6] and data from oligodendrocytes in brain data [8] to conduct the numerical experiments in Figures 1a-b and 3a-b, respectively. For each cell type, we first selected the top 2,000 highly expressed genes and top 20,000 highly accessible peaks, and used a cis-region of width 1Mb to select candidate peak-gene pairs. To save computational costs, we randomly sampled 1,000 pairs when evaluating the type I errors and precision-recall curves. In simulation experiments, we applied scMultiMap to estimate the parameters $\{\mu_{1,j}, \mu_{2,j'}\}$ and $\{\sigma_{1,jj}, \sigma_{2,j'j'}\}$ for the mean and variance of underlying gene expression z_{ij} and underlying peak accessibility $v_{ij'}$ respectively, and $\sigma_{12,jj'}$ for the covariance between gene j and peak j'. We further set $\sigma_{12,jj'} = 0$ for those pairs whose p-values are greater than 0.05. When multiple subjects are present in the data, we specified heterogeneous $\mu_{1,jk}$'s across subjects $k = 1, \ldots, K$, such that the standard deviation across K subjects is 20% of the average mean $1/K \sum_{k=1}^{k} \mu_{1,jk}$, and that the largest overdispersion $(\sigma^2_{1,jj}/\mu^2_{1,jk})$ is 1 across subjects. We repeated the same procedure for peaks. Given the parameters, we then used a Gamma-Poisson mixture model combined with a Gaussian copula [9] to simulate correlated counts for each peak-gene pair [10].

Details in the analysis of gene regulatory trios

To maximize statistical power for inferring gene regulatory trios in PBMC, we combined cells from four PBMC datasets from 10x Genomics [5, 6, 7, 11] and run scMultiMap with binary indicators for datasets as covariates. We used the following significance thresholds on raw p-values of scMultiMap and CS-CORE [10] for prioritizing regulatory trios: 0.05 for CD14 monocytes in PBMC data and 0.001, 0.001, 0.001, 0.05 and 0.1 for excitatory neurons, inhibitory neurons, oligodendrocytes, astrocytes, and microglia in brain data. We considered the top 5 enriched GO:BP driver terms (BH-adjusted p-value < 0.05) for the top 10 TFs with the greatest number of enriched trios in each cell type. Terms with relevant biological functions of each cell type are highlighted in Figure 2c,3f-g and Supplementary Figure 13.

Details of defining cell-type-specific enhancer-gene pairs from orthogonal datasets

We used the Supplementary Table 5 of [12] to extract enhancers, promoters and interactomes on neuronal cells, oligodendrocytes and microglia. We adapted the codes from [13] and defined a pair of interacting peaks that covers a cell-type-specific enhancer and a promoter to be a enhancer-gene pair in that cell type. We used the results on neuronal cells to validate the findings in excitatory and inhibitory neurons.

Supplementary Discussion

Discussion on the group effects

Here, we clarify the impact of incorporating subject-specific means on estimating peak-gene associations across groups (e.g. samples from AD and control groups). The subject-specific mean will not remove group effects in peak-gene associations (i.e., covariances). This is because mean and covariance are two distinct components of the data. While subject-specific means adjust for individual variation in the mean, they do not address differences in the covariance between groups.

When estimating the peak-gene association, we assume that all subjects come from the same population and share the same covariance. When there are multiple populations (e.g., AD and control groups), we estimate the covariance and peak-gene associations separately for each group, which does not eliminate the covariance differences between them. For example, in the analysis of microglia from AD and control subjects in Results, peak-gene associations were estimated separately for each group.

Evaluation of SCENIC+

SCENIC+ [14] is a tool for gene regulatory network inference, which also includes an intermediate step that identifies associated peak-gene pairs from single-cell multimodal data. While all the methods we have considered (scMultiMap, Signac, SCENT, scMEGA) rely on statistical tests and p-values to identify associated pairs, SCENIC+ uses variable importance scores from gradient boosting machines, combined with heuristic procedures, to binarize these scores and determine associated pairs. For example, when using all neighboring peaks to predict gene expression in a gradient boosting machine, SCENIC+ assigns peaks with importance scores above 85th, 90th, and 95th quantiles as associated with the corresponding gene. This approach does not consider the uncertainty in estimated associations (importance scores) and does not provide theoretical guarantee for controlling false positives.

Since SCENIC+ is a ranking based method without statistical significance evaluation, it will always identify the same number of associated pairs at a given cutoff, even when there is no actual relationship between the peaks and genes (i.e., when all peaks and genes are independent). This makes it challenging to conduct a meaningful and fair comparison with other methods based on statistical tests. For example, if an approach simply claims all pairs as associated, its power will be exactly 1, performing seemingly better than all other methods. However, its type I error will also be 1. Therefore, a meaningful comparison of the true discoveries across different approaches is only possible when their type I error rates are comparable. For these reasons, SCENIC+ was evaluated only on a few representative benchmark datasets, and any comparison in this manuscript involving SCENIC+ should be interpreted with these caveats in mind.

Despite the relatively poor performance of SCENIC+ in our evaluation (Supplementary Figures 4,6,10,12), we would like to clarify a few distinctions between the procedure evaluated here and the original proposed use of SCENIC+. First, the original SCENIC+ paper applies this procedure to cells across all cell types. This approach is more likely to identify pairs of cell-type marker genes and marker peaks, rather than peak-gene pairs that co-vary within a specific cell type. Hence, the original proposed use of SCENIC+ is distinct from the goal of this paper. Secondly, SCENIC+ imputes the chromatin accessibility data before evaluating the association between peaks and genes. While imputation may address some data limitations, it does not resolve the lack of false positive control inherent in this approach. Since imputation is not utilized by other methods considered in this manuscript and has been shown to induce false positives when inferring associations between genes [15], evaluating its impact requires a thorough and separate investigation, which we leave for future work.

Finally, SCENIC+ uses two additional sets of procedures to binarize the variable importance scores: top $5, 10, 15$ peaks per gene and a customized implementation of BASC [16]. However, the former approach will identify a large proportion of pairs as associated on our brain data (e.g. 50%-89% associated pairs on excitatory neurons from [8]), which is unrealistic, and the latter approach relies on an external tool to binarize the scores which also fails to consider the uncertainty in score estimates.

Validation with ENCODE enhancers

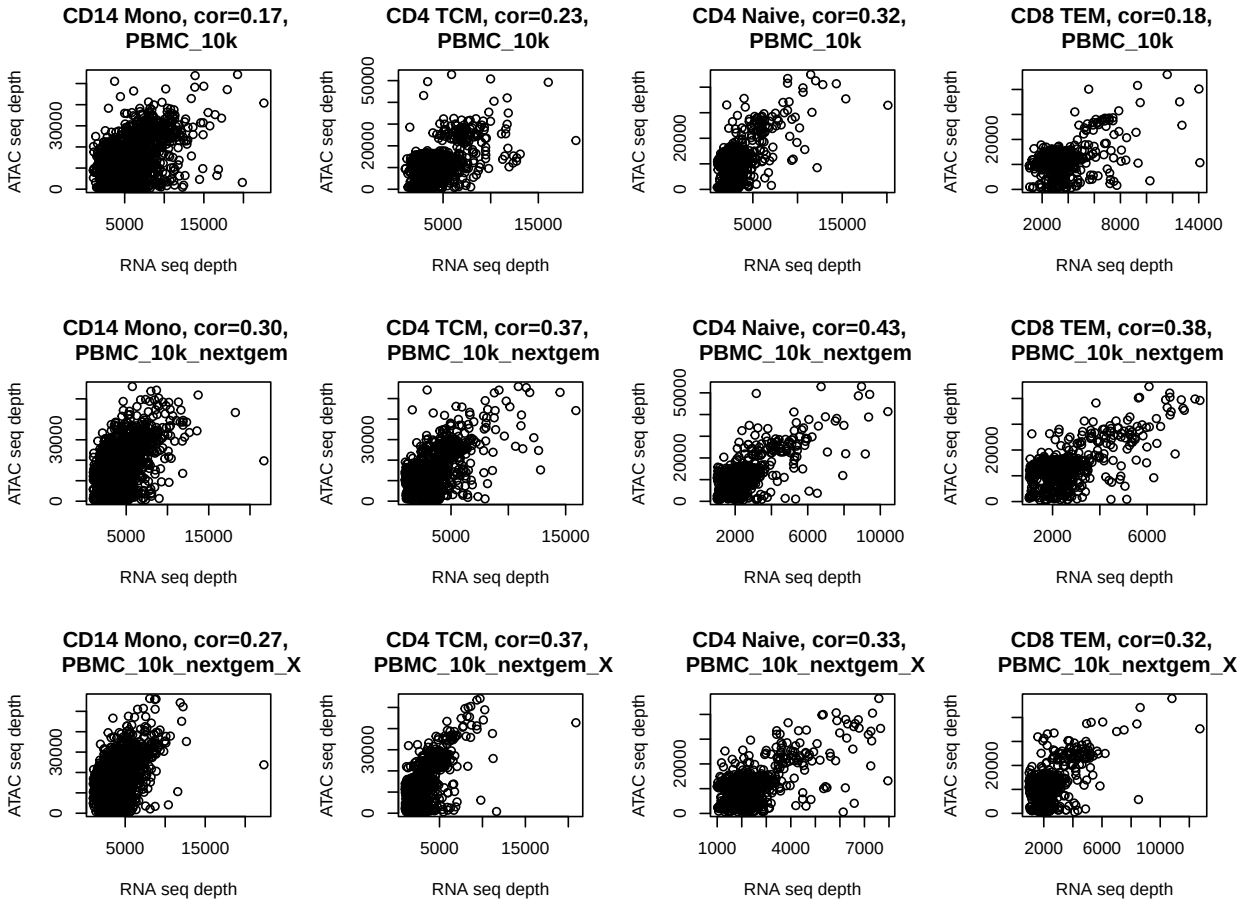
There are two caveats to interpreting the consistency analysis with ENCODE enhancers (Supplementary Figure 11). First, these findings only validate the peaks as putative enhancers, but do not validate the association between peaks and their target genes. In other words, these results could not inform the performance of mapping enhancers to target genes. Second, the ENCODE annotation data were generated from bulk brain tissues, which may lack cell-type-specificity in identified enhancers. Given these considerations, we suggest that (Supplementary Figure 11) may not be the most appropriate for comparing the performance of the three methods.

Supplementary Table

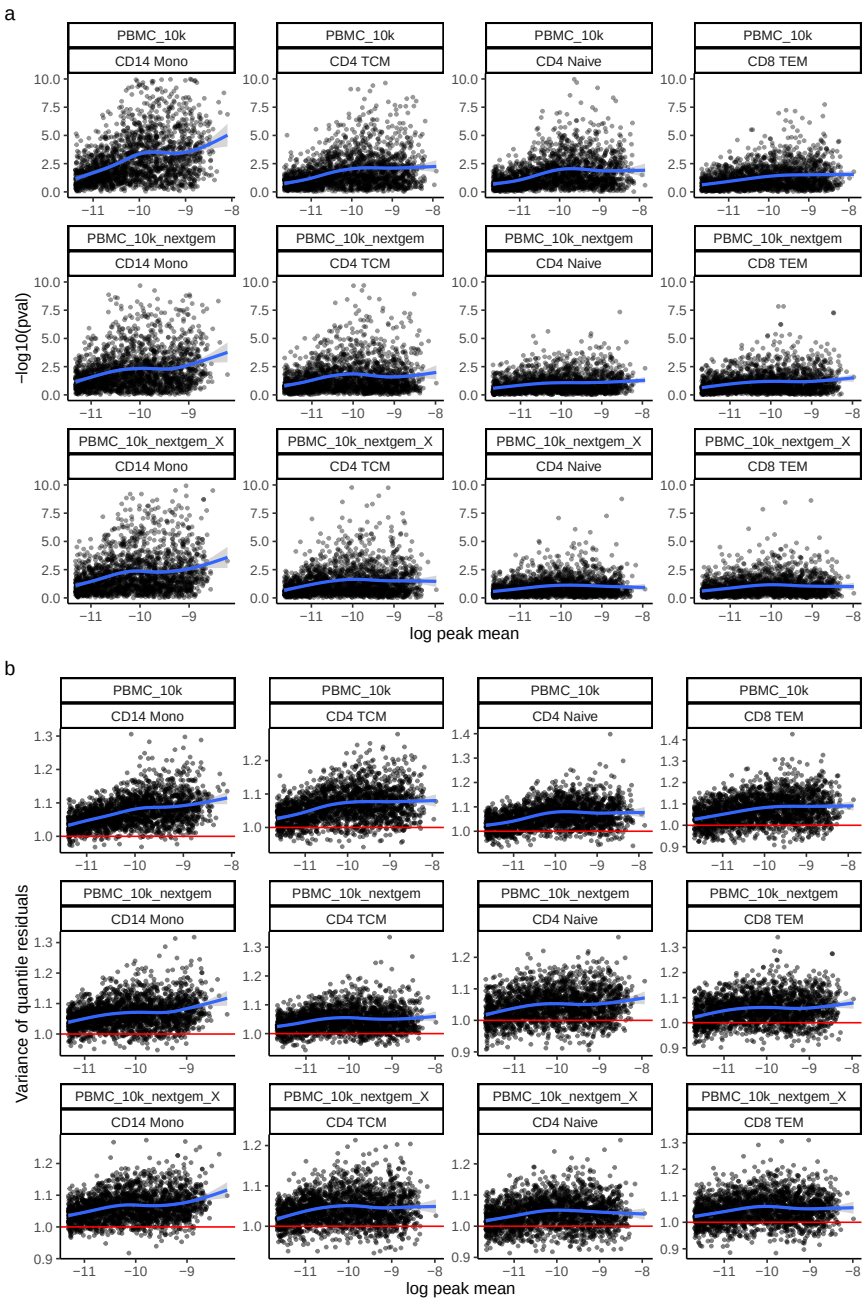
Supplementary Table 1: Summary of single-cell (nucleus) multimodal data used in analyses. Cell types with more than 100 cells are considered.

Data sets	10x [5]	10x [6]	10x [7]	10x [11]	GSE214979 [8]	Brain data from Kellis and Tsai labs [13]
Tissue	PBMC	PBMC	PBMC	PBMC	Brain	Brain
#cells/nuclei	11,457	9,229	9,517	2,590	105,332	65,046
#cell types	16	14	15	7	8	8
Median RNA depth	3,810	2,833	2,664	3,843.5	9,593	3,097
Median ATAC depth	13,030	13,826	13,156	10,180	6,638	2,159
#genes	36,601	36,601	36,601	36,601	36,601	36,601
#peaks	155,060	140,253	140,460	97,957	380,154	258,671
#samples	1	1	1	1	15	19

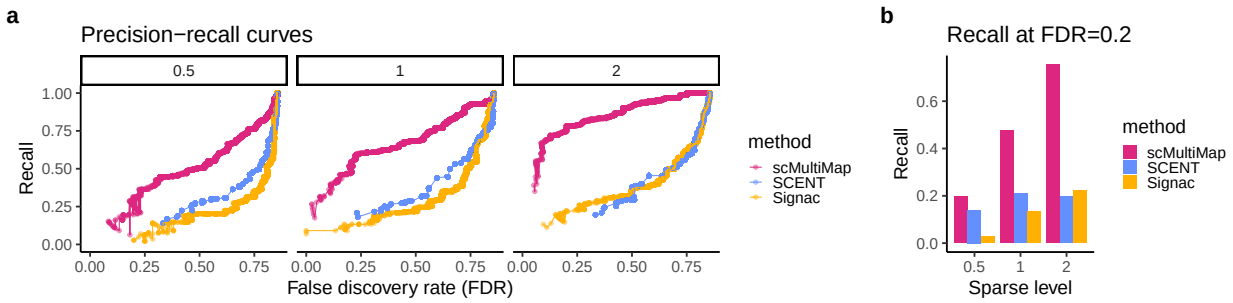
Supplementary Figures



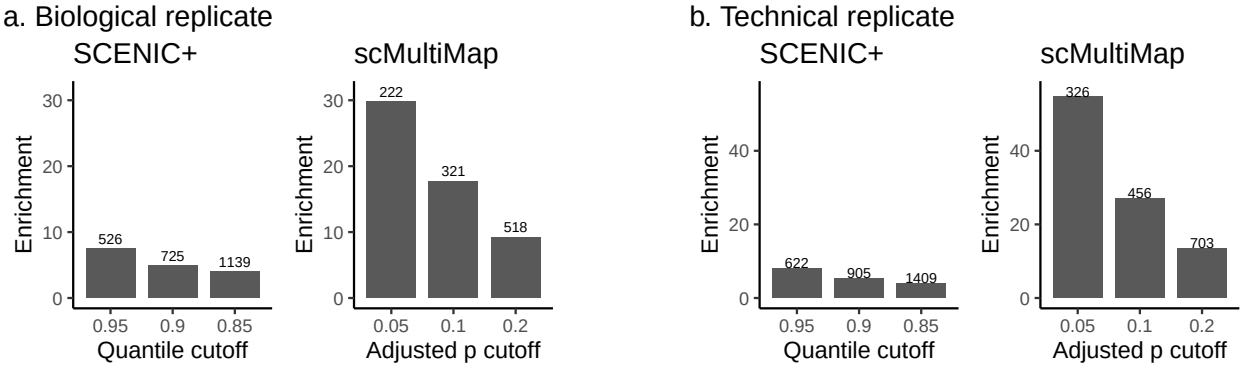
Supplementary Figure 1: Spearman correlations between sequencing depths of ATAC and RNA in three single-cell multimodal datasets on PBMC from 10x Genomics: PBMC_10k [5], PBMC_10k_nextgem [6], PBMC_10k_nextgem_X [7] across four cell types in PBMC: CD14+ Monocyte (CD14 Mono), CD4+ Memory T Cell (CD4 TCM), CD4+ Naive T cell (CD4 Naive), CD8+ Effector Memory T Cell (CD8 TEM). Sequencing depths are computed as the total number of gene counts for RNA and peak counts for ATAC. Source data are provided as a Source Data file.



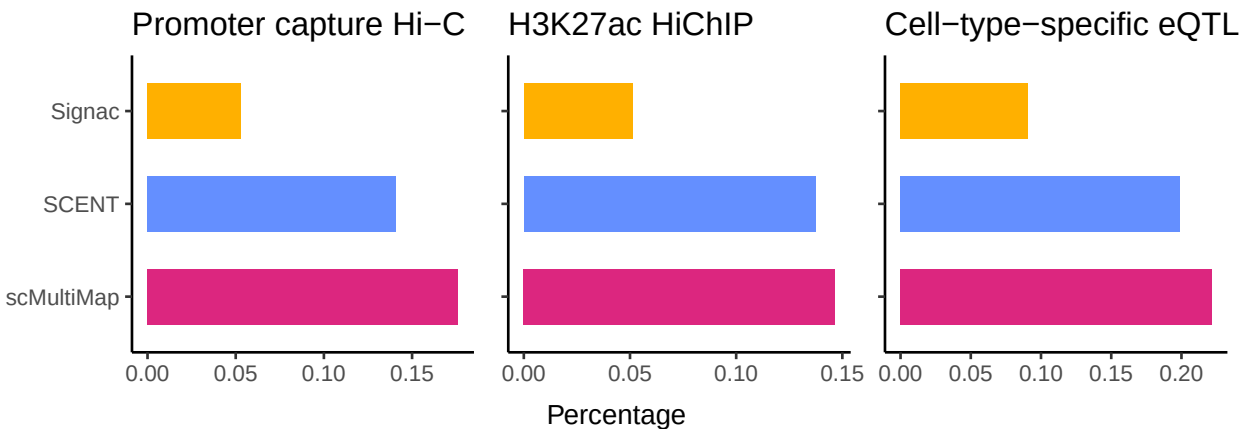
Supplementary Figure 2: Evidence for overdispersion in peak counts in three single-cell multimodal datasets from 10x Genomics and four cell types (same as Supplementary Figure 1). **a**. Results from the goodness of fit test against Poisson for peaks with varying mean accessibility. -$\log_{10}$ p-values were truncated at 10 for visualization. **b**. Variance of quantile residuals for peaks with varying mean accessibility. Variance above 1 (denoted by the red line) indicates overdispersion, i.e. variance that cannot be explained by a Poisson model. Blue curves show the smooth fit by generalized additive models. See Supplementary Methods for more details on this analysis. Source data are provided as a Source Data file.



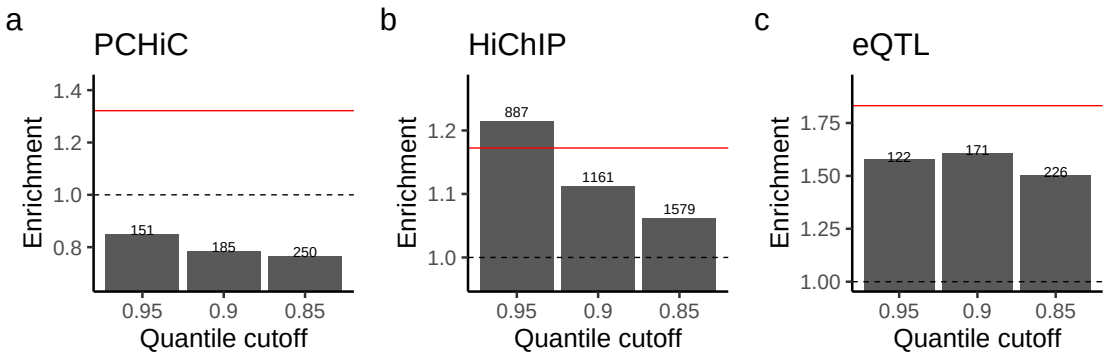
Supplementary Figure 3: Precision-recall curves (**a**) and recall at FDR=0.2 (**b**) of scMultiMap, SCENT, and Signac across sparsity levels of scATAC-seq data. Different sparsity levels were simulated by setting the simulation sequencing depths to be the sequencing depths observed on CD14 monocytes from PBMC data [6] multiplied by 0.5, 1, and 2, such that the simulated peaks have a median sparsity of 0.78, 0.62 and 0.42, respectively. Source data are provided as a Source Data file.



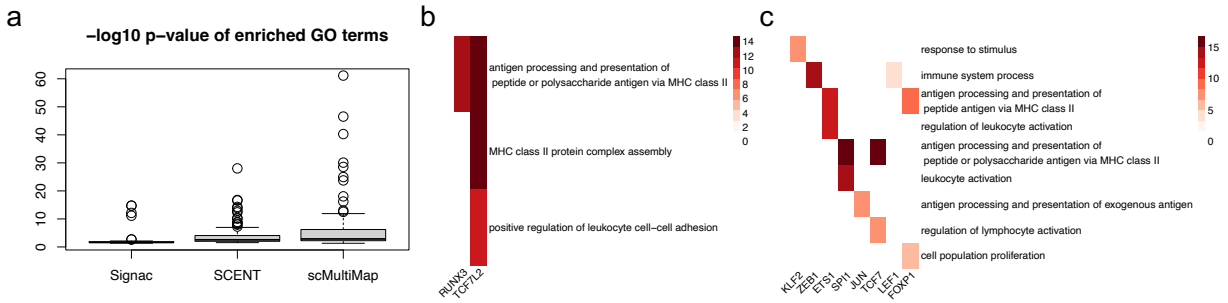
Supplementary Figure 4: Enrichment of reproduced pairs between biological replicates (**a**) and between technical replicates (**b**) of single-cell multimodal data on CD14 monocytes across varying cutoffs for quantiles of variable importance scores (SCENIC+) and for adjusted p-values (scMutliMap). Enrichment is quantified by odds ratio. Numbers above bars give the counts of overlapping pairs. Source data are provided as a Source Data file.



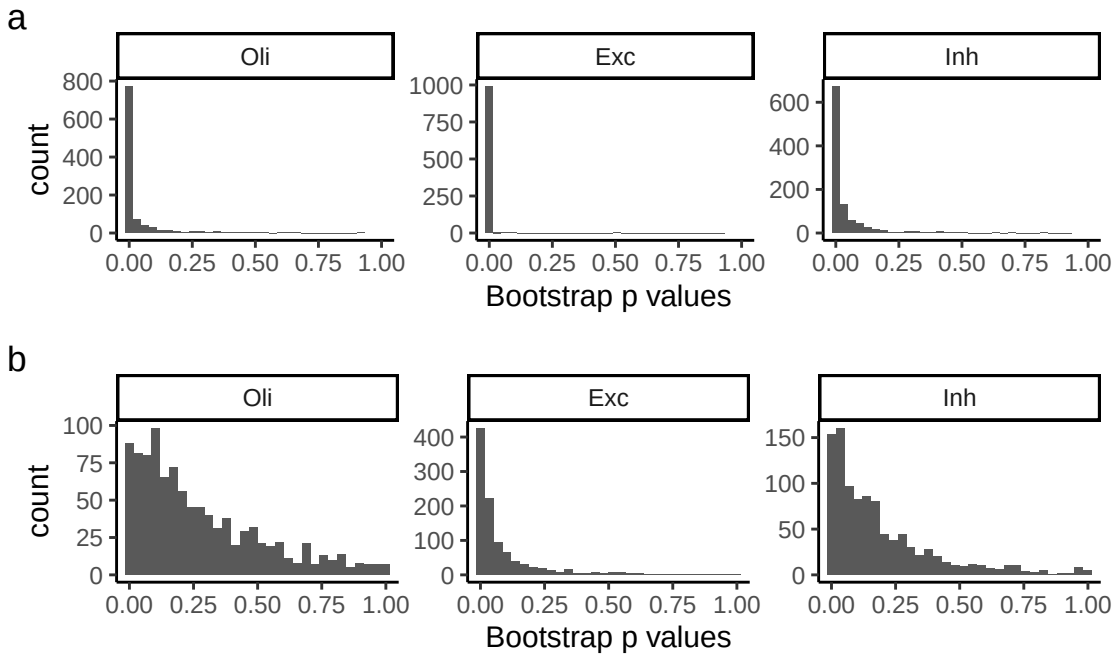
Supplementary Figure 5: Consistency of scMultiMap findings with orthogonal data modalities in percentages. The rest of the figure legend is the same as Figure 2b. Source data are provided as a Source Data file.



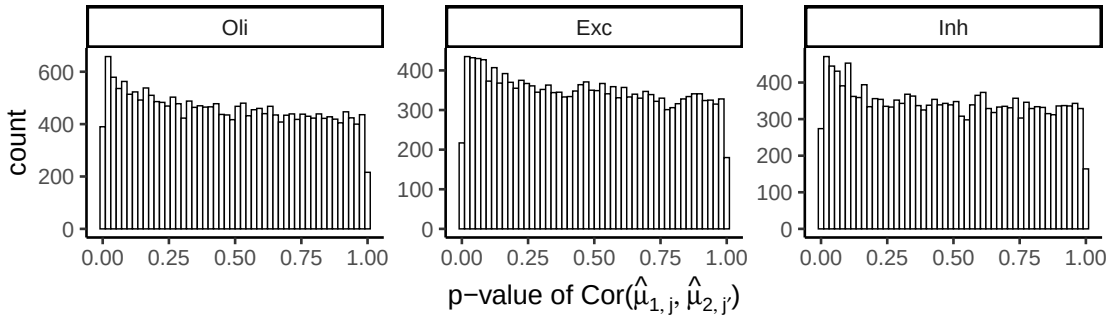
Supplementary Figure 6: Consistency of SCENIC+ identified pairs in CD14 monocytes with orthogonal data modalities across varying cutoffs for quantiles of variable importance scores. Data modalities considered here are the same as in Figure 2b, including PCHiC (**a**), HiChIP (**b**), and eQTL (**c**). The red solid horizontal lines denote the consistency by scMultiMap in Figure 2b. Enrichment is quantified by odds ratio. Numbers above bars give the counts of overlapping pairs. Source data are provided as a Source Data file.



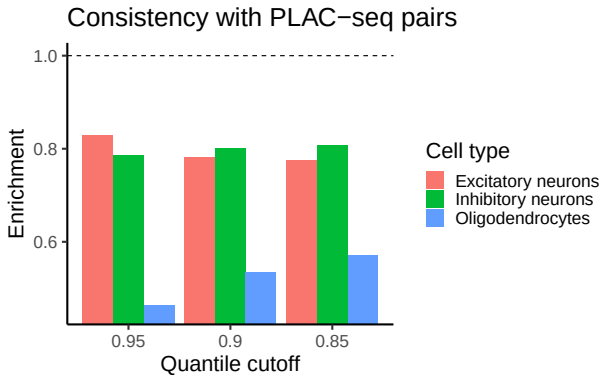
Supplementary Figure 7: GO enrichment results for Signac and SCENT for the same analysis in Figure 2c. **a**. −log10 p-values of enriched GO terms among TF trios for Signac, SCENT and scMultiMap. Boxplots display the median (center), the first to the third quartiles (box), and whiskers extending to values within 1.5× the distance between the quartiles. Points indicate outliers beyond this range. **b-c**. Enrichment of GO biological processes among the target genes in trios identified for each TF in CD14 monocytes for Signac (**b**) and SCENT (**c**), respectively. Source data are provided as a Source Data file.



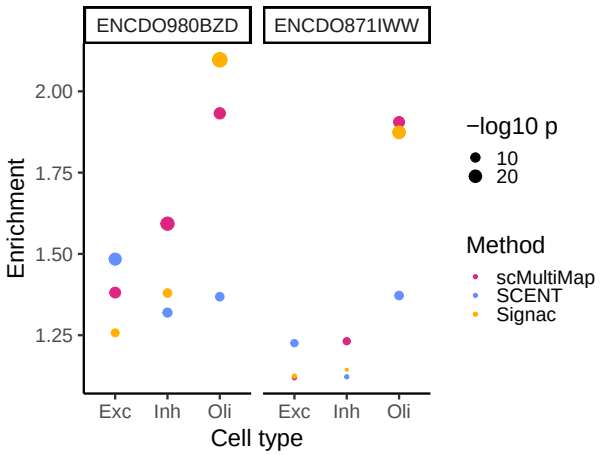
Supplementary Figure 8: Significant variations in mean gene expression (**a**) and mean peak accessibility (**b**) across biological samples based on single-cell multimodal dataset on brain from [8]. We evaluated the statistical significance of mean variations using one-sided bootstrap p-values of F-test statistics. See Supplementary Methods for more details on this analysis. Source data are provided as a Source Data file.



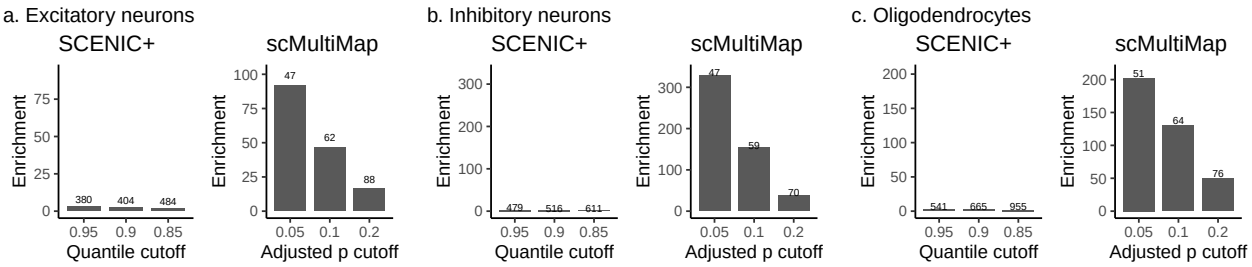
Supplementary Figure 9: Evidence for correlated variations between peaks and genes based on single-cell multimodal dataset on brain from [8]. We evaluated the statistical significance of correlations between peak and gene variations with two-sided p-values of Pearson correlations. See Supplementary Methods for more details on this analysis. Source data are provided as a Source Data file.



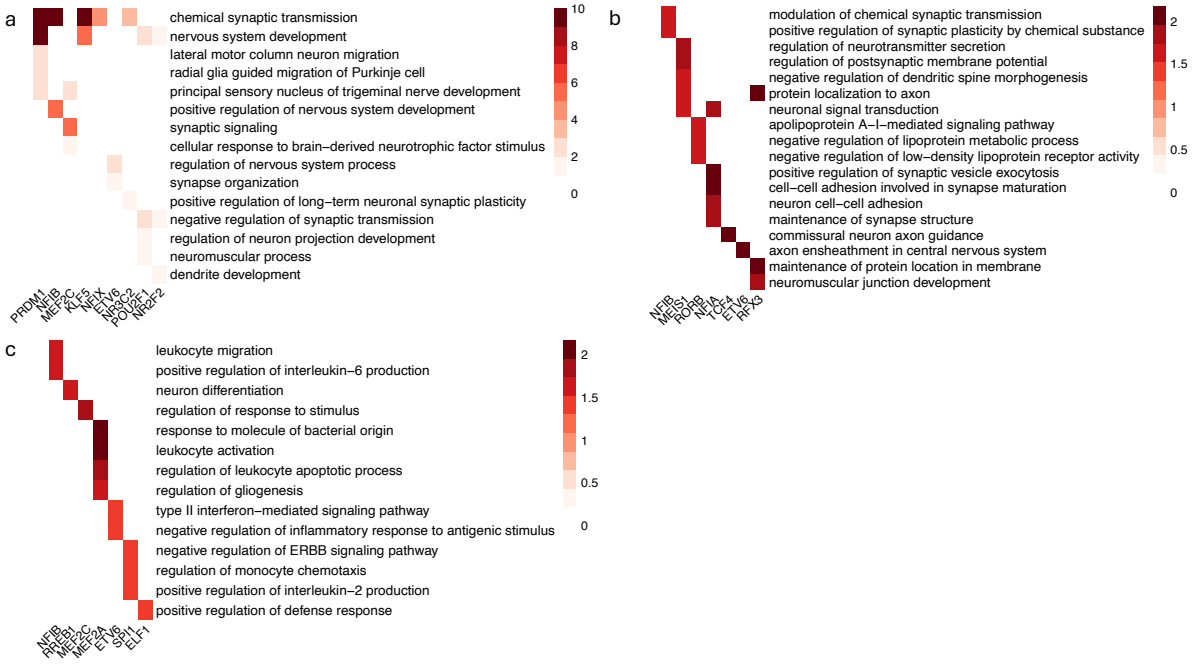
Supplementary Figure 10: Consistency of SCENIC+ identified pairs in three major brain cell types with PLAC-seq [12] across varying cutoffs for quantiles of variable importance scores. The same analysis is performed for scMultiMap in Figure 3c. Enrichment is quantified by odds ratio. Source data are provided as a Source Data file.



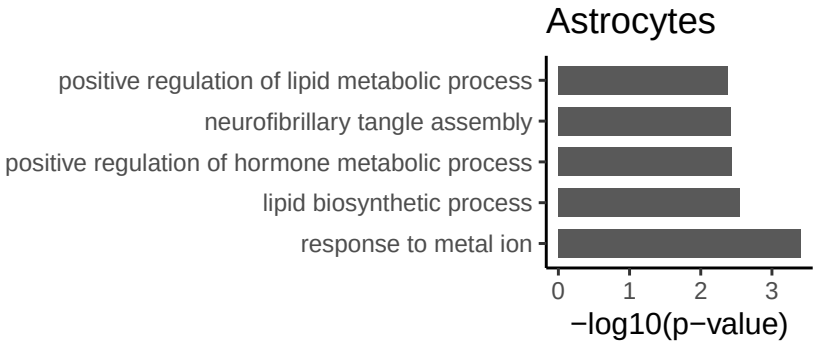
Supplementary Figure 11: Consistency of peaks from significant pairs (BH-adjusted p-value$<$ 0.2) with annotated enhancers in dorsolateral prefrontal cortex tissues from two aged donors (ENCDO980BZD and ENCDO871IWW) in the ENCODE consortium [17] across three major brain cell types: excitatory neurons (Exc), inhibitory neurons (Inh) and oligodendrocytes (Oli). We use the regions with annotations EnhA1 and EnhA2 from a 18-state ChromHMM [18] model, which are enhancer states with strong H3K27ac signals [19]. The enrichment is quantified by odds ratio and p-values are from one-sided Fisher exact tests. Source data are provided as a Source Data file.



Supplementary Figure 12: Enrichment of reproduced pairs between two single-cell multimodal datasets [8, 13] on major brain cell types (**a**. Excitatory neurons, **b**. Inhibitory neurons, **c**. Oligodendrocytes) across varying cutoffs for quantiles of variable importance scores (SCENIC+) and for adjusted p-values (scMultiMap). Enrichment is quantified by odds ratio. Numbers above bars give the counts of overlapping pairs. Source data are provided as a Source Data file.



Supplementary Figure 13: Enrichment of GO biological processes among the target genes in trios identified for each TF in inhibitory neurons (**a**), oligodendrocytes (**b**), and microglia (**c**). 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.



Supplementary Figure 14: Enrichment of GO biological processes among the genes from significantly differential peak-gene pairs in astrocytes (Supplementary Methods). BH-adjusted p-values from one-sided Fisher exact tests are shown. Source data are provided as a Source Data file.

Supplementary References

[1] Van der Vaart, A. W. *Asymptotic statistics*, vol. 3 (Cambridge university press, 2000).

[2] Choudhary, S. & Satija, R. Comparison and evaluation of statistical error models for scrna-seq. *Genome biology* **23**, 27 (2022).

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

[4] Dunn, P. K. & Smyth, G. K. Randomized quantile residuals. *Journal of Computational and graphical statistics* **5**, 236–244 (1996).

[5] Pbmc from a healthy donor - granulocytes removed through cell sorting (10k). URL `https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard` Date Published: 2020-09-09.

[6] 10k human pbmcs, multiome v1.0, chromium controller. URL `https://www.10xgenomics.com/datasets/10-k-human-pbm-cs-multiome-v-1-0-chromium-controller-1-standard-2-0-0`. Date Published: 2021-08-09.

[7] 10k human pbmcs, multiome v1.0, chromium x. URL `https://www.10xgenomics.com/datasets/10-k-human-pbm-cs-multiome-v-1-0-chromium-x-1-standard-2-0-0`. Date Published: 2021-08-09.

[8] Anderson, A. G. *et al.* Single nucleus multiomics identifies zeb1 and mafb as candidate regulators of alzheimer's disease-specific cis-regulatory elements. *Cell Genomics* **3** (2023).

[9] Sun, T., Song, D., Li, W. V. & Li, J. J. scdesign2: a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome biology* **22**, 163 (2021).

[10] Su, C. *et al.* Cell-type-specific co-expression inference from single cell rna-sequencing data. *Nature Communications* **14**, 4846 (2023).

[11] Pbmc from a healthy donor - granulocytes removed through cell sorting (3k). URL `https://www.10xgenomics.com/datasets/`

pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-3-k-1-standard-
Date Published: 2020-09-09.

[12] Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).

[13] Xiong, X. *et al.* Epigenomic dissection of alzheimer's disease pinpoints causal variants and reveals epigenome erosion. *Cell* **186**, 4422–4437 (2023).

[14] Bravo González-Blas, C. *et al.* Scenic+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature methods* **20**, 1355–1367 (2023).

[15] Zhang, R., Atwal, G. S. & Lim, W. K. Noise regularization removes correlation artifacts in single-cell rna-seq data preprocessing. *Patterns* **2** (2021).

[16] Hopfensitz, M. *et al.* Multiscale binarization of gene expression data for reconstructing boolean networks. *IEEE/ACM transactions on computational biology and bioinformatics* **9**, 487–498 (2011).

[17] Moore, J. E. *et al.* Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

[18] Ernst, J. & Kellis, M. Chromhmm: automating chromatin-state discovery and characterization. *Nature methods* **9**, 215–216 (2012).

[19] Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).