nature portfolio

Peer Review File

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

Corresponding Author: Dr Chang Su

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Assess the correlation between ATAC-seq peaks and gene expression using multi-omic data is a critical step to understand gene regulations. scMultiMap, proposed in this paper, addresses the inflated type I error, reduced power, and heavy computational burdens of existing methods. The method is very carefully evaluated, and the results demonstrate the clear advantages of scMultiMap. I have a few questions about the method and more minor concerns for the presentations.

1. For the equation after equation (4), an additive noise assumption is used. This is not ideal for count data, and it is not consistent with equation (3)? What is the distribution of \epsilon_{2,i,j'}?

2. Does the asymptotic result for the distribution of T_{jj'} depend on the range of the counts? for example, would the results consistently hold for negative binomial with large or small over-dispersion?

3. When accounting for individual specific effects, as described in the middle of page 17, if subject-specific mean is used, would it remove potential group effects, such as difference between AD and control subjects?

Minor concerns.
1. At page 6: "Additionally, correlations of marginally normalized data are known to be biased by mean and overdispersion [14, 25], and random sampling is inadequate to adjust for such bias and may generate invalid p-values for inference"

This is an important message and worth to elaborate.

2. Figure 1A. What is the meaning of each point. The Type I error should be a proportion and thus each point should the proportion of p-values smaller than 0.05 in a group of tests? From the description of Methods, one point should be one replicate. though it would be helpful to clarify it around Figure 1.

3. Figure 2C shows the GO enrichment. Is that only for scMultiMap, what about Signac and SCENT?

4. for the part of "scMultiMap mapped GWAS variants of Alzheimer", it is not clear which datasets was used to infer candidate cis-regulatory elements.

5. For the GO enrichment analysis in Figure 4C. it appears most GO terms are very generic. It may not be surprising since most GWAS associated genes are not enriched in certain functional categories. I am wondering whether gene expression abundance is a possible confounding factor. For example, the author may choose gene sets with highest expression in microglia and test whether they are enriched in some GO terms.

6. It might be helpful to mention other resources to identify/confirm the functional roles of regulatory elements and genes are through gene knockout studies, such as large consortium IGVF and MorPhiC.

7. For the journal of NC, many audiences may not be statisticians, and thus it would be helpful to clarify the derivation of the variance term in equation (2) and (3). For example, by model (1) x_ij follows a Poisson distribution given z_ij.

$Var(x_{ij}) = E(Var(x_{ij}|z_{ij})) + Var(E(x_{ij}|z_{ij})) = E(s_i z_{ij}) + Var(s_i z_{ij}) = s_i \mu_{ij} + s_i^2 \sigma_{jj}^2$

In addition, it looks like the authors are using \sigma, instead of \sigma^2 to indicate variance. it would be important to clarify this somewhat unconventional notation choice.

(Remarks on code availability)
it is clearly organized with appropriate example.

Reviewer #2

(Remarks to the Author)
The authors present scMultiMap, a statistical method and R package designed to map enhancer-gene associations using single-cell multimodal data. The authors convincingly show that previous approaches have not been able to maintain type I error rates and have reduced power. Unlike previous approaches, scMultiMap addresses challenges such as data sparsity and sequencing depth variation for greater power and maintenance of type I error rates. Most importantly, the author have significantly reduced computational complexity of the problem. I think the authors have been thorough in their presentation, benchmarking and application examples and scMultiMap will be a valuable tool, but I do have some minor comments:

1) In the introduction the authors do not account for the possibility of using independently generated single cell RNA-sequencing and ATAC-seq data from the same sample. While this setup is inferior to single cell multimodal data, it has been shown that cell-type specific enhancers and promoters can be still be accurately identified when multiple samples are available (10.1016/j.cell.2021.07.039).

2) I really like the elegant statistical model the authors derived for the computation of correlations between gene expression and enhancer accessibility, but for completeness please provide a justification for the choice of a Poisson distribution to model the gene count in a single cell. Did the authors consider a Negative Binomial distribution for the gene count or the accessibility data?

3) I would suggest that in the Overview section for the statistical model the authors provide details on how the model can be extended to include data from multiple samples with the help of indicators as described in the Supplementary. It would also be great if the authors could comment on the computational implementation of their model.

4) The author only discuss the Signac and SCENT, as alternative methods for the identification of gene and enhancer pairs. The following tools are also popular and should be added to the comparison:
- SCENIC+ (10.1038/s41592-023-01938-4)
- scMEGA (10.1093/bioadv/vbad003)

5) Given the severe sparsity of scATAC-seq data, it would be interesting to simulate this effect to assess the impact on the power of scMultiMap. This could be easily achieved for the PBMC datasets.

6) It is odd that the results in Figure 2B are presented as only a count and not a percentage of correctly identified peak-gene pairs from the total of each capture type. For the cell-type specific pairs, the figure legend and ideally figure should identify that this result was derived for monocytes.

7) To ensure that real enhancers are indeed identified in the brain it would be great to see the overlap with ENCODE predicted enhancers across fetal and adult human tissue via chromHMM (10.1038/nprot.2017.124).

8) In the AD GWAS variant analysis, it would be interesting to see whether the associated genes have already been largely identified by single cell differential expression analysis across different cell types (10.1038/s41586-024-07606-7). This would lead further evidence that scMultiMap is capable of identification of real enhancer gene associations.

9) The associated R – package should be updated to include more extensive documentation and vignettes.

(Remarks on code availability)
As stated in my comments to the authors the documentation could be improved but overall I think this is a great effort.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
the authors have addressed all my concerns. Congratulation for this important work! Wei Sun.

(Remarks on code availability)
well organized and documented.

Reviewer #2

(Remarks to the Author)
All my comments have been addressed. It was a pleasure to review this work and I look forward to using scMultiMap in my own work.

(Remarks on code availability)

Summary of Revisions

We sincerely appreciate the valuable suggestions from the senior editor and the two reviewers. We have revised the manuscript by incorporating all suggestions and addressing all comments. Below, we briefly summarize the main revisions, followed by detailed point-by-point responses to the reviewers.

- **New numerical analysis for method evaluation**
  - Evaluation of SCENIC+ on benchmarking datasets
  - Validation based on ENCODE chromatin state annotations from ChromHMM
  - Evaluation of the impact of sparsity on scMultiMap's power
- **Enhanced clarity in describing statistical and computational methods**
  - Derivation of moment-based regressions
  - Clarification on distributional assumptions
  - Details of computational implementation
- **Improved functionality of software implementation**
  - New vignette: scMultiMap for disease-control studies
  - New vignette: scMultiMap for integrative analysis with GWAS results
  - Updated vignette: Introduction to scMultiMap
  - More software updates can be found at GitHub repo `https://github.com/ChangSuBiostats/scMultiMap` and the associated package website `https://changsubiostats.github.io/scMultiMap/`
  - All code for reproducing results in the paper are provided in a GitHub repo: `https://github.com/ChangSuBiostats/scMultiMap_analysis`

Response to Reviewer 1

We sincerely thank you for the careful reading of our manuscript and the valuable and constructive comments. In the following, your comments are shown in *italics*, followed by our point-by-point responses. For ease of reference, we have highlighted the major changes in blue in the revised manuscript.

Questions about the method

"*Assess the correlation between ATAC-seq peaks and gene expression using multi-omic data is a critical step to understand gene regulations. scMultiMap, proposed in this paper, addresses the inflated type I error, reduced power, and heavy computational burdens of existing methods. The method is very carefully evaluated, and the results demonstrate the clear advantages of scMultiMap. I have a few questions about the method and more minor concerns for the presentations.*

Response: Thank you for your positive and encouraging review.

1. "*For the equation after equation (4), an additive noise assumption is used. This is not ideal for count data, and it is not consistent with equation (3)? What is the distribution of* $\epsilon_{2,i,j'}$*?*"

Response: Thank you for this question. Equation (3) does not have an error term as it describes the expected value of the response variable and the error term has an expected value of zero. Equation (3) and the equation after (4) are consistent: the equation after (4) models the random variable $y_{ij'}$, while equation (3) models the expected value $\mathrm{E}(y_{ij'})$ of $y_{ij'}$. The mean zero error term is defined as $\epsilon_{2,i,j'} = y_{ij'} - E(y_{ij'})$, and we do not impose any distributional assumptions on $\epsilon_{2,ij'}$'s in our approach. We give more details below.

To provide context for our discussion, we first rewrite equation (3) and the equation after (4). Equation (3) in the manuscript consists of the following two components:

$$\mathrm{E}(y_{ij'}) = r_i \mu_{2,j'}, \tag{R1}$$

$$\mathrm{Var}(y_{ij'}) = r_i \mu_{2,j'} + r_i^2 \sigma_{2,j'j'}. \tag{R2}$$

The equation after (4) models the observed counts $y_{ij'}$ and has the two following components:

$$y_{ij'} = r_i \mu_{2,j'} + \epsilon_{2,ij'}, \tag{R3}$$

$$(y_{ij'} - r_i \mu_{2,j'})^2 = r_i \mu_{2,j'} + r_i^2 \sigma_{2,j'j'} + \eta_{2,ij'}. \tag{R4}$$

It is seen that (R3) models $y_{ij'}$ and (R1) describes $\mathrm{E}(y_{ij'})$. Correspondingly, the noise term $\epsilon_{2,ij'} = y_{ij'} - \mathrm{E}(y_{ij'})$ is a mean zero random variable. Equations (R3) and (R1) are analogous to $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$ and $\mathrm{E}(y_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ in classical linear regressions, and these two equations are consistent. Similarly, (R4) models random variables $(y_{ij'} - r_i \mu_{2,j'})^2$ and (R2) describes the expected values $\mathrm{Var}(y_{ij'}) = \mathrm{E}[(y_{ij'} - r_i \mu_{2,j'})^2]$.

In our approach, we do not make any parametric assumptions on the distribution of $\epsilon_{2,ij'}$, and our proposed estimation and inference framework in scMultiMap can be applied to any distribution of $\epsilon_{2,ij'}$. This flexibility is enabled by the use of moment-based regressions (e.g., (R3)-(R4)), which only relies on the first two moments of the observed counts, instead of the full likelihood. Furthermore, we note that this flexible framework can accommodate commonly used distribution as special cases, such as the negative binomial distribution. For example, when $v_{ij'}$ follows a gamma distribution, $y_{ij'}$ will follow a negative binomial distribution, and the distribution of $\epsilon_{2,ij'}$ will follow correspondingly (see Results section 'Overview of scMultiMap' for more details).

To more clearly describe our method, we have incorporated the above discussion to the Methods section 'scMultiMap method' in the revised manuscript.

2. "*Does the asymptotic result for the distribution of $T_{jj'}$ depend on the range of the counts? For example, would the results consistently hold for negative binomial with large or small overdispersion?*"

Response: Thank you for this thoughtful question. The asymptotic result for the distribution of $T_{jj'}$ does not depend on the range of counts. In particular, the null distribution of $T_{jj'}$ is derived based on the theory of M-estimation, and it does not depend on the range of the counts or the overdispersion parameter of the negative binomial distribution [1]. We have added a clarification on this point in the Supplementary Methods, which is also referenced in Methods section 'scMultiMap method' in the main text.

3. "*When accounting for individual specific effects, as described in the middle of page 17, if subject-specific mean is used, would it remove potential group effects, such as difference between AD and control subjects?*"

Response: Thank you for this question. 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 our paper, peak-gene associations were estimated separately for each group.

To clarify this point, we have incorporated the discussion above to Methods and to Supplementary Discussion.

Minor concerns for the presentations

1. "*At page 6: "Additionally, correlations of marginally normalized data are known to be biased by mean and overdispersion [14, 25], and random sampling is inadequate to adjust for such bias and may generate invalid p-values for inference" This is an important message and worth to elaborate.*"

Response: Thank you for this helpful suggestion. We give more details below.

According to [2, 3] ([14,25] in the manuscript), correlations of marginally normalized data may be more attenuated for genes or peaks with lower abundance levels or smaller overdispersion. This is because genes and peaks with lower abundance or smaller overdispersion suffer more from the sequencing-induced measurement errors, which attenuate the association between correlated genes and peaks. Hence, to achieve 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. However, this can hardly be guaranteed by random sampling.

We have added the above discussion to the Results section 'scMultiMap has better detection accuracy and computational efficiency' in the revised manuscript.

2. "*Figure 1A. What is the meaning of each point. The Type I error should be a proportion and thus each point should the proportion of p-values smaller than 0.05 in a group of tests? From the description of Methods, one point should be one replicate. though it would be helpful to clarify it around Figure 1.*"

Response: In Figure 1A, each point represents a peak-gene pair, and each boxplot contains a total of 1,000 points corresponding to 1,000 randomly selected peak-gene pairs from CD14 monocytes. These pairs were sampled from the full set of candidate peak-gene pairs, with varying mean and variance parameters. As you noted, the value of each point represents the proportion of p-values smaller than 0.05 across 100 null replicates generated from a permutation-based procedure for the corresponding pair.

We have added the above details to the captions of Figures 1 and 3 and to the Methods section 'Experiments for evaluating type I errors and power' in the revised manuscript.

3. "*Figure 2C shows the GO enrichment. Is that only for scMultiMap, what about Signac and SCENT?*"

Response: Thank you for this question. The results presented in Figure 2C are only for scMultiMap. In our revision, we have added results for Signac and SCENT.

To evaluate GO enrichment by Signac, we constructed trios by matching TF-peak associations with peak-target gene associations inferred by Signac, along with target gene-TF associations inferred by CS-CORE [2]. We then applied the same strategy as in the analysis presented in Figure 2C to perform the GO enrichment analysis. A similar analysis was also conducted for SCENT. The results are shown in Figure R1 below.

We found that the statistical significance of enriched GO terms by Signac and SCENT's trios are generally weaker than scMultiMap's (Figure R1A), indicating that the inferred trios are biologically less meaningful. Furthermore, we applied the same procedure as in Figure 2C to visualize the enrichment for Signac and SCENT. Fig-

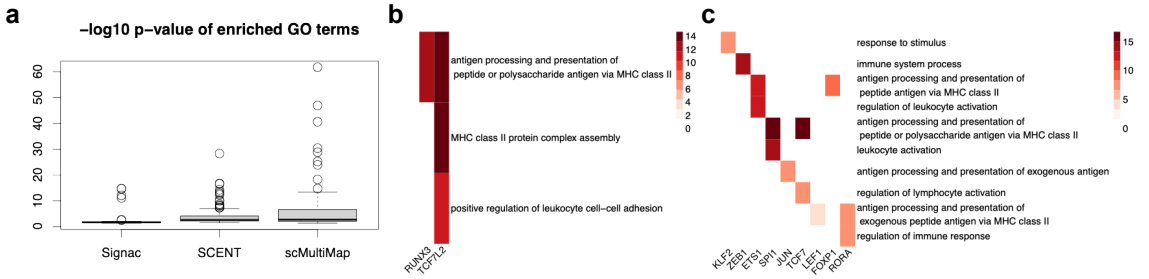

Figure R1: GO enrichment results for Signac and SCENT for the same analysis in Figure 2C. **a**. −log10 p-values of enriched GO terms among TF trios for Signac, SCENT and scMultiMap. **b-c**. Enrichment of GO biological processes among the target genes in trios identified for each TF in CD14 monocytes for Signac (**b**) and SCENT (**c**), respectively.

ure R1 B-C shows fewer immune related and enriched GO terms for TFs by these two methods, which is consistent with the weaker enrichment results observed in Figure R1A.

We have incorporated these results into the Results section 'scMultiMap has higher reproducibility across independent datasets' and Supplementary Figures.

4. "*For the part of "scMultiMap mapped GWAS variants of Alzheimer", it is not clear which datasets was used to infer candidate cis-regulatory elements.*"

Response: Thank you for pointing this out. We used the same datasets as in the previous section entitled "scMultiMap identified biologically relevant gene regulatory mechanisms in brain cells", that is, data from [4]. To clarify this, we have updated the first paragraph of the Results section 'scMultiMap mapped GWAS variants of Alzheimer's disease to target genes in microglia' in the revised manuscript.

5. "*For the GO enrichment analysis in Figure 4C, it appears most GO terms are very generic. It may not be surprising since most GWAS associated genes are not enriched in certain functional categories. I am wondering whether gene expression abundance is a possible confounding factor. For example, the author may choose gene sets with highest expression in microglia and test whether they are enriched in some GO terms.*"

Response: Thank you for this insightful question. Per your suggestion, we have

evaluated the enrichment among top 500 highly expressed genes in microglia. Three GO terms in Figure 4C were found to be significantly enriched (BH adjusted $p-$value < 0.05), including 'regulation of biological process', 'phosphorus metabolic process', and 'ERBB signaling pathway', all of which are generic functional terms. We also considered the top 2,000 and 5,000 highly expressed genes in microglia, and the results remain similar.

To adjust for the potential confounding effect of gene expression levels, we additionally performed the GO enrichment analysis using the top 5,000 highly expressed genes as the background gene set, with the choice of 5,000 being consistent with that considered in our differential association analysis. Two GO terms in Figure 4C remained significantly enriched: 'regulation of biological process' and 'high-density lipoprotein particle assembly', suggesting that the original results were robustness to potential confounding effects from gene expression abundance. A similar analysis was performed for astrocytes, where three out of five GO terms remained significant after using highly expressed genes in astrocytes as the background gene set: 'response to metal ion', 'positive regulation of hormone metabolic process', and 'positive regulation of lipid metabolic process'.

We have updated the Results section 'scMultiMap mapped GWAS variants of Alzheimer's disease to target genes in microglia' to discuss this potential issue with interpreting GO results, and to discuss its robustness when highly expressed genes are used as the background.

6. "*It might be helpful to mention other resources to identify/confirm the functional roles of regulatory elements and genes are through gene knockout studies, such as large consortium IGVF and MorPhiC.*"

Response: Thank you for this great suggestion! In the revised Discussion section, we discuss both IGVF and MorPhiC and how scMultiMap may contribute to these large consortium efforts.

7. "*For the journal of NC, many audiences may not be statisticians, and thus it would be helpful to clarify the derivation of the variance term in equation (2) and (3).*

For example, by model (1) x_{ij} follows a Poisson distribution given z_{ij}.

$$\mathrm{Var}(x_{ij}) = \mathrm{E}(\mathrm{Var}(x_{ij}|z_{ij})) + \mathrm{Var}(E(x_{ij}|z_{ij})) = \mathrm{E}(s_i z_{ij}) + \mathrm{Var}(s_i z_{ij}) = s_i \mu_{ij} + s_i^2 \sigma_{jj}^2$$

In addition, it looks like the authors are using σ, instead of σ^2 to indicate variance. it would be important to clarify this somewhat unconventional notation choice. "

Response: Thank you for this helpful suggestion. In the revised manuscript, we have added the derivation of mean, variance, and covariance terms and clarified the definition of σ_{jj} as variance in the Methods section 'scMultiMap method'.

Response to Reviewer 2

We appreciate your valuable suggestions. In the following, your comments are shown in *italics*, followed by our point-by-point responses. For ease of reference, we have highlighted the major changes in blue in the revised manuscript.

"*The authors present scMultiMap, a statistical method and R package designed to map enhancer-gene associations using single-cell multimodal data. The authors convincingly show that previous approaches have not been able to maintain type I error rates and have reduced power. Unlike previous approaches, scMultiMap addresses challenges such as data sparsity and sequencing depth variation for greater power and maintenance of type I error rates. Most importantly, the author have significantly reduced computational complexity of the problem. I think the authors have been thorough in their presentation, benchmarking and application examples and scMultiMap will be a valuable tool.*

Response: We greatly appreciate your positive and encouraging review.

1. "*In the introduction the authors do not account for the possibility of using independently generated single cell RNA-sequencing and ATAC-seq data from the same sample. While this setup is inferior to single cell multimodal data, it has been shown that cell-type specific enhancers and promoters can be still be accurately identified when multiple samples are available (10.1016/j.cell.2021.07.039).*"

Response: Thank you for pointing us to this body of literature. Per your suggestion, we have discussed this approach as an alternative method to map enhancers and target genes in Introduction section, and commented on the challenge of using scMultiMap on such data in the Discussion section of the revised manuscript.

2. "*I really like the elegant statistical model the authors derived for the computation of correlations between gene expression and enhancer accessibility, but for completeness please provide a justification for the choice of a Poisson distribution to model the gene count in a single cell. Did the authors consider a Negative Binomial distribution for the gene count or the accessibility data?*"

Response: Thank you for your question. Our model has two components: the expression model and the measurement model. While the Poisson distribution is used in our measurement model, the resulting counts from both model components are generally not Poisson, due to the additional variability introduced by the expression model. Although we do not impose parametric assumptions on the expression model to allow for greater modeling flexibility, under certain distributional assumptions, the resulting gene or peak counts can follow a Negative Binomial distribution. We give more details below.

Recall that we assume the following model for gene and peak counts:

$$x_{ij}|z_{ij} \sim \text{Poisson}(s_i z_{ij}), y_{ij'}|v_{ij'} \sim \text{Poisson}(s_i v_{ij'}), \quad (z_{i1}, \ldots, z_{ip}, v_{i1}, \ldots, v_{iq}) \sim F_{p+q}(\cdot), \tag{R5}$$

where $x_{ij}, y_{ij'}$ are the observed counts of gene j and peak j' in cell i, $z_{ij}, v_{ij'}$ are the underlying abundance level of gene j and peak j' in cell i, assumed to follow an unknown nonnegative $p+q$-variate distribution $F_{p+q}(\cdot)$, and s_i is the sequencing depth. As the underlying expressions z_{ij}'s are random variables, the marginal distribution of gene count x_{ij} is usually not Poisson and has a higher dispersion than Poisson depending on $F_{p+q}(\cdot)$. In fact, if $F_{p+q}(\cdot)$ is Gamma, the distribution of count x_{ij} is negative binomial under (R5) [5]. Similar conclusions will also hold for the peak counts $y_{ij'}$'s. Hence, our model includes modeling x_{ij}'s and $y_{ij'}$'s using negative binomial as a special case and is more flexible.

In Equation (R5), we choose Poisson to model the conditional distribution of $x_{ij}|z_{ij}$, referred to as the measurement model in scMultiMap, as it agrees with the existing literature [6, 7, 5] that a Poisson distribution is usually sufficient to characterize the variations introduced by the sequencing in droplet-based scRNA-seq. Similar observations were also made for scATAC-seq data in [8, 9], which supports the use of Poisson for the conditional distribution of peak counts $y_{ij'}|v_{ij'}$. In the case that mRNA molecules or DNA fragments exhibit greater variations than Poisson from the sequencing step, it can be useful to adapt scMultiMap to model $x_{ij}|z_{ij}$ and $y_{ij'}|v_{ij'}$ using alternative distributions, such as the Negative Binomial, that allow for higher dispersion. This would require re-deriving the moment conditions used in scMultiMap

(given in Equations (2)-(3) in the paper) under the new distribution for $x_{ij}|z_{ij}$ and $y_{ij'}|v_{ij'}$, 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.

We have added the above discussion to the Discussion section of the revised manuscript.

3. "*I would suggest that in the Overview section for the statistical model the authors provide details on how the model can be extended to include data from multiple samples with the help of indicators as described in the Supplementary. It would also be great if the authors could comment on the computational implementation of their model.*"

Response: Thank you for these helpful suggestions. We have updated the Overview section of scMultiMap (Pages 5-6) to provide details on how the model can be extended to include data from multiple samples, and to comment on the computational implementation of the model. For your convenience, we've included the revised text below:

"For genes $j \in [p]$, we estimate the mean parameters $\mu_{1,j}$'s and variance parameters $\sigma_{1,jj}$'s using moment-based regressions for the first and second moments, iteratively until convergence. In this process, $\sigma_{1,jj}$' s are estimated with $\hat{\mu}_{1,j}$'s and $\mu_{1,j}$'s are estimated with weights involving $\hat{\sigma}_{1,jj}$'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 $\sigma_{2,j'j'}$'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."

4. "*The author only discuss the Signac and SCENT, as alternative methods for the identification of gene and enhancer pairs. The following tools are also popular*

and should be added to the comparison: - SCENIC+ (10.1038/s41592-023-01938-4) - scMEGA (10.1093/bioadv/vbad003)"

Response: Thank you for your valuable suggestion. SCENIC+ [10] and scMEGA [11] are two popular tools for gene regulatory network inference, and scMEGA includes an intermediate step that identifies candidate gene-enhancer pairs. In this step, scMEGA [11] uses the method from ArchR [12] to link peaks to genes by computing Pearson's correlations between peak and gene expression data. From this view, this step in scMEGA adopts the same methodology as in Signac evaluated in our benchmarking analysis, as Signac also computes the Pearson's correlations for identifying gene-enhancer pairs, and is expected to have the same performance. We have updated the Results section "scMultiMap has better detection accuracy and computational efficiency" to make this connection when we introduce the Signac method.

SCENIC+ takes a different approach to identifying gene-enhancer pairs and comparing it with other methods can be challenging. While all the methods we have considered (scMultiMap, Signac, SCENT, scMEGA and ArchR) 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 $95\%, 90\%$ and 85% 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

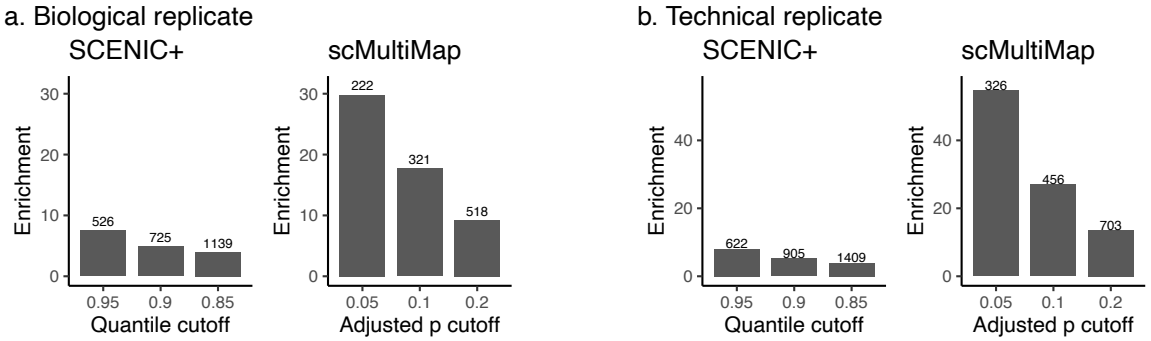

Figure R2: Enrichment of reproduced pairs between biological replicates (**a**) and between technical replicates (**b**) of single-cell multimodal data on CD14 monocytes across varying cutoffs for quantiles of variable importance scores (SCENIC+) and for adjusted p-values (scMutliMap). Enrichment is quantified by odds ratio. Numbers above bars give the counts of overlapping pairs.

different approaches is only possible when their type I error rates are comparable.

Nevertheless, with the caveats mentioned above, we evaluated SCENIC+ in reproducibility and consistency analyses using orthogonal data modalities on PBMC and brain data, as in Figures 2A-B and 3C-D in the manuscript, and compared its performance with scMultiMap on the same tasks.

When applied to technical and biological replicates of single-cell multimodal data on CD14 monocytes, the enrichment of reproduced discoveries by SCENIC+ is consistently lower than scMultiMap across cutoffs (Figure R2). When considering similar numbers of discoveries, SCENIC+ also yields lower enrichment (Figure R2). Figure R3 further shows that the enrichment among three sets of orthogonally identified peak-gene pairs are also generally lower compared to scMultiMap's results in Figure 2B, which is marked with horizontal red lines. In particular, the enrichment is smaller than 1 for all cutoffs when compared against Promoter capture Hi-C.

Across three major brain cell types, the enrichment of reproduced discoveries by SCENIC+ (Figure R4) is again consistently lower than scMultiMap (Figure 3C). The number of reproduced pairs are smaller for scMultiMap due to the much shallower sequencing depths in [13] data (Supplementary Table 1), yet SCENIC+ identified a similar number of pairs regardless of the amount of information in the data. When comparing SCENIC+ identified pairs against PLAC-seq data [14] on major brain cell

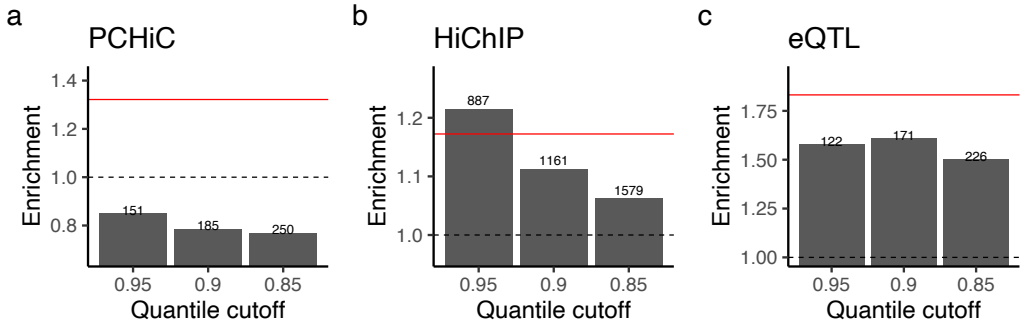

Figure R3: Consistency of SCENIC+ identified pairs in CD14 monocytes with orthogonal data modalities across varying cutoffs for quantiles of variable importance scores. Data modalities considered here are the same as in Figure 2B, and the vertical lines denote the consistency by scMultiMap in Figure 2B. Enrichment is quantified by odds ratio. Numbers above bars give the counts of overlapping pairs.

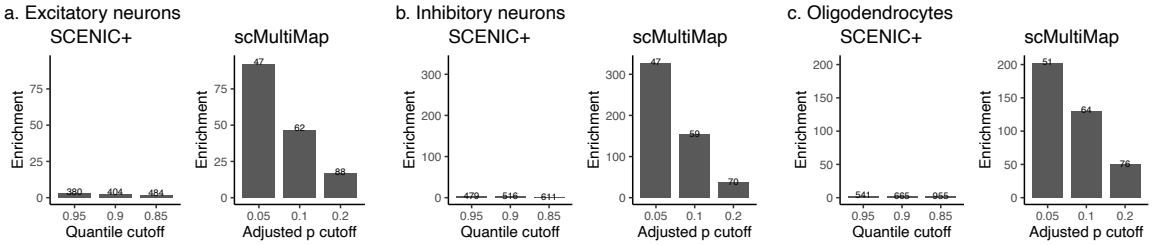

Figure R4: Enrichment of reproduced pairs between two single-cell multimodal datasets [4, 13] on major brain cell types (**a**. Excitatory neurons, **b**. Inhibitory neurons, **c**. Oligodendrocytes) across varying cutoffs for quantiles of variable importance scores (SCENIC+) and for adjusted p-values (scMutliMap). Enrichment is quantified by odds ratio. Numbers above bars give the counts of overlapping pairs.

types, the enrichment of SCENIC+ is smaller than 1 across all cell types and quantile cutoffs (Figure R3), suggesting relative poor performance for identifying true peak-gene pairs.

Despite the relatively poor performance of SCENIC+ in our evaluation, 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 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

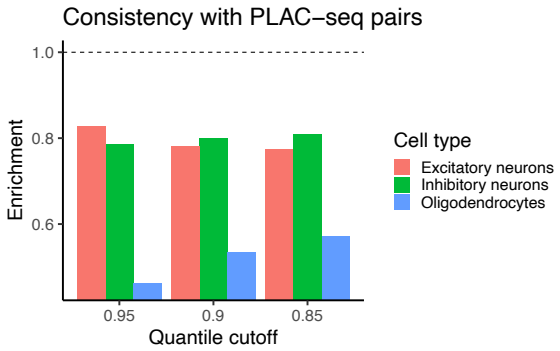

Figure R5: Consistency of SCENIC+ identified pairs in three major brain cell types with PLAC-seq [14] across varying cutoffs for quantiles of variable importance scores. The same analysis is performed for scMultiMap in Figure 3C. Enrichment is quantified by odds ratio.

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 [4]), 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.

We have updated the Introduction section to discuss SCENIC+ as an alternative approach that generates peak-gene pairs. The evaluation results presented above have been added to Supplementary Figures 3-4, 7-8 and discussed in the Results section. Other discussions have been added to Supplementary Discussion.

5. "*Given the severe sparsity of scATAC-seq data, it would be interesting to simulate this effect to assess the impact on the power of scMultiMap. This could be easily achieved for the PBMC datasets.*"

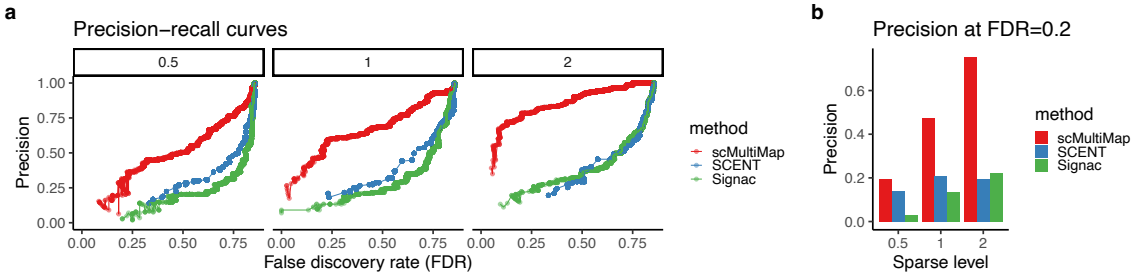

Figure R6: Power of scMultiMap across sparsity levels of scATAC-seq data. Different sparsity levels were simulated by setting the simulation sequencing depths to be the sequencing depths observed on CD14 monocytes from PBMC data [17] multiplied by 0.5, 1, and 2, such that the simulated peaks have a median sparsity of 0.78, 0.62 and 0.42, respectively.

Response: Per your suggestion, we conducted a new set of simulations with varying scATAC-seq sparsity. We leveraged the observed sequencing depths on CD14 monocytes from the PBMC data [17] and the same set of peaks and genes as in Figure 1b. We then simulated scATAC-seq data with different sparsity levels by varying the simulation sequencing depths to be either 0.5 or 2 times the observed values on real data, such that the median sparsity across peaks are 0.78 and 0.42, respectively, while the median sparsity is 0.62 at the observed sequencing depths. We then evaluated scMultiMap, Signac and SCENT on these three sets of simulated peak counts with varying sparsity. The power of scMultiMap consistently outperforms the other two methods at different sparsity levels as shown by the overall precision-recall curves (Figure R6a) and the precision at FDR=0.2 (Figure R6b). Moreover, the improvement by scMultiMap over the other two methods is greater when the data are less sparse. The power of all methods increases as the data become less sparse, with the exception of SCENT, which binarizes peak counts and benefits less from larger nonzero counts.

We have incorporated these results into the Results section 'scMultiMap has better detection accuracy and computational efficiency'.

6. "*It is odd that the results in Figure 2B are presented as only a count and not a percentage of correctly identified peak-gene pairs from the total of each capture type. For the cell-type specific pairs, the figure legend and ideally figure should identify that*

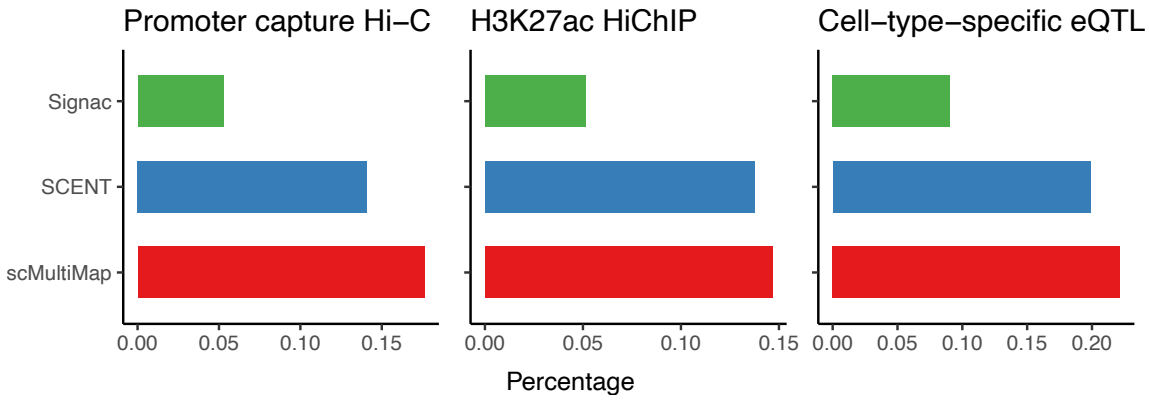

Figure R7: Consistency of scMultiMap findings with orthogonal data modalities in percentages. The rest of the figure legend is the same as Figure 2b.

this result was derived for monocytes."

Response: Per your suggestion, we also calculated the percentage of correctly identified peak-gene pairs among total number of pairs of each capture type for three methods. Figure R7 shows the consistency with alternative methods presented by percentages. As the total number of pairs of each capture type is the same for all three peak-gene association methods considered, the comparison among methods remain the same as Figure 2b. The percentage values are moderate, likely due to the extreme sparsity and the limited number of cells in the single-cell multimodal data analyzed in this manuscript (8,484 CD14 monocytes). With more consortium efforts ongoing that generates large-scale single-cell multimodal data, we expect that scMultiMap will be a powerful tool to better and more comprehensively define enhancer-gene pairs on larger datasets.

We have incorporated this figure as Supplementary Figure 6 and discussed it in the Results section 'scMultiMap has higher reproducibility across independent datasets'. We have also updated the legend of Figure 2b to clarify that the results were derived for monocytes.

7. "*To ensure that real enhancers are indeed identified in the brain it would be great to see the overlap with ENCODE predicted enhancers across fetal and adult human tissue via chromHMM (10.1038/nprot.2017.124).*"

Response: Thank you for this excellent suggestion. To enhance the validity of our results, we performed an annotation search at the ENCODE portal with the following filters: keyword = brain, annotation type = chromatin state, Encyclopedia version = current and v4, analysis = chromHMM. Among 68 results from the search, we focus on annotations from dorsolateral prefrontal cortex tissues and from aged donors to match with the brain region and the donor age of the single-cell multimodal data analyzed in the manuscript [4]. This results in two sets of annotation files:

- Annotation file set: ChromHMM 18-state model of BSS01272: dorsolateral prefrontal cortex (previously identified as middle frontal area 46) male adult (81 year) from donor(s) ENCDO980BZD. URL: `https://doi.org/doi:10.17989%2FENCSR675ELD`

- Annotation file set: ChromHMM 18-state model of BSS01271: dorsolateral prefrontal cortex (previously identified as middle frontal area 46) female adult (75 years) from donor(s) ENCDO871IWW. URL: `https://doi.org/doi:10.17989%2FENCSR064TJH`

These two annotation files are based on 1 male adult aged 81 years old (ENCDO980BZD) and 1 female adult aged 75 years old (ENCDO871IWW). 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]. Figure R8 shows that across three major brain cell types, the peaks from peak-gene pairs identified by scMultiMap are significantly enriched for annotated enhancers from these two samples, with odds ratio greater than 1. The only exception is the overlap of findings on excitatory neurons with ENCDO871IWW.

There are two caveats to the interpretation of these results. 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 Figure R8 may not be the most appropriate for comparing the performance of the three methods.

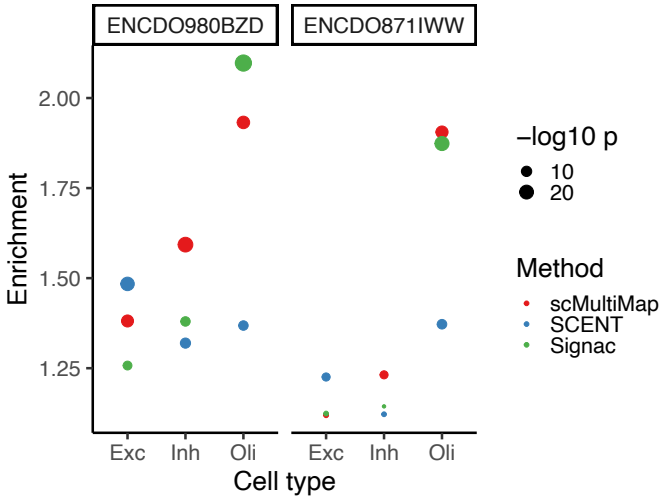

Figure R8: Consistency of peaks from significant pairs (BH-adjusted p-value< 0.2) with annotated enhancers in dorsolateral prefrontal cortex tissues from two aged donors (ENCDO980BZD and ENCDO871IWW) in the ENCODE consortium [20] across three major brain cell types: excitatory neurons (Exc), inhibitory 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.

We have incorporated the above results into Supplementary Figure 11 and Results section 'scMultiMap identified biologically relevant gene regulatory mechanisms in brain cells'.

8. "*In the AD GWAS variant analysis, it would be interesting to see whether the associated genes have already been largely identified by single cell differential expression analysis across different cell types (10.1038/s41586-024-07606-7). This would lead further evidence that scMultiMap is capable of identification of real enhancer gene associations.* "

Response: This is a great suggestion. In this analysis, we focus on microglia—the cell type most enriched for AD GWAS signals—and genes involved in differentially co-expressed gene pairs, which highlight genes with dysregulated associations to neighboring peaks in AD compared to controls. We compared these genes with differential expression results from two papers: [21] (i.e. 10.1038/s41586-024-07606-7) and [4].

In both datasets, the overlaps are highly significant in microglia ([21]: OR=1.58, p-value=$2.55 \cdot 10^{-6}$; [4]: OR=1.51, p-value=0.01; OR denotes odds ratio).

We have incorporated these results to the Results section 'scMultiMap mapped GWAS variants of Alzheimer's disease to target genes in microglia'.

9. "*The associated R – package should be updated to include more extensive documentation and vignettes.*"

Response: Thank you for your suggestion. We have improved the documentation of our R package and added the following new analyses to demonstrate the functionality of scMultiMap:

1. Validating inferred peak-gene pairs with orthogonal data modalities: please visit Introduction to scMultiMap
2. Identifying differentially associated peak-gene pairs in disease versus control groups: please visit scMultiMap for disease-control studies
3. Using scMultiMap to map the regulatory functions of GWAS variants in disease-related cell types: please visit scMultiMap for integrative analysis with GWAS results

Please find the updated GitHub repo at `https://github.com/ChangSuBiostats/scMultiMap` and the updated vignettes at `https://changsubiostats.github.io/scMultiMap/`.

Additionally, we have organized and uploaded all code necessary to reproduce the analytical results in the paper to the GitHub repository: `https://github.com/ChangSuBiostats/scMultiMap_analysis`.

References

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

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

[3] Wang, Y., Hicks, S. C. & Hansen, K. D. Addressing the mean-correlation relationship in co-expression analysis. *PLoS computational biology* **18**, e1009954 (2022).

[4] 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).

[5] Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature genetics* **53**, 770–777 (2021).

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

[7] Lause, J., Berens, P. & Kobak, D. Analytic pearson residuals for normalization of single-cell rna-seq umi data. *Genome biology* **22**, 1–20 (2021).

[8] Miao, Z. & Kim, J. Uniform quantification of single-nucleus atac-seq data with paired-insertion counting (pic) and a model-based insertion rate estimator. *Nature Methods* **21**, 32–36 (2024).

[9] Martens, L. D., Fischer, D. S., Yépez, V. A., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell atac-seq analysis. *Nature Methods* **21**, 28–31 (2024).

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

[11] Li, Z., Nagai, J. S., Kuppe, C., Kramann, R. & Costa, I. G. scmega: single-cell multi-omic enhancer-based gene regulatory network inference. *Bioinformatics Advances* **3**, vbad003 (2023).

[12] Granja, J. M. *et al.* Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics* **53**, 403–411 (2021).

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

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

[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] 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.

[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).

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

[21] Mathys, H. *et al.* Single-cell multiregion dissection of alzheimer's disease. *Nature* **632**, 858–868 (2024).