nature portfolio

Peer Review File

Best practices for differential accessibility analysis in single-cell epigenomics

Open Access This file is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. In the cases where the authors are anonymous, such as is the case for the reports of anonymous peer reviewers, author attribution should be to 'Anonymous Referee' followed by a clear attribution to the source work. The images or other third party material in this file are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit http://creativecommons.org/licenses/by/4.0/.

Editorial Note: This manuscript was submitted to *Nature Communications* as a Registered Report. The rebuttals for pre-study and post-study are indicated within this peer review file.

Review Study Plan (Stage 1)

Reviewer #1 (Remarks to the Author):

Here, the authors propose a systematic benchmark of differential accessibility testing methods for single-cell ATAC-seq. They show two key pieces of preliminary data that motivates the study: 1) a wide range of methods are being used in different scATAC-seq studies, there appears to be little convergence in the field; 2) an analysis of publicly available scATAC-seq data demonstrates a high number of false positive hits when using some of the most common testing methods (binomial, LR, Fisher exact, Wilcoxon). The authors have presented a thorough examination of the literature that shows scATAC-seq experiments are becoming increasingly common. This provides strong support for the significance of the proposed study.

The authors propose a series of experiments that focus on using real data to comprehensively benchmark the performance of different DA testing methods. The major goals are: 1) assess concordance between bulk and single-cell ATAC DA regions; 2) assess concordance between gene expression and DNA accessibility; 3) assess false discovery rates for DA methods; 4) assess biases such as peak length; 5) assess whether scATAC-seq should be binarized for DA analysis; 6) benchmark computational requirements and scalability of different methods.

These experiments are well designed and I agree with the approach of using real data wherever possible, rather than relying on simulations. There are three aspects that I would like to see further explored:

- Sensitivity analysis: one important question in the field is how many reads and/or cells are required to perform the DA test. Another experiment could be added to perform downsampling of the scATAC-seq data (counts per cell and/or total number of cells) and test how the concordance with bulk ATAC and/or gene expression changes.
- Measures of effect size and the relationship between P-value and effect size. For many cases where a significant P-value is returned by a test, the effect size may be very small. In practice these DA results can easily be filtered out by applying a filter based on effect size (eg, log2 fold change). It would be useful to see, for example, the number of false positive DA results that remain after applying different fold-change filters. This analysis could be added to several of the proposed experiments (eg, stratify results by effect size filtering levels: no filtering, 1-fold change, 2-fold change, etc).
- Handling of replicate and batch information. Some of the DA methods have the capacity to handle replicate information, which should help in identifying accurate DA regions in the presence of batch variation. For example, the LR method can include covariates. This would be particularly useful to explore for the false positive analysis (experiment 3).

Other comments:
- What data goes into the test. This is touched on in the discussion of binarization, but more detail is needed in the methods used for other sections. Will counts or normalized values be used, and if not counts what transformation will be applied to the data?
- For the comparison between gene expression and ATAC: how will bidirectional promoters be handled?
- For the comparison between gene expression and ATAC: lowly expressed genes may be difficult to detect as DE, but may be easier to detect in DA due to the more similar dynamic range across different peaks. This could result in some false positive DA peaks appearing, due to inability to accurately detect the change in gene expression. Focusing analysis on highly expressed genes may be better
- For DA tests, a one-vs-all-other comparison can be performed to avoid running an extremely large number of pairwise tests
- For computational resource benchmarking: different software implementations can also impact RAM and runtime, aside from the statistical test. Since different software is being used for different tests (Signac, SnapATAC, Libra) there would be some confounding between the software and test used.
- More information is needed on how the results from all experiments will be summarized into a set of recommendations.

Reviewer #2 (Remarks to the Author):

The registered report has as aim the analysis of differential accessibility analysis in single cell chromatin data. This is a relevant topic and there is clearly no consensus in the literature on how to deal with this issue. Also, most DA methods have been implemented from transcriptomics data, which is potentially problematic due to the distinct distribution characteristics of DNA (accessibility) and RNA. While the proposed report is mostly technically sound and the authors did a good job in reviewing recent single cell ATAC-seq literature, it is very similar study by the authors for single cell RNA-seq. Particularly problematic is the fact authors mostly ignore ATAC-seq specific issues and a large literature of bulk ATAC-seq (and related chromatin protocols) on their analysis (see points below). This is possibly a reflection the lack of practical experience of authors on the analysis or method developmentof chromatin data.

1) A major distinction of ATAC in contrast to RNA is the fact it is not clear a priori where genomic features (peaks) are found. Indeed, the problem of how to represent scATAC-seq data is still an open topic (gene-based representation, peak based, or window based; see archR tool or Chen et al, Genome Biology, 2019). Currently, the study only considers peak based matrices. Moreover, there is a literature of how to do peak calling in ATAC-seq, which is not addressed here (see the blog post below). This also include considering quality aspects of ATAC-seq such as fragment insert size. (see https://bigmonty12.github.io/peak-calling-benchmark for an interesting review). Altogether, the question of how these decisions affect DA analysis is a relevant one.

2) There is also a long-standing literature on the closely related differential peak calling problem for bulk ATAC and ChiP-seq. This is ignored by the authors. There, several discussions are of

potential interest to this work.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02686-y
https://academic.oup.com/bib/article/17/6/953/2453197?login=false

This also includes a work with a similar approach of this work on bulk data.

https://www.nature.com/articles/s41598-020-66998-4

3) Further of example a specific artifact not considered here is CG bias. This is also likely to affect promoters and enhancers regions differently.

https://pubmed.ncbi.nlm.nih.gov/36452861/

4) It is unclear what DA problem is addressed altogether. Finding cell specific DARs or differential test (comparing accessibility on a cell between two biological conditions). These are two distinct problems as cell specific DAs are likely to have clearer differences than "comparative" Das.

5) A major issue for studies benchmarking accessible regions is the lack of gold standards. The authors proposed using of parallel data (bulk atac or multiome data) to address this. I would consider this rather a silver standard, as these approaches require the use of statistical methods for DA and DE analysis in the parallel data sets. This generates a circular problem; validation is based on predictions made by the same method. Random selection of methods is a way to mitigate this, but I wonder if this approach does not simply favor DA methods, which predictions are more similar to other methods. Moreover, chromatin and transcription are known to have distinct dynamics in particular in cell differentiation (https://pubmed.ncbi.nlm.nih.gov/33098772/). This would be a further limitation on the use of multiome data as silver standard.

Specific points

Authors do not include Scenic (https://scenicplus.readthedocs.io/) related analysis (topic modelling derived DARs) in their analysis.

ArchR corrects for cell specific artifacts when finding DARs (read depth and other QC measures). This should be considered. https://www.archrproject.com/bookdown/identifying-marker-peaks-with-archr-1.html

Reviewer #3 (Remarks to the Author):

Detecting differential signals in sequencing-based assays is a foundational analysis in genomics. While in bulk assays, standard analytical methods of differential analysis have emerged (e.g., DESeq2, etc), no such practices have been established for single-cell data, especially from chromatin accessibility assays (ATAC-seq). Here, Yang Teo et al. propose to

comprehensively evaluate all published methods to detect differential accessibility from single-cell experiments. To do this, they will make use of publicly available data, in which both single-cell and matched bulk-derived data is available. Overall, I think this is a timely and important endeavor, and will help the field coalesce around a best-practice.

Overall, I find the proposed experiments/analysis well designed and find no technical problems and the authors should proceed as planned.

I do have two concerns regarding the design/considerations of analysis:

1) Unlike RNA-seq (single cell or bulk), chromatin accessibility assays do not have a fixed annotation that is used to compare two experiments (GENCODE, Refseq, etc.), and peak detection is first needed to define "informative" genomic regions. They authors should consider peak detection as an important part of the single cell DA detection analysis. The number of features tested will affect the multiple test correction approaches as well as normalization (see pt. 2 below). Because chromatin accessibility in different cell types can vary drastically on top of the considerable technical variation, this could be a significant problem to overcome. In addition, how are peaks to be defined in the single-cell data -- will a pseudobulk procedure be used? Alternatively, one could use a fixed reference scaffold as proposed in Meuleman et al., 2020 Nature.

2) Although not discussed much in the literature for chromatin accessibility data, the required normalization of samples is likely quite different that the methods used for RNA-seq data. While, DESeq2 typically uses a single scale factor (geometric mean of read depth) for an entire sample to normalize RNA-seq data, this is likely not appropriate for accessibility data because of regional differences (promoters more accessible that distal elements), most elements are very weak (distribution of peak densities are exponentially decaying in a sample), and binary nature of single cell data (theoretically at most 2 read per region). Because normalization can have large impact on DA analyses, the authors should consider evaluating how different approaches affect results.

Response to reviewers

We would like to sincerely thank the reviewers for their careful review of our Registered Report, "**Best practices for differential accessibility analysis in single-cell epigenomics**," and for their thoughtful feedback. We were grateful for the opportunity to respond to the points they raised, and have revised our proposal in order to incorporate their comments. We have highlighted all the changes to the text in the accompanying manuscript, but also included the new text in the following responses for your convenience.

The most important changes to our proposal are as follows:

- We have added two new experiments suggested by reviewers to the revised proposal. In Experiment 5, we propose to test the impact of filtering regions according to log-fold change thresholds. In Experiment 7, we propose to assess the minimum data requirements for single-cell DA analysis by downsampling analyses of sequencing depth or the number of cells profiled.
- We have also added several new analyses to our previously proposed experiments in order to address points raised by the reviewers. These analyses will allow us to assess points including the effects of different peak calling strategies, of different approaches to normalization, of incorporating replicate as a covariate in regression models, and of correcting for technical artefacts such as GC bias, among others.
- In the "Conclusions" section, we now elaborate on how we propose to quantitatively summarize the results of all the proposed experiments in order to formulate a data-driven set of recommendations for single-cell DA analysis.
- We have added a "Limitations" section that we are committed to retaining in the published report in which we acknowledge several of the limitations of our framework.

Throughout this document, reviewer comments are shown in **blue**, with our own response in **black**, and changes to manuscript text or figure captions in **red**.

Reviewer #1 (Remarks to the Author):

Here, the authors propose a systematic benchmark of differential accessibility testing methods for single-cell ATAC-seq. They show two key pieces of preliminary data that motivates the study: 1) a wide range of methods are being used in different scATAC-seq studies, there appears to be little convergence in the field; 2) an analysis of publicly available scATAC-seq data demonstrates a high number of false positive hits when using some of the most common testing methods (binomial, LR, Fisher exact, Wilcoxon). The authors have presented a thorough examination of the literature that shows scATAC-seq experiments are becoming increasingly common. This provides strong support for the significance of the proposed study.

The authors propose a series of experiments that focus on using real data to comprehensively benchmark the performance of different DA testing methods. The major goals are: 1) assess concordance between bulk and single-cell ATAC DA regions; 2) assess concordance between gene expression and DNA accessibility; 3) assess false discovery rates for DA methods; 4) assess biases such as peak length; 5) assess whether scATAC-seq should be binarized for DA analysis; 6) benchmark computational requirements and scalability of different methods.

These experiments are well designed and I agree with the approach of using real data wherever possible, rather than relying on simulations. There are three aspects that I would like to see further explored:

Thank you for your careful review of our manuscript and your encouraging comments. Below, we outline how we propose to address each of your suggestions to improve our analysis.

- Sensitivity analysis: one important question in the field is how many reads and/or cells are required to perform the DA test. Another experiment could be added to perform downsampling of the scATAC-seq data (counts per cell and/or total number of cells) and test how the concordance with bulk ATAC and/or gene expression changes.

This is an excellent suggestion. To address this comment, we propose two new analyses. First, we will take advantage of the relatively high sequencing depth of the datasets used in Experiment 1 (concordance to bulk ATAC-seq) to carry out a downsampling analysis of the counts per cell, as suggested. This will allow us to address the relationship between sequencing depth and the accuracy of DA analysis. Second, we will take advantage of the large numbers of cells in the datasets used in Experiment 2 (concordance within multiome data) to carry out a downsampling analysis of the numbers of cells per dataset, also as suggested. This will allow us to address the relationship between the number of cells and the accuracy of DA analysis.

We have incorporated these new analyses as a new experiment in the revised proposal, Experiment 7: Data requirements for accurate DA analysis. The text of this new section is reproduced in full below.

> Experiment 7: Data requirements for single-cell DA analysis
> *Rationale:* The scale of scATAC-seq datasets is increasing exponentially (**Fig. 1b**). On one hand, the availability of hundreds of thousands—or even millions—of cells per dataset could increase the statistical power of DA analysis. On the other hand, past work in single-cell transcriptomics[77] reported that the number of cell types reported per study was closely linked to the total number of individual cells sequenced in that study. This observation suggests that the increased statistical power afforded by a greater number of cells could be offset by an increased granularity of DA comparisons, each of which might leverage proportionally fewer cells. Moreover, many of the largest datasets are extremely sparse: for instance, the single-cell map of chromatin accessibility across 30 tissues reported by Zhang et al.[26] had a median of just ~2,800 fragments per cell. These trends raise the question of what minimum sequencing depth or number of cells are required for accurate DA analysis. In other words, what are the minimum data requirements for single-cell DA analysis? We will address this question through downsampling analyses of the sequencing depth or number of cells per dataset.
>
> *Experiments and datasets:* The datasets in Experiment 1 were collected primarily using plate-based techniques, and therefore provide an excellent platform for downsampling the total number of fragments per cell. Accordingly, in Experiment 7a, we will downsample each dataset to a mean of 500, 1000, 2000, 5000, or 10,000 fragment counts per cell, using the 'downsampleMatrix' function from the 'scuttle' R package to perform downsampling on the entire matrix rather than on a per-cell basis; this will maintain the natural variation in depth from one cell to another that is characteristic of all single-cell datasets. We will then repeat the AUCC analysis for each dataset and visualize the dependency between sequencing depth (number of counts per cell) and biological accuracy (AUCC). The datasets will be the same as those used in Experiment 1.
>
> The multiome datasets in Experiment 2 were collected by droplet-based technology, providing a much larger number of cells per dataset, and dozens of pairwise comparisons are supported by at least 1,000 cells in each group. These comparisons provide a platform to downsample the number of cells per dataset. Accordingly, in Experiment 7b, we will downsample each comparison to consider only 20, 50, 100, 200, 500, or 1,000 cells per group. We will again repeat the AUCC analysis for each dataset and visualize the dependency between the number of cells per dataset and the AUCC. The datasets used in this analysis will comprise the subset of comparisons from Experiment 2 that entail comparisons of at least 1,000 cells per group.

Data analysis plan: The primary outcomes of this experiment will be the AUCC for each DA method at each level of fragment or cell downsampling. We will then characterize the dependency of the AUCC on the sequencing depth (Experiment 7a) or number of cells (Experiment 7b).

In summary, the list of proposed analyses is as follows:

1. Impact of downsampling fragments on the concordance between single-cell and bulk DA, as quantified by the AUCC
2. Impact of downsampling cells on the concordance between single-cell DA and DE, as quantified by the AUCC

- Measures of effect size and the relationship between P-value and effect size. For many cases where a significant P-value is returned by a test, the effect size may be very small. In practice these DA results can easily be filtered out by applying a filter based on effect size (eg, log2 fold change). It would be useful to see, for example, the number of false positive DA results that remain after applying different fold-change filters. This analysis could be added to several of the proposed experiments (eg, stratify results by effect size filtering levels: no filtering, 1-fold change, 2-fold change, etc).

Thank you for this suggestion. We agree that assessing the effect of filtering DA results by effect size would be an interesting analysis. We also agree that it would be important to extend this analysis throughout multiple experiments, and not just Experiment 3 (false discoveries): while the number of false discoveries will obviously be reduced by increasingly stringent log-fold change filters, this will expectantly come at the expense of a decrease in the number of true-positives. Therefore, understanding how log-fold change filtering affects the biological accuracy of the results, as captured by Experiments 1 and 2 (concordance to bulk ATAC-seq and multiome data), is equally critical. We will therefore address this comment by assessing the impact of log-fold change filtering on Experiments 1, 2, and 3 (concordance to bulk ATAC-seq, concordance to multiome, and false discoveries, respectively).

In carrying out these analyses, we can foresee three challenges that we think are worth noting here:

1. One challenge is that different DA methods often implement different approaches to the calculation of log-fold change differently; most notably, the log-fold change computed by pseudobulk methods is inherently different from the log-fold change computed by single-cell methods. To address this issue and allow a fair comparison between methods, we propose to standardize log-fold change calculation for all DA methods using the method implemented in Signac.
2. A second challenge is that the datasets we propose to analyze involve biological comparisons with markedly different biological effect sizes, which will expectantly lead to different distributions of log-fold changes from one dataset to another. We propose to address this by filtering log-fold changes whose absolute value is below a given quantile in each dataset (e.g., removing the peaks in the bottom 20% of log-fold changes from consideration). This approach does have the inherent limitation that it will not allow us to suggest a specific log-fold change threshold; given that log-fold changes will vary across datasets depending on the underlying biology, we think this is appropriate. This approach will, however, allow us to comment on whether some degree of filtering by log-fold change does generally have a positive effect.
3. A third challenge is that the outcomes we are proposing to calculate (AUCC and number of false discoveries, respectively) are both sensitive to the absolute number of peaks being tested. This introduces a potential confounding factor, in that simply filtering peaks at random would also be expected to increase the AUCC and decrease the number of false discoveries. To account for this effect, for each log-fold change threshold, we will test the effect of removing an equal number of peaks from the dataset at random. We will then compute the difference in the AUCC or number of false discoveries (respectively) when filtering by log-fold change versus at random. If filtering by log-fold change has a positive effect, we will expect that the AUCC will be higher (and/or the number of false discoveries lower) than when removing an equivalent number of random peaks.

We have incorporated the above analyses into the revised proposal as a new experiment in the revised proposal, Experiment 5: Impact of log-fold change filtering. The text of this new section is reproduced in full below.

> Experiment 5: Impact of log-fold change filtering
> *Rationale:* In addition to the statistical significance of DA estimated by any given method, the difference in the number of reads overlapping a particular genomic region can be quantified by a measure of effect size, most commonly the log-fold change. Differential analyses of ATAC-seq data may discard differentially accessible regions with a log-fold change below an arbitrary threshold, on the grounds that regions with small fold changes are unlikely to be biologically relevant[44]. Similar practices are ubiquitous in the analysis of other sequencing-based technologies, such as RNA-seq, where the choice of log-fold change threshold has been reported to alter the biological interpretation of a given experiment[81]. These observations motivate an empirical assessment of the impact of log-fold change filtering on single-cell DA analysis. Specifically, we will repeat several of the analyses described above after imposing more or less stringent thresholds on the minimum log-fold change, discarding regions without such a difference.
>
> *Experiments and datasets:* In Experiments 5a and 5b, we will evaluate the effect of imposing more or less stringent log-fold change cutoffs on the biological accuracy of DA analysis. To do this, we will make use of the matching bulk ATAC-seq and multiome datasets employed in Experiments 1 and 2, respectively. In Experiment 5a, we will repeat our comparison of the AUCC between single-cell and bulk ATAC-seq data after filtering peaks according to their deciles of log-fold change in the single-cell data, removing the bottom 10% to 90% of peaks with the lowest log-fold change in each dataset. In Experiment 5b, we will repeat our comparison of the AUCC between the ATAC and RNA modalities of single-cell multiome data after filtering genes according to their deciles of log-fold change in the single-cell data, again removing the bottom 10% to 90% of genes with the lowest log-fold change in each dataset.
>
> In parallel, we will also evaluate the effect of imposing more or less stringent log-fold change cutoffs on the number of false discoveries in single-cell DA analysis. In Experiment 5c, we will repeat our null comparisons of random replicates from published scATAC-seq datasets. We will then compare the number of false discoveries returned by each single-cell DA method at each level of log-fold change filtering with its biological accuracy at the same threshold.
>
> The outcomes discussed above (AUCC and number of false discoveries, respectively) are both sensitive to the absolute number of peaks being tested. This introduces a potential confounding factor, in that simply filtering peaks at random would also be expected to increase the AUCC and decrease the number of false discoveries. To account for this effect, for each log-fold change threshold, we will test the effect of removing an equal number of peaks from the dataset at random. We will then compute the difference in the AUCC or number of false discoveries (respectively) when filtering by log-fold change versus at random.
>
> *Data analysis plan:* Datasets will be preprocessed or generated as described above for Experiments 1, 2, 3, and 4. Concordance will be measured as described above, after filtering the results from the previous analyses to remove the bottom one to nine deciles of peaks or genes with the lowest absolute log-fold changes. Our proposal to remove peaks and genes based on relative rather than absolute thresholds is based on the consideration that the datasets we propose to analyze involve biological comparisons with markedly different biological effect sizes. This will expectantly lead to different distributions of log-fold changes from one dataset to another. This approach does have the inherent limitation that it will not allow us to suggest a specific log-fold change threshold; however, given that log-fold changes will vary across datasets depending on the underlying biology, we believe that this is appropriate. Moreover, log-fold change estimates differ across software implementations, most notably between DA methods that operate on individual cells versus on so-called pseudobulks. For this analysis, we will standardize log-fold change calculation for all DA methods using the code implemented in Signac.
>
> In summary, the list of proposed analyses is as follows:
>
> 1. Impact of log-fold change filtering on concordance between single-cell and bulk DA
> 2. Impact of log-fold change filtering on concordance between DA and DE
> 3. Impact of log-fold change filtering on false discoveries in null comparisons of published scATAC-seq data

- Handling of replicate and batch information. Some of the DA methods have the capacity to handle replicate information, which should help in identifying accurate DA regions in the presence of batch

variation. For example, the LR method can include covariates. This would be particularly useful to explore for the false positive analysis (experiment 3).

This is an excellent suggestion. To address it, we will add two additional analyses to Experiment 3, as requested. First, we will explore the impact of adding replicate (batch) as a covariate to $LR_{clusters}$ and negative binomial regression, using the 'latent.vars' argument implemented in Signac. Second, because it may be more appropriate to consider replicate as a random effect rather than a fixed effect, we will also explore the impact of fitting negative binomial generalized linear mixed models (GLMMs), using the fast approximations in the recently described NEBULA package (He et al., *Commun. Biol.* 2022, doi: 10.1038/s42003-021-02146-6) that now make these computationally tractable even for scATAC-seq datasets with $>10^4$ cells and $>10^5$ peaks. We incorporated these new analyses into the Data analysis plan as shown below:

> We previously showed that, in DE analysis of single-cell RNA-seq data, accounting for variation between biological replicates is required to achieve control of the false discovery rate. Because some of the single-cell DA methods analyzed here are capable of accounting for this effect, we will perform additional comparisons whereby we will incorporate the effect of biological replicate into the underlying model. Specifically, we will explore the impact of adding replicate as a covariate to $LR_{clusters}$ and negative binomial regression, using the 'latent.vars' argument implemented in Signac. Moreover, because it may be more appropriate to consider replicate as a random effect rather than a fixed effect, we will also explore the impact of fitting negative binomial generalized linear mixed models (GLMMs), using the fast approximations in the NEBULA package[74].

Other comments:
- What data goes into the test. This is touched on in the discussion of binarization, but more detail is needed in the methods used for other sections. Will counts or normalized values be used, and if not counts what transformation will be applied to the data?

The majority of the DA methods that we propose to analyze, including LR_{peaks}, negative binomial regression, Fisher's exact test, binomial test, permutation test, SnapATAC::findDAR, DESeq2, and edgeR, operate directly on the count matrix. Three methods, including the t-test, Wilcoxon rank-sum test, and $LR_{clusters}$, can operate directly on the count matrix but expectantly require normalization to produce the most accurate results. In our original proposal, we had proposed to use the implementation of these tests in Signac, which by default scales counts to 10,000 per cell and log-transforms the scaled values (log-TP10K normalization).

Your comment, as well as one from reviewer 3, spurred us to consider the effects of alternative normalization approaches on the accuracy of these three methods. Accordingly, we reviewed the approaches to normalization implemented by published scATAC-seq analysis packages (**Fig. 1b**). Signac additionally provides methods for TF-IDF normalization as well as a TP10K normalization without log-transformation ("relative counts"). ArchR implements a subsampling procedure to match cells in the two groups being compared based on read depth, then further scales counts in each cell to the median number of reads in TSSs across cells by default (MarkerFeatures.R, lines 207-218). EpiScanpy implements a similar normalization strategy to the relative counts method in Signac by default, except that instead of scaling counts to a fixed depth (e.g., 10,000 counts per cell), the median of total counts across all cells is used as the scaling factor, analogously to ArchR. Log-transformation can then optionally be applied afterwards. The remaining packages (SnapATAC, scABC, MAESTRO, Scasat, scATAC-pro, Destin, chromSCape) either perform DA analysis directly on the count matrix or implement one of the approaches to normalization described above.

Based on these observations, we now propose to test the impact of the following additional normalization strategies on the t-test, Wilcoxon rank-sum test, and $LR_{clusters}$:

1. Raw counts

2. Scaling counts to a total of 10,000 per cell without log-transformation, using the “relative counts” implementation in Signac
3. Scaling counts to the median number of counts per cell without log-transformation
4. Scaling counts to the median number of counts per cell with log-transformation
5. TF-IDF normalization
6. Last, in response to a comment from reviewer 2, we will also consider the effect of accounting for GC bias in the normalization procedure using a method recently developed for bulk ATAC-seq analysis (‘smooth GC-full-quantile’; van den Berge et al., *Cell Rep. Methods* 2022).

We have incorporated these new analyses as a complement to the binarization analyses in Experiment 6 (previously 5). The text added to this section is reproduced below:

> Another important question, particularly when scATAC-seq data are not binarized, is whether and how the data should be normalized before DA analysis. The majority of the DA methods that we propose to analyze (including LR_{peaks}, negative binomial regression, Fisher’s exact test, binomial test, permutation test, SnapATAC::findDAR, DESeq2, and edgeR) operate directly on either counts or binarized data as input. However, the remaining three methods (t-test, Wilcoxon rank-sum test, and $LR_{clusters}$) expectantly require normalization to produce the most biologically accurate results. We therefore propose to evaluate the impact of the most widely used normalization strategies for scATAC-seq data on these three DA methods by again repeating several of the analyses described above.
>
> To identify the most widely used methods for normalizing scATAC-seq data for DA analysis, we reviewed the approaches implemented by published scATAC-seq analysis packages[33–43]. Based on this review, we propose to complement our analyses of the log-counts per 10,000 normalization implemented by default in Signac with additional DA analyses using the following additional normalization strategies:
>
> 1. Raw counts
> 2. Scaling counts to a total of 10,000 per cell without log-transformation, using the “relative counts” implementation in Signac
> 3. Scaling counts to the median number of counts per cell[33,36] without log-transformation
> 4. Scaling counts to the median number of counts per cell with log-transformation
> 5. TF-IDF normalization
>
> Finally, some normalization methods that have been applied to bulk ATAC-seq data[81] explicitly seek to regress out technical properties of the underlying sequencing data, notably GC bias. To evaluate the impact of these approaches, we will also consider the effect of accounting for GC bias in the normalization procedure using a method recently developed for bulk ATAC-seq data[81] (‘smooth GC-full-quantile’).
>
> *Experiments and datasets:* [...] In Experiments 6d, 6e, and 6f, we will repeat the analyses of Experiments 6a-c but instead evaluate the effects of normalization. We will analyze the concordance to bulk ATAC-seq data, the number of false discoveries, and the biases of these three DA methods using the six normalization strategies described above. Software implementations from the Signac package will be used for all normalization strategies except smooth GC-full-quantile, for which the implementation in the qsmooth package will be used. [...]
>
> *Data analysis plan:* Datasets will be preprocessed or generated as described above for Experiments 1, 2, 3, and 4. For our analysis of binarization (Experiments 6a-c), several of the DA methods we propose to evaluate already operate on binarized data, precluding a comparison between qualitative and quantitative representations. These methods include Fisher’s exact test, the binomial test, the permutation test, and logistic regression using binary peak accessibility as the independent variable (LR_{peaks}), and will be excluded from the experiments proposed in this section. For the remaining seven DA methods, DA analysis will be performed as described above, but peak count matrices will be binarized prior to analysis. Conversely, our evaluation of normalization approaches will consider only the t-test, Wilcoxon rank-sum test, and $LR_{clusters}$. For our analysis of the QC matching procedure implemented in ArchR (Experiment 6g), we will repeat Experiment 1, but here perform DA analysis on equal numbers of cells from each condition matched by their TSS enrichment scores and $\log_{10}$(# of fragments), adapting code from ArchR for this purpose. Cells will be sampled from the ‘background’ (that is, the condition with more cells) in order to match the number of cells in the ‘foreground’ (that is, the condition with fewer cells).

- For the comparison between gene expression and ATAC: how will bidirectional promoters be handled?

Thank you for raising this point. In our original submission, we had not proposed to take into account the potential for overlap between adjacent promoters, which has the potential to confound our analysis when reads overlapping a TSS cannot be unambiguously assigned to a single gene. To address this point, we will perform an additional sensitivity analysis in Experiment 2, wherein we remove all genes with overlapping windows around their TSSs. We added this point to the Data analysis plan of Experiment 2:

> *Data analysis plan:* For the scATAC-seq data, FASTQ files have already been downloaded from GEO and demultiplexed using SnapATAC to confirm their availability for the proposed analyses, except for the Luecken *et al.* dataset, for which only aligned BAM files are available, and our own in-house dataset, for which aligned BAM files will likewise be used. For the remaining datasets, demultiplexed FASTQ files will be trimmed with Trim Galore and aligned to the genome using bwa, as implemented within SnapATAC. Reads will be aggregated around transcription start sites (TSSs), using TSS definitions from GENCODE (human and mouse) or FlyBase (Drosophila) and a window of 10 kbp. To exclude the possibility that our results are confounded by bidirectional promoters, we will perform a sensitivity analysis whereby we exclude all genes with overlapping windows around their TSSs will be excluded from the AUCC calculation. For the scRNA-seq data, annotated count matrices have already been downloaded from GEO to confirm their availability for the proposed analyses, and will be used directly to perform DE analysis. A complete list of accessions or URLs is provided in the Data availability statement.

Separately, while considering this point we realized that the multiome dataset of Floc'hlay *et al.* that we had originally proposed to analyze is probably ill-suited for this experiment, as the short intergenic distances in fly would lead to a large number of overlapping promoter regions compared to human or mouse. We have therefore removed this dataset from the proposal.

- For the comparison between gene expression and ATAC: lowly expressed genes may be difficult to detect as DE, but may be easier to detect in DA due to the more similar dynamic range across different peaks. This could result in some false positive DA peaks appearing, due to inability to accurately detect the change in gene expression. Focusing analysis on highly expressed genes may be better

This is an interesting suggestion, and we agree that it would be of interest to exclude the possibility that our results are confounded by false-negatives among lowly-expressed genes in the scRNA-seq component of the multiome data. To address this point, we will perform an additional sensitivity analysis in Experiment 2, wherein we will repeat our concordance analysis after removing the bottom tercile of lowly-expressed genes. The text added to the Data analysis plan is reproduced below:

> Previous work has shown that inferences about DE are generally more accurate for highly expressed genes[73,74]. Conversely, identifying instances of true DE among lowly-expressed genes can be highly challenging[49]. These observations raise the possibility that inaccurate inferences about DE for lowly-expressed genes could confound our analysis. Therefore, we will perform a final sensitivity analysis for both the gene- and GO-level concordance, whereby we will re-evaluate concordance after excluding the bottom tercile of lowly-expressed genes.

- For DA tests, a one-vs-all-other comparison can be performed to avoid running an extremely large number of pairwise tests

This is a thoughtful suggestion, but we don't believe it is necessary: in Experiment 2, we had proposed to randomly select a maximum of 100 pairwise cell type comparisons per dataset, which will be computationally tractable given the resources we have access to. We do think that in general, it is preferable to run pairwise comparisons in this experiment, as this setup will significantly increase the

number of data points in our comparison of DA methods (e.g., a dataset with 20 cell types would have 20 one-vs-all comparisons but 190 pairwise comparisons).

- For computational resource benchmarking: different software implementations can also impact RAM and runtime, aside from the statistical test. Since different software is being used for different tests (Signac, SnapATAC, Libra) there would be some confounding between the software and test used.

The point that different software implementations can also impact RAM and runtime is certainly a valid one. However, we don't believe this is a confounding factor *per se* in our proposed analysis, simply because no two packages above implement the same tests: SnapATAC is the only software package that implements the SnapATAC DA method, whereas Libra overlaps with Signac only insofar as Libra provides a thin wrapper around some of the Signac implementations (for the t-test, Wilcoxon rank-sum test, negative binomial regression, and $LR_{clusters}$).

Nonetheless, we are aware of some instances in which multiple widely used software packages in single-cell epigenomics implement the exact same method, but with different implementations of the same test are reported to be markedly different in performance. A prominent example would be the Wilcoxon rank-sum test, where an implementation that claims a ~1,000-fold speed-up over the default implementation in Signac is available ('presto,' doi: 10.1101/653253) and implemented by default in ArchR. Other examples include the implementations of the t-test in ArchR vs. Signac, as well as the negative binomial regression implementations in Signac vs. glmGamPoi.

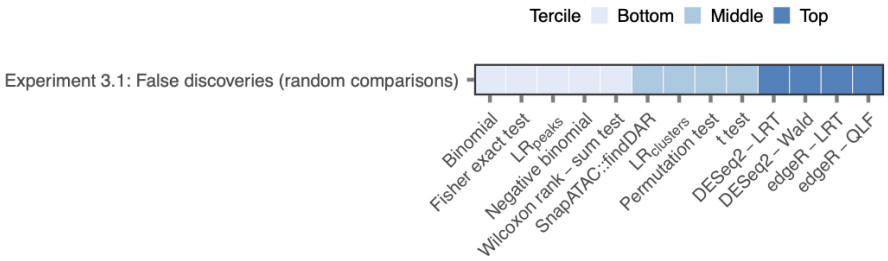
We will therefore address this point by including multiple implementations of the same test whenever impossible, as outlined above. We revised the text of the Experiment 8 (previously Experiment 6) section to clarify this as shown:

> *Experiments and datasets:* In the course of carrying out Experiments 1 and 2, we will quantify the computational resources required by each DA method for each individual dataset. Peak memory usage will be monitored using the 'peakRAM' R package. Wall time will be monitored using the base R function 'system.time'. Datasets will be the same as those described above for Experiments 1 and 2, and accessions or URLs are provided below in the Data availability statement. For some of the DA methods we propose to test, different implementations are provided in widely-used software packages, some of which have been reported to offer speed-ups of several orders of magnitude[78]. These methods include the Wilcoxon rank-sum test (Signac vs. ArchR), the t-test (Signac vs. ArchR), and negative binomial regression (Signac vs. glmGamPoi). For these three DA methods, we will test both implementations described above.

- More information is needed on how the results from all experiments will be summarized into a set of recommendations.

Thank you for the opportunity to elaborate on this important point. In reflecting on this question, we sought to summarize these results in as data-driven a manner as possible, avoiding subjective judgements. We recognized that the outcomes of Experiments 1, 2, 3 and 8 (formerly 6) all lend themselves naturally to a single, quantitative summary statistic: we can simply take the average AUCC (Experiments 1 and 2) or number of false discoveries (Experiment 3) per DA method. We further reasoned that we could summarize the biases of each DA method (as evaluated in Experiment 4) into a single summary statistic, as follows: for Experiment 4a, we propose to rank methods based on their biases (i.e., towards highly-accessible or wide peaks), and then average the rank across all three peak properties (mean count, % of cells open, and peak width). For Experiment 4b, we propose to compute the total proportion of false discoveries arising from the top decile of each property (e.g., from the top-10% widest peaks), and then average these proportions across peak properties.

Then, following the style established by seminal benchmarks in the field of single-cell genomics (Soneson and Robinson, *Nat. Methods* 2018) that has now become widely adopted, we could divide DA methods into terciles of performance and establish top-, middle-, and bottom-performing groups of methods in each experiment. To provide an example of this, we show below a single row of the proposed visualization. This row shows the pilot data presented in **Figure 2** of our submission, i.e., false discoveries in random comparisons of scATAC-seq data:



Response Fig. 1.1 | Summarization of experiments into a set of recommendations. The performance of the DA methods evaluated in **Figure 2** on random comparisons of scATAC-seq data is summarized by grouping DA methods into terciles of performance.

Finally, we will then tally the total number of experiments in which DA methods were ranked in the top-, middle-, and bottom-performing groups. DA methods that are consistently ranked in the top tercile across experiments would be the most logical to recommend, but the proposed visualization would also allow readers to select DA methods according to specific characteristics (i.e., biological accuracy vs. FDR control vs. biases).

The remaining experiments do not involve comparisons of DA methods *per se*, but rather seek to provide more general guidelines for DA analysis that are independent of the specific DA method used: e.g., the impact of binarization or various normalization strategies, of filtering by log-fold change. For these, we will summarize our recommendations by estimating the effect size of the given strategy on the outcome of interest, for example, the impact of binarization on the concordance between single-cell and bulk DA, in Experiment 6). Effect sizes will be derived from a mixed-effects model that treats DA methods as random effects and the strategy in question (e.g., binarization) as a fixed effect. We will produce a figure that visualizes all of the estimated effect sizes for these experiments.

Together, these approaches will allow us to summarize the results of the proposed experiments in a data-driven yet interpretable manner.

Reviewer #2 (Remarks to the Author):

The registered report has as aim the analysis of differential accessibility analysis in single cell chromatin data. This is a relevant topic and there is clearly no consensus in the literature on how to deal with this issue. Also, most DA methods have been implemented from transcriptomics data, which is potentially problematic due to the distinct distribution characteristics of DNA (accessibility) and RNA. While the proposed report is mostly technically sound and the authors did a good job in reviewing recent single cell ATAC-seq literature, it is very similar study by the authors for single cell RNA-seq. Particularly problematic is the fact authors mostly ignore ATAC-seq specific issues and a large literature of bulk ATAC-seq (and related chromatin protocols) on their analysis (see points below). This is possibly a reflection the lack of practical experience of authors on the analysis or method developmentof chromatin data.

Thank you for your comments highlighting the importance of this topic and the lack of consensus in the field. We appreciated the opportunity to incorporate the suggestions for improvement that you made, and expand the scope of our study to address additional questions in the analysis of scATAC-seq data.

1) A major distinction of ATAC in contrast to RNA is the fact it is not clear a priori where genomic features (peaks) are found. Indeed, the problem of how to represent scATAC-seq data is still an open topic (gene-based representation, peak based, or window based; see archR tool or Chen et al, Genome Biology, 2019). Currently, the study only considers peak based matrices. Moreover, there is a literature of how to do peak calling in ATAC-seq, which is not addressed here (see the blog post below). This also include considering quality aspects of ATAC-seq such as fragment insert size. (see https://bigmonty12.github.io/peak-calling-benchmark for an interesting review). Altogether, the question of how these decisions affect DA analysis is a relevant one.

We certainly agree that the issue of peak calling is critical to the analysis of scATAC-seq data, and in particular the biological interpretation of DA results. At the same time, our intention was not to propose a benchmark of peak-calling approaches. In our reasoning, so long as every DA method is provided with the same set of peaks as input, all DA methods should be on an ‘even playing field’. The specific set of peaks that are provided as input to the methods is therefore only relevant to a comparison of DA methods insofar as an interaction could exist between the relative performance of a given DA method and the set of genomic features provided as input. An example of such an interaction could entail a scenario where (for instance) the Wilcoxon rank-sum test only performs well with a specific set of peaks as input.

Your comment spurred us to consider the possibility that such an interaction might exist. To address this possibility, we have added another proposed sensitivity analysis to Experiment 1. This experiment requires us to use an identical set of peak definitions between the single-cell and matching bulk ATAC-seq data in order to quantify the concordance of DA between the two. In the original submission, we proposed to take advantage of the availability of bulk ATAC-seq by calling peaks in the bulk data, and using these peaks as a reference to assemble peak count matrices in the single-cell data. In the revised submission, we additionally propose to test another peak-calling strategy that is more representative of how single-cell datasets are generally analyzed: namely, using the single-cell data only. Specifically, we will call peaks using MACS2 on pseudobulks from the single-cell data, mirroring the approach used in Signac, ArchR, and SnapATAC. We will present the resulting concordance results and then formally test for an interaction term between DA method performance and the peak set used as input.

We have revised the Data analysis plan for Experiment 1 to incorporate this new analysis as shown below:

Measuring concordance between DA analyses of single-cell and bulk ATAC-seq requires a matching set of peaks to be defined in both datasets. In the analyses described above, we have proposed to obtain such a peak set by calling peaks in the matched bulk ATAC-seq data, under the hypothesis that analyzing bulk data would yield a higher-quality set of peak definitions. However, this procedure is at odds with the design of most scATAC-seq studies, which instead call peaks in 'pseudobulks' of single-cell data; this is the strategy implemented in widely used software packages such as Signac, ArchR, and SnapATAC. This divergence raises the possibility that the relative performance of single-cell DA methods may depend on the definitions of the peaks provided as input. To evaluate this possibility, we will perform a final sensitivity analysis in which we repeat the calculation of concordance using peaks called from 'pseudobulks' of single-cell data, adapting code from Signac. In addition to presenting the resulting concordance results, we will then use a linear model to formally test for an interaction term between DA method performance and the peak set used as input.

In summary, the list of proposed analyses is as follows:

1. Concordance between single-cell and bulk DA, as quantified by the AUCC
   a. Sensitivity analysis: impact of removing bulk DA methods
   b. Sensitivity analysis: impact of the parameter *k*
   c. Sensitivity analysis: impact of filtering
   d. Sensitivity analysis: impact of peak definitions

Additionally, we respond to two other points raised by this comment. First, the reviewer points out the need to perform quality control of scATAC-seq data. We agree, but feel this is not relevant to our proposed experiments as we are using datasets that have already been QC'd (in other words, we will re-process the reads but match the resulting single-cell profiles to barcodes and cell type labels provided by the authors of the original studies). Second, we want to point out that it is not correct to say we are only considering peak-based matrices: our proposed Experiment 2 leverages single-cell multiome data to directly compare gene-based representations with differential RNA expression in the same cells. The use of window-based matrices (also called tile matrices), while ubiquitous, does not naturally lend itself to comparison of DA methods, as it is not clear how tile-level DA would be interpreted biologically, and having reviewed the literature extensively, we are unaware of an instance where such an analysis has been attempted.

2) There is also a long-standing literature on the closely related differential peak calling problem for bulk ATAC and ChiP-seq. This is ignored by the authors. There, several discussions are of potential interest to this work.

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02686-y
https://academic.oup.com/bib/article/17/6/953/2453197?login=false

This also includes a work with a similar approach of this work on bulk data.

https://www.nature.com/articles/s41598-020-66998-4

Thank you for highlighting these valuable references. We have revised the Introduction of our manuscript to place our proposal in the context of these works and also discuss what sets the proposed experiments apart from prior knowledge. In addition, and in response to comments from this reviewer and others, we have revised our proposal to incorporate some of the issues that these studies tackled, including the impact of sequencing depth (Experiment 7 in the revised proposal) and log-fold change filtering (Experiment 5 in the revised proposal). The text added to the Introduction is reproduced below:

Methods for differential analysis of bulk ChIP-seq and ATAC-seq datasets have previously been benchmarked[44–46]. These benchmarks not only compared individual DA methods, but also evaluated the impact of factors such as sequencing depth, number of replicates, or the characteristics of the underlying signal (e.g., broad histone modifications vs. sharp transcription factor (TF)-binding events[45,46]). However,

these benchmarks generally relied on simulations or a small number of case studies with unclear ground truth, and did not address the new challenges raised by single-cell epigenomics, including the analysis of markedly sparser datasets comprising thousands of cells.

3) Further of example a specific artifact not considered here is CG bias. This is also likely to affect promoters and enhancers regions differently.

https://pubmed.ncbi.nlm.nih.gov/36452861/

We agree that sample-specific GC bias has the potential to confound single-cell DA analysis, as suggested by this paper, which analyzed GC bias in the context of bulk ATAC-seq data. We do note that one of the conclusions of this paper is that "exploratory data analysis is essential to guide the choice of an appropriate normalization method for a given dataset." While we agree in principle with this statement, unfortunately we do not think that performing EDA of every dataset is compatible with the design of the Registered Report format. We also wish to highlight that, having reviewed the methods for normalization implemented in scATAC-seq analysis package, we are unable to identify any instance in which DA analysis is performed on a GC bias-corrected peak matrix. Nonetheless, we agree that beginning to tackle the question of whether GC bias correction improves DA analysis of scATAC-seq data would be informative for the field. Consequently, we have incorporated the smooth GC-full-quantile normalization method introduced in the cited reference into our benchmark of normalization methods. The text added to Experiment 6 (previously 5) is reproduced below:

Another important question, particularly when scATAC-seq data are not binarized, is whether and how the data should be normalized before DA analysis. The majority of the DA methods that we propose to analyze (including LR_{peaks}, negative binomial regression, Fisher's exact test, binomial test, permutation test, SnapATAC::findDAR, DESeq2, and edgeR) operate directly on either counts or binarized data as input. However, the remaining three methods (t-test, Wilcoxon rank-sum test, and $LR_{clusters}$) expectantly require normalization to produce the most biologically accurate results. We therefore propose to evaluate the impact of the most widely used normalization strategies for scATAC-seq data on these three DA methods by again repeating several of the analyses described above.

To identify the most widely used methods for normalizing scATAC-seq data for DA analysis, we reviewed the approaches implemented by published scATAC-seq analysis packages[33–43]. Based on this review, we propose to complement our analyses of the log-counts per 10,000 normalization implemented by default in Signac with additional DA analyses using the following additional normalization strategies:

6. Raw counts
7. Scaling counts to a total of 10,000 per cell without log-transformation, using the "relative counts" implementation in Signac
8. Scaling counts to the median number of counts per cell[33,36] without log-transformation
9. Scaling counts to the median number of counts per cell with log-transformation
10. TF-IDF normalization

Finally, some normalization methods that have been applied to bulk ATAC-seq data[81] explicitly seek to regress out technical properties of the underlying sequencing data, notably GC bias. To evaluate the impact of these approaches, we will also consider the effect of accounting for GC bias in the normalization procedure using a method recently developed for bulk ATAC-seq data[81] ('smooth GC-full-quantile').

Experiments and datasets: [...] In Experiments 6d, 6e, and 6f, we will repeat the analyses of Experiments 6a-c but instead evaluate the effects of normalization. We will analyze the concordance to bulk ATAC-seq data, the number of false discoveries, and the biases of these three DA methods using the six normalization strategies described above. Software implementations from the Signac package will be used for all normalization strategies except smooth GC-full-quantile, for which the implementation in the qsmooth package will be used. [...]

Data analysis plan: Datasets will be preprocessed or generated as described above for Experiments 1, 2, 3, and 4. For our analysis of binarization (Experiments 6a-c), several of the DA methods we propose to evaluate already operate on binarized data, precluding a comparison between qualitative and quantitative representations. These methods include Fisher's exact test, the binomial test, the

permutation test, and logistic regression using binary peak accessibility as the independent variable (LR_{peaks}), and will be excluded from the experiments proposed in this section. For the remaining seven DA methods, DA analysis will be performed as described above, but peak count matrices will be binarized prior to analysis. Conversely, our evaluation of normalization approaches will consider only the t-test, Wilcoxon rank-sum test, and $LR_{clusters}$. For our analysis of the QC matching procedure implemented in ArchR (Experiment 6g), we will repeat Experiment 1, but here perform DA analysis on equal numbers of cells from each condition matched by their TSS enrichment scores and $\log_{10}$(# of fragments), adapting code from ArchR for this purpose. Cells will be sampled from the 'background' (that is, the condition with more cells) in order to match the number of cells in the 'foreground' (that is, the condition with fewer cells).

4) It is unclear what DA problem is addressed altogether. Finding cell specific DARs or differential test (comparing accessibility on a cell between two biological conditions). These are two distinct problems as cell specific DAs are likely to have clearer differences than "comparative" Das.

We agree that there are important distinctions between detecting cell-type-specific versus condition-specific DARs, and that the former are likely to have larger effect sizes in general. Unfortunately, to the best of our knowledge (and we have extensively reviewed the literature in developing this proposal), the number of studies with matching single-cell and bulk ATAC-seq data precludes us from limiting our analysis to either problem alone. In this sense, we are facing similar challenges as initial benchmarks of DE methods for scRNA-seq data (i.e. Soneson and Robinson, *Nat. Methods* 2018). We have revised the manuscript to discuss this as a potential limitation of our study, in a new "Limitations" section that we are committed to retaining in the final, published report, reproduced below:

> Limitations
> Our study has a number of limitations that we are committed to acknowledging in the final, published version of this report. First, our proposed comparisons to parallel bulk ATAC-seq or multiome data leverage high-throughput assays carried out on matching samples (or even cells) under identical experimental conditions, but a limitation of these analyses is the requirement of the use of statistical methods for DA and DE analysis in these parallel data sets. We propose sensitivity analyses to test the impact of leaving out any given bulk DA or DE method. However, we cannot formally exclude the possibility that these comparisons will favor single-cell DA methods that yield output more similar to that of other methods in general for technical rather than biological reasons. Second, our comparison to multiome data is based on the premise that genes that are DE across biological conditions will also tend to have DA promoters, but exceptions to this assumption are known, notably during differentiation[64]. The majority of the comparisons that we propose entail the comparison of two fully differentiated cell types, mitigating this concern to some degree. Third, the relative paucity of datasets with paired bulk ATAC-seq data from identical biological conditions means that the datasets to be used in Experiment 1 comparisons involve comparisons both between and within cell types. However, these are comparisons with different underlying biological motivations, and in particular, the effect sizes of comparisons between cell types are likely to be larger than those within cell types and across conditions.

5) A major issue for studies benchmarking accessible regions is the lack of gold standards. The authors proposed using of parallel data (bulk atac or multiome data) to address this. I would consider this rather a silver standard, as these approaches require the use of statistical methods for DA and DE analysis in the parallel data sets. This generates a circular problem; validation is based on predictions made by the same method. Random selection of methods is a way to mitigate this, but I wonder if this approach does not simply favor DA methods, which predictions are more similar to other methods. Moreover, chromatin and transcription are known to have distinct dynamics in particular in cell differentiation (https://pubmed.ncbi.nlm.nih.gov/33098772/). This would be a further limitation on the use of multiome data as silver standard.

We certainly agree that using parallel bulk or multiome data does not constitute a perfect comparison (and we note that we did not use the term 'gold standard' anywhere in the manuscript). With that said, we do believe the experiments we propose constitute the best possible 'silver standard'. Simulations

can be a useful tool (and, indeed, we propose to incorporate them into our analyses), but we believe that repeated benchmarks of DE analysis in scRNA-seq data using simulated data have yielded the clear lesson that simulation studies are far from a gold standard either and—conversely—risk simply recapitulating the model from which the simulations were generated. We also note that even collecting our own, new 'gold standard' datasets would not address the reviewer's fundamental concern regarding the limitations of comparisons to parallel data.

We also appreciate that there might be a perception of circular reasoning, given the need to carry out parallel analyses of bulk ATAC-seq or scRNA-seq datasets in order to provide a reference for scATAC-seq analyses. We do feel it is worth underscoring that these are completely independent datasets, meaning that the reasoning is not circular in a literal sense. Furthermore, we are proposing to test the impact of leaving out each bulk DA or DE method (sensitivity analyses in Experiments 1 and 2) to mitigate the possibility that these comparisons might be affected by any universal biases of one particular method.

With this said, the reviewer does raise a valid caveat that this set of gold standards might favour approaches that are more similar to other methods in general. We acknowledge this and have revised the manuscript to address it as a potential limitation, in the new "Limitations" section that we are committed to retaining in the final, published report, as shown in response to the comment above.

Separately, we also acknowledge chromatin and transcription can display distinct dynamics. Most of the comparisons in Experiment 2 are between fully differentiated cell types, which mitigates this concern somewhat. We have, however, revised the manuscript to address it as a potential limitation, as shown in the response above.

Authors do not include Scenic (https://scenicplus.readthedocs.io/) related analysis (topic modelling derived DARs) in their analysis.

We made the conscious decision to exclude SCENIC from our evaluation as this package is not designed to identify individual differentially accessible peaks; instead, as the reviewer notes, SCENIC performs topic modeling of a peak matrix to identify clusters of peaks (or 'topics'). The resulting matrix of topic contributions per cell can then be used as input to a variety of downstream analysis tasks, including motif enrichment analysis. We have revised the manuscript to clarify this point and cite the SCENIC papers:

> For the t-test, Wilcoxon rank-sum test, negative binomial regression, and $LR_{clusters}$, implementations from Signac[35] will be used. The findDAR method will be drawn from the SnapATAC package. For LR_{peaks}, Fisher's exact test, binomial test, and permutation test, we have developed optimized implementations for sparse single-cell matrices, which will be used for the analysis. Finally, for DESeq2 and edgeR, we will use the implementations from our previously developed Libra package[46]. We excluded computational tools that do not perform statistical comparisons of individual genomic regions. For instance, we excluded SCENIC, which performs topic modeling of peak matrices to simultaneously cluster peaks and cells[55,61].

ArchR corrects for cell specific artifacts when finding DARs (read depth and other QC measures). This should be considered.
https://www.archrproject.com/bookdown/identifying-marker-peaks-with-archr-1.html

The strategy implemented in ArchR is to perform a statistical test for DARs on a subset of all cells available for analysis. Specifically, for a given set of 'foreground' cells (e.g., cells of a user-specified type), ArchR constructs a set of 'background' cells of equal size that are matched according to some set of QC features. By default, TSS enrichment and $\log_{10}$(# of fragments) are used to select a matching set of 'background' cells. This approach entails a trade-off whereby correcting for potentially

confounding artefacts comes at the cost of using only a subset of the available data for analysis. We therefore agree with the reviewer that it is of interest to test this approach within the framework of our proposed study. Consequently, we have incorporated it in the revised proposal into Experiment 6 (Best practices for scATAC-seq analysis, formerly Experiment 5). The text added to this section is reproduced below:

> One widely used software package, ArchR[33], does not use all available cells in DA analysis but rather performs a statistical test for DARs on a subset of all cells. Specifically, for a given set of 'foreground' cells (e.g., cells of a user-specified type), ArchR constructs a set of 'background' cells of equal size that are matched according to some set of QC features. By default, TSS enrichment and $\log_{10}$(# of fragments) are used to select a matching set of 'background' cells. This approach entails a trade-off whereby correcting for potentially confounding artefacts comes at the cost of using only a subset of the available data for analysis. We will evaluate the impact of this approach in Experiment 6g.
>
> *Data analysis plan:* Datasets will be preprocessed or generated as described above for Experiments 1, 2, 3, and 4. For our analysis of binarization (Experiments 6a-c), several of the DA methods we propose to evaluate already operate on binarized data, precluding a comparison between qualitative and quantitative representations. These methods include Fisher's exact test, the binomial test, the permutation test, and logistic regression using binary peak accessibility as the independent variable (LR_{peaks}), and will be excluded from the experiments proposed in this section. For the remaining seven DA methods, DA analysis will be performed as described above, but peak count matrices will be binarized prior to analysis. For our analysis of the QC matching procedure implemented in ArchR (Experiment 6g), we will repeat Experiment 1, but here perform DA analysis on equal numbers of cells from each condition matched by their TSS enrichment scores and $\log_{10}$(# of fragments), adapting code from ArchR for this purpose. Cells will be sampled from the 'background' (that is, the condition with more cells) in order to match the number of cells in the 'foreground' (that is, the condition with fewer cells).

Reviewer #3 (Remarks to the Author):

Detecting differential signals in sequencing-based assays is a foundational analysis in genomics. While in bulk assays, standard analytical methods of differential analysis have emerged (e.g., DESeq2, etc), no such practices have been established for single-cell data, especially from chromatin accessibility assays (ATAC-seq). Here, Yang Teo et al. propose to comprehensively evaluate all published methods to detect differential accessibility from single-cell experiments. To do this, they will make use of publicly available data, in which both single-cell and matched bulk-derived data is available. Overall, I think this is a timely and important endeavor, and will help the field coalesce around a best-practice.

Overall, I find the proposed experiments/analysis well designed and find no technical problems and the authors should proceed as planned.

Thank you for your careful review of our manuscript and your very encouraging remarks.

I do have two concerns regarding the design/considerations of analysis:

1) Unlike RNA-seq (single cell or bulk), chromatin accessibility assays do not have a fixed annotation that is used to compare two experiments (GENCODE, Refseq, etc.), and peak detection is first needed to define "informative" genomic regions. They authors should consider peak detection as an important part of the single cell DA detection analysis. The number of features tested will affect the multiple test correction approaches as well as normalization (see pt. 2 below). Because chromatin accessibility in different cell types can vary drastically on top of the considerable technical variation, this could be a significant problem to overcome. In addition, how are peaks to be defined in the single-cell data -- will a pseudobulk procedure be used? Alternatively, one could use a fixed reference scaffold as proposed in Meuleman et al., 2020 Nature.

We certainly agree that the issue of peak calling is critical to the analysis of scATAC-seq data, and in particular the biological interpretation of DA results. At the same time, our intention was not to propose a benchmark of peak-calling approaches. In our reasoning, so long as every DA method is provided with the same set of peaks as input, all DA methods should be on an ‘even playing field’. The specific set of peaks that are provided as input to the methods is therefore only relevant to a comparison of DA methods insofar as an interaction could exist between the relative performance of a given DA method and the set of genomic features provided as input. An example of such an interaction could entail a scenario where (for instance) the Wilcoxon rank-sum test only performs well with a specific set of peaks as input.

Your comment spurred us to consider the possibility that such an interaction might exist. To address this possibility, we have added another proposed sensitivity analysis to Experiment 1. This experiment requires us to use an identical set of peak definitions between the single-cell and matching bulk ATAC-seq data in order to quantify the concordance of DA between the two. In the original submission, we proposed to take advantage of the availability of bulk ATAC-seq by calling peaks in the bulk data, and using these peaks as a reference to assemble peak count matrices in the single-cell data. In the revised submission, we additionally propose to test another peak-calling strategy that is more representative of how single-cell datasets are generally analyzed: namely, using the single-cell data only. Specifically, we will call peaks using MACS2 on pseudobulks from the single-cell data, mirroring the approach used in Signac, ArchR, and SnapATAC. We will present the resulting concordance results and then formally test for an interaction term between DA method performance and the peak set used as input.

We have revised the Data analysis plan for Experiment 1 to incorporate this new analysis as shown below:

> Measuring concordance between DA analyses of single-cell and bulk ATAC-seq requires a matching set of peaks to be defined in both datasets. In the analyses described above, we have proposed to obtain such a peak set by calling peaks in the matched bulk ATAC-seq data, under the hypothesis that analyzing bulk data would yield a higher-quality set of peak definitions. However, this procedure is at odds with the design of most scATAC-seq studies, which instead call peaks in 'pseudobulks' of single-cell data; this is the strategy implemented in widely used software packages such as Signac, ArchR, and SnapATAC. This divergence raises the possibility that the relative performance of single-cell DA methods may depend on the definitions of the peaks provided as input. To evaluate this possibility, we will perform a final sensitivity analysis in which we repeat the calculation of concordance using peaks called from 'pseudobulks' of single-cell data, adapting code from Signac. In addition to presenting the resulting concordance results, we will then use a linear model to formally test for an interaction term between DA method performance and the peak set used as input.
>
> In summary, the list of proposed analyses is as follows:
>
> 2. Concordance between single-cell and bulk DA, as quantified by the AUCC
>   a. Sensitivity analysis: impact of removing bulk DA methods
>   b. Sensitivity analysis: impact of the parameter *k*
>   c. Sensitivity analysis: impact of filtering
>   d. Sensitivity analysis: impact of peak definitions

2) Although not discussed much in the literature for chromatin accessibility data, the required normalization of samples is likely quite different that the methods used for RNA-seq data. While, DESeq2 typically uses a single scale factor (geometric mean of read depth) for an entire sample to normalize RNA-seq data, this is likely not appropriate for accessibility data because of regional differences (promoters more accessible that distal elements), most elements are very weak (distribution of peak densities are exponentially decaying in a sample), and binary nature of single cell data (theoretically at most 2 read per region). Because normalization can have large impact on DA analyses, the authors should consider evaluating how different approaches affect results.

We agree that normalization is an important point. The majority of the DA methods that we propose to analyze, including LR_{peaks}, negative binomial regression, Fisher's exact test, binomial test, permutation test, SnapATAC::findDAR, DESeq2, and edgeR, operate directly on the count matrix. Three methods, including the t-test, Wilcoxon rank-sum test, and $LR_{clusters}$, can operate directly on the count matrix but expectantly require normalization to produce the most accurate results. In our original proposal, we had proposed to use the implementation of these tests in Signac, which by default scales counts to 10,000 per cell and log-transforms the scaled values (log-TP10K normalization).

Your comment, as well as one from reviewer 1, spurred us to consider the effects of alternative normalization approaches on the accuracy of these three methods. Accordingly, we reviewed the approaches to normalization proposed by published scATAC-seq analysis packages (**Fig. 1b**). Signac additionally provides methods for TF-IDF normalization as well as a TP10K normalization without log-transformation ("relative counts"). ArchR implements a subsampling procedure to match cells in the two groups being compared based on read depth, then further scales counts in each cell to the median number of reads in TSSs across cells by default (MarkerFeatures.R, lines 207-218). EpiScanpy implements a similar normalization strategy to the relative counts method in Signac by default, except that instead of scaling counts to a fixed depth (e.g., 10,000 counts per cell), the median of total counts across all cells is used as the scaling factor, analogously to ArchR. Log-transformation can then optionally be applied afterwards. The remaining packages (SnapATAC, scABC, MAESTRO, Scasat, scATAC-pro, Destin, chromSCape) either perform DA analysis directly on the count matrix or implement one of the approaches to normalization described above.

Based on these observations, we now propose to test the impact of the following additional normalization strategies on the t-test, Wilcoxon rank-sum test, and $LR_{clusters}$:

1. Raw counts
2. Scaling counts to a total of 10,000 per cell without log-transformation, using the "relative counts" implementation in Signac
3. Scaling counts to the median number of counts per cell without log-transformation
4. Scaling counts to the median number of counts per cell with log-transformation
5. TF-IDF normalization
6. Last, in response to a comment from reviewer 2, we will also consider the effect of accounting for GC bias in the normalization procedure using a method recently developed for bulk ATAC-seq analysis ('smooth GC-full-quantile'; van den Berge et al., *Cell Rep. Methods* 2022).

We have incorporated these new analyses as a complement to the binarization analyses in Experiment 6 (previously 5). The text added to this section is reproduced below:

> Another important question, particularly when scATAC-seq data are not binarized, is whether and how the data should be normalized before DA analysis. The majority of the DA methods that we propose to analyze (including LR_{peaks}, negative binomial regression, Fisher's exact test, binomial test, permutation test, SnapATAC::findDAR, DESeq2, and edgeR) operate directly on either counts or binarized data as input. However, the remaining three methods (t-test, Wilcoxon rank-sum test, and $LR_{clusters}$) expectantly require normalization to produce the most biologically accurate results. We therefore propose to evaluate the impact of the most widely used normalization strategies for scATAC-seq data on these three DA methods by again repeating several of the analyses described above.
>
> To identify the most widely used methods for normalizing scATAC-seq data for DA analysis, we reviewed the approaches implemented by published scATAC-seq analysis packages[33–43]. Based on this review, we propose to complement our analyses of the log-counts per 10,000 normalization implemented by default in Signac with additional DA analyses using the following additional normalization strategies:
>
> 11. Raw counts
> 12. Scaling counts to a total of 10,000 per cell without log-transformation, using the "relative counts" implementation in Signac
> 13. Scaling counts to the median number of counts per cell[33,36] without log-transformation
> 14. Scaling counts to the median number of counts per cell with log-transformation
> 15. TF-IDF normalization
>
> Finally, some normalization methods that have been applied to bulk ATAC-seq data[81] explicitly seek to regress out technical properties of the underlying sequencing data, notably GC bias. To evaluate the impact of these approaches, we will also consider the effect of accounting for GC bias in the normalization procedure using a method recently developed for bulk ATAC-seq data[81] ('smooth GC-full-quantile').
>
> *Experiments and datasets:* [...] In Experiments 6d, 6e, and 6f, we will repeat the analyses of Experiments 6a-c but instead evaluate the effects of normalization. We will analyze the concordance to bulk ATAC-seq data, the number of false discoveries, and the biases of these three DA methods using the six normalization strategies described above. Software implementations from the Signac package will be used for all normalization strategies except smooth GC-full-quantile, for which the implementation in the qsmooth package will be used. [...]
>
> *Data analysis plan:* Datasets will be preprocessed or generated as described above for Experiments 1, 2, 3, and 4. For our analysis of binarization (Experiments 6a-c), several of the DA methods we propose to evaluate already operate on binarized data, precluding a comparison between qualitative and quantitative representations. These methods include Fisher's exact test, the binomial test, the permutation test, and logistic regression using binary peak accessibility as the independent variable (LR_{peaks}), and will be excluded from the experiments proposed in this section. For the remaining seven DA methods, DA analysis will be performed as described above, but peak count matrices will be binarized prior to analysis. Conversely, our evaluation of normalization approaches will consider only the t-test, Wilcoxon rank-sum test, and $LR_{clusters}$. For our analysis of the QC matching procedure implemented in ArchR (Experiment 6g), we will repeat Experiment 1, but here perform DA analysis on equal numbers of cells from each condition matched by their TSS enrichment scores and $\log_{10}$(# of fragments), adapting code from ArchR for this purpose. Cells will be sampled from the 'background' (that is, the condition with more cells) in order to match the number of cells in the 'foreground' (that is, the condition with fewer cells).

Finally, regarding the reviewer's comments on size factors in DESeq2, we agree that this is an interesting direction for future research. However, we are not aware of work exploring this assumption or proposing a mixture of per-sample scaling factors, particularly in the context of ATAC-seq data. Therefore, we believe that new method development would be required to examine these assumptions, and this would fall out of scope of our proposed comparison of existing methods.

Reviewer #1 (Remarks to the Author):

The authors have fully addressed all of my comments on the proposal. I appreciate the thorough and thoughtful revision, and look forward to reading the results of the study.

Reviewer #2 (Remarks to the Author):

I have read the registered report and the authors' reply. Authors have only partially addressed my requests. Also, I noticed additional issues with the study after considering the revised version and replies.

Major points:

1. None of the analyses based on real ATAC-seq data proposed here have a "ground truth" (or a “gold standard”). Authors need to refrain from using the term "ground truth" unless referring to simulated data.

2. I notice a problem with the "ground truth" labels proposed in experiment 1, which are based on parallel bulk ATAC-seq data. This assumes that bulk ATAC-seq has better quality than single-cell ATAC-seq data. Interestingly, my experience is that the quality of single-cell ATAC-seq data is better than bulk ATAC-seq. For instance, pseudo-bulk scATAC-seq data tends to have higher FRIP scores whenever a reasonable number of cells (<200) is present in the cluster. This is because it's possible to remove low-quality, unviable cells, or contaminants after QC and clustering analysis of scATAC-seq data. I invite the authors to check this in their data (FRIP scores of pseudo-bulk vs. bulk). Another exercise is to examine bulk (pseudo or not) ATAC-seq profiles in a genome browser and contrast predicted peaks. They should also provide evidence that bulk-specific peaks are more likely to represent biological signals than artifacts. This approach is likely to be valid only when the number of cells in a cluster is low (>50).

3. I also miss in the report an analysis on the impact of differential accessibility (DA) peaks in enhancers vs. promoter regions. Usually, such peaks have distinct characteristics, such as different sizes or CG bias. Of particular biological relevance are enhancers, as they are highly cell-specific and potentially harder to detect due to low read sizes. Authors should evaluate DA performance for enhancers and promoters separately.

4. Authors need to clarify the scope of the DA analysis evaluated in each experiment (cell type comparison or condition comparison). The report (or final manuscript) would benefit from a schematic figure describing all the evaluated scenarios with major study factors (type of data, standard used for benchmarking, type of genomic intervals used, etc.). A good example is Fig. 2 of Cheng et al. 2018.

Reviewer #3 (Remarks to the Author):

The authors suggested changes will substantially improve the proposed research and have addressed my concerns.

Response to reviewers

We thank the reviewers again for their thoughtful feedback on our Registered Report, "**Best practices for differential accessibility analysis in single-cell epigenomics**." Here, we respond to the additional points raised in the second round of review. We have highlighted all the changes to the text in the accompanying manuscript, but also included the new text in the following responses for convenience. Throughout this document, reviewer comments are shown in **blue**, with our own response in **black**, and changes to manuscript text or figure captions in **red**.

Reviewer #1 (Remarks to the Author):

The authors have fully addressed all of my comments on the proposal. I appreciate the thorough and thoughtful revision, and look forward to reading the results of the study.

Thank you for your encouraging comments and thoughtful feedback.

Reviewer #2 (Remarks to the Author):

I have read the registered report and the authors' reply. Authors have only partially addressed my requests. Also, I noticed additional issues with the study after considering the revised version and replies.

Major points:

1. None of the analyses based on real ATAC-seq data proposed here have a "ground truth" (or a "gold standard"). Authors need to refrain from using the term "ground truth" unless referring to simulated data.

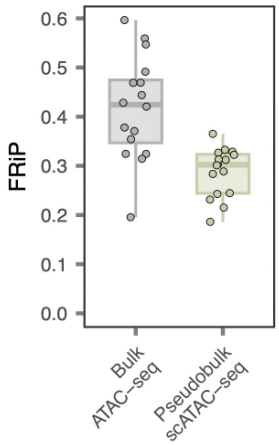
We revised the text to remove the use of the term "ground truth" and clarify that, except in our simulation studies, we are referring to the availability of matching bulk ATAC-seq or scRNA-seq. This change applies throughout the manuscript; for example, in the abstract:

> **[...] Here, we propose to conduct a systematic evaluation of every statistical method that has been applied to identify DA regions in single-cell ATAC-seq (scATAC-seq) data. We will leverage a compendium of scATAC-seq experiments with matching bulk ATAC-seq or scRNA-seq in order to assess the accuracy, bias, robustness, and scalability of each statistical method. [...]**

2. I notice a problem with the "ground truth" labels proposed in experiment 1, which are based on parallel bulk ATAC-seq data. This assumes that bulk ATAC-seq has better quality than single-cell ATAC-seq data. Interestingly, my experience is that the quality of single-cell ATAC-seq data is better than bulk ATAC-seq. For instance, pseudo-bulk scATAC-seq data tends to have higher FRIP scores whenever a reasonable number of cells (<200) is present in the cluster. This is because it's possible to remove low-quality, unviable cells, or contaminants after QC and clustering analysis of scATAC-seq data. I invite the authors to check this in their data (FRIP scores of pseudo-bulk vs. bulk). Another exercise is to examine bulk (pseudo or not) ATAC-seq profiles in a genome browser and contrast predicted peaks. They should also provide evidence that bulk-specific peaks are more likely to represent biological signals than artifacts. This approach is likely to be valid only when the number of cells in a cluster is low (>50).

With respect to the reviewer's first request, we were not entirely sure whether it was in line with the spirit of a Registered Report, in which the central idea is to carry out a preregistered analysis. Nonetheless, because the characteristics of the peaks are ancillary to the focus of our analysis, which

is on comparing DA methods—a point on which we elaborate further below—we have undertaken the effort of calling peaks on bulk and pseudo-bulk data from the matching bulk and single-cell datasets, and performed an exploratory data analysis to investigate the characteristics of these peaks. The results show that the FRiP is significantly higher for bulk peaks as compared to pseudo-bulk peaks ($p = 9.2 \times 10^{-6}$, paired t-test). It is not clear to us whether this is a generalizable result, given the relatively small number of paired bulk and single-cell ATAC-seq datasets investigated here, but in general we feel this result substantiates our proposal to use peaks called in bulk ATAC-seq as the reference in our primary analysis (although we have also taken the reviewer's suggestion to consider pseudobulk peaks as a separate reference, a point that we also discuss below). We will include this plot in the final manuscript.



Response Fig. 2.1. Fraction of reads in peaks in bulk ATAC-seq (with peaks called from bulk ATAC-seq data) versus single-cell ATAC-seq (with peaks called from pseudobulks of single-cell ATAC-seq data).

Separately, with respect to the reviewer's request that we demonstrate "bulk-specific peaks are more likely to represent biological signals than artifacts," it is not entirely clear to us how this would be relevant to our proposed analysis. We think it is helpful to reiterate the goal of the proposed work. Our goal is not to perform a comparison of peak-calling methods. Instead, we wish to compare methods for identifying differentially accessible peaks. We feel that, so long as every DA method is provided with the same set of peaks as input, all DA methods should be on an 'even playing field'. Moreover, in our previous revision, we had proposed an additional experiment to evaluate the impact of the peak definitions on the DA results, whereby we proposed to test the accuracy of DA methods with peaks called from pseudobulks of the single-cell data, as the reviewer had requested. This would directly address the reviewer's concern that peak-calling on pseudobulks may yield more accurate results than peak-calling on matching bulk ATAC-seq.

Nonetheless, after reflecting on this comment, we realized that in addition to this analysis, we could also directly test the possibility that artifactual peaks could influence the *relative* performance of single-cell DA methods, through a new analysis which we have incorporated into the revised proposal. Briefly, we propose to perform peak calling with MACS2 using a more lenient q-value threshold of 0.1. We would then add the peaks detected only with this q-value threshold to the set of peaks detected under the normal q-value threshold, and repeat the analysis. We reasoned that this would be preferable to using random genome regions because the newly-added peaks would have some fragments associated with them, but would be disproportionately likely to be artifacts relative to the original peak set. Repeating DA analysis on the expanded peak set would therefore allow us to directly measure the impact of including artifactual peaks on the relative performance of single-cell DA methods. We added this as a sensitivity analysis to Experiment 1 in the revised proposal:

> A related issue in the analysis of both single-cell and bulk ATAC-seq data that represent artifacts of data processing rather than true biological signal. To test whether these artifactual peaks could influence the relative rankings of single-cell DA methods, we will perform an additional sensitivity analysis in which we deliberately introduce a certain amount of artifactual peaks into the dataset. We will achieve this by performing a second round of peak calling with MACS2 in the matched bulk ATAC-seq data at a deliberately increased q-value threshold of 0.1. To avoid analyzing broader peaks throughout the genome, we will remove peaks that overlap with the original peak set. We will then combine these remaining peaks with the original peak set and repeat DA analysis at the expanded peak set, in which a substantial proportion of the additional peaks are expected to be artifactual. We will repeat the analysis of concordance between single-cell and bulk DA as described above, and use a linear model to formally test for an interaction term between DA method performance and the peak set used as input.

In summary: through exploratory data analysis, we have shown that peaks called from bulk and single-cell ATAC-seq are of at least comparable quality, if not even higher quality in the bulk data. We have argued that although peak-calling is certainly an important consideration in the analysis of scATAC-seq data, it is relevant to the goals of our study only insofar as it might affect the relative performance of DA methods, i.e. through an interaction between peak characteristics and the accuracy of a given DA method. We had previously proposed to study this possibility by testing the performance of DA methods on peaks called from both bulk and single-cell data. We now additionally propose to directly test the effect of artifactual peaks by adding peaks called at a lower q-value threshold to the original peak set. Together, we believe these analyses should enable a thorough investigation of how different peak calling strategies affect methods for single-cell DA analysis, the latter being the focus of our proposal.

3. I also miss in the report an analysis on the impact of differential accessibility (DA) peaks in enhancers vs. promoter regions. Usually, such peaks have distinct characteristics, such as different sizes or CG bias. Of particular biological relevance are enhancers, as they are highly cell-specific and potentially harder to detect due to low read sizes. Authors should evaluate DA performance for enhancers and promoters separately.

Thank you for this suggestion. We agree it would be of interest to evaluate DA performance for enhancers and promoters separately. We added this experiment as a sensitivity analysis to Experiment 1, as shown below:

> A final issue that is relevant to single-cell DA analysis is whether single-cell DA methods exhibit differential performance on peaks located in enhancer versus promoter regions. Enhancer elements are generally supported by fewer read counts than promoters in both bulk and single-cell ATAC-seq data[2,32], which raises the possibility that DA inference will likely be less accurate in general for enhancer elements. However, the question of whether certain DA methods are better-suited for analysis of enhancer elements in single-cell data has not, to our knowledge, previously been addressed. To evaluate this possibility, we will repeat our analysis of DA concordance between single-cell and matched bulk ATAC-seq data, but calculating concordance separately for peaks located in enhancer versus promoter elements, as derived from ENCODE Registry of Candidate cis-Regulatory Elements[64].

4. Authors need to clarify the scope of the DA analysis evaluated in each experiment (cell type comparison or condition comparison). The report (or final manuscript) would benefit from a schematic figure describing all the evaluated scenarios with major study factors (type of data, standard used for benchmarking, type of genomic intervals used, etc.). A good example is Fig. 2 of Cheng et al. 2018.

Thank you for this suggestion. We have added a new figure to the revised manuscript in which we specify (i) the datasets used in the different evaluation scenarios, (ii) the standard being used for benchmarking (i.e. matched bulk data versus scRNA-seq data), (iii) the genomic intervals being used (i.e. peaks called from bulk ATAC-seq, peaks called from pseudobulk scATAC-seq, or genomic intervals around the TSS), and (iv) the biological nature of the comparison (i.e. cell type or condition

DA). This is provided as **Fig. 2** in the revised proposal (and is reproduced below) and of course, we are committed to including this figure in the final manuscript.

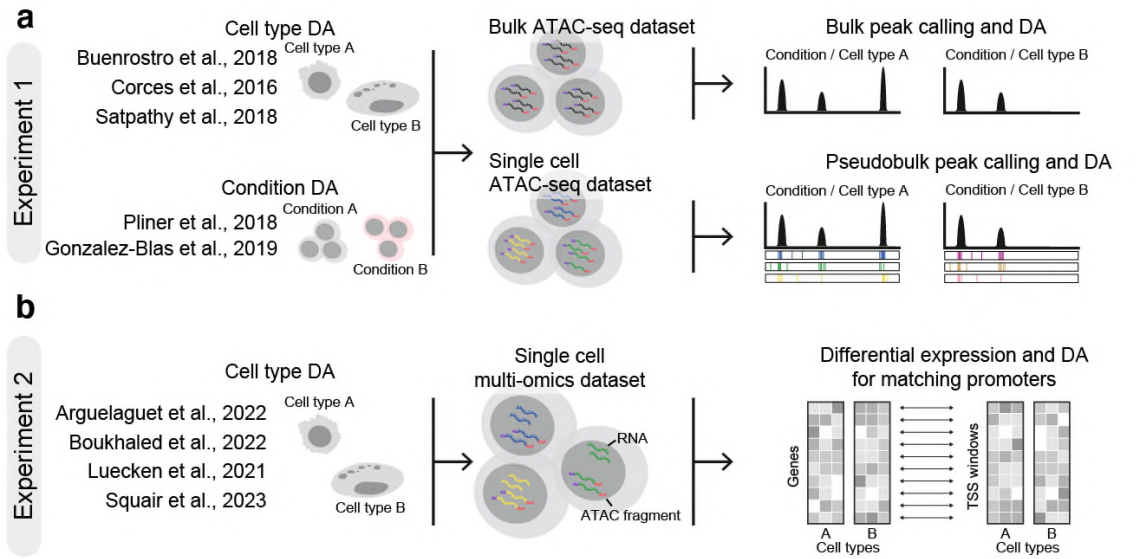


Fig. 2 | Design of Experiments 1 and 2.
a, Design of Experiment 1. DA analysis will be performed between cell types or conditions for single-cell datasets with matching bulk ATAC-seq as a reference. Peaks will be called either in the bulk ATAC-seq data or in pseudobulk single-cell data.
b, Design of Experiment 2. DA analysis will be performed between cell types using matched scRNA-seq data from the same cell as a reference, comparing DA of genomic intervals around the TSS to gene-level differential expression.

Reviewer #3 (Remarks to the Author):

The authors suggested changes will substantially improve the proposed research and have addressed my concerns.

Thank you for your encouraging comments and thoughtful feedback.

Reviewer #2 (Remarks to the Author):

I am satisfied with the authors change and I am looking forward to the final manuscript version.

Review Study Plan (Stage 2)

Reviewer #1 (Remarks to the Author):

In their registered report, Teo et al. present a benchmarking study of differential accessibility methods for single-cell ATAC-seq analysis. They focus on 8 experiments using real and simulated datasets, aiming to assess the accuracy, false positive rate, bias, scalability, and influence of data normalization on DA identification. While these experiments are comprehensive, some of the presentation of results and methodological approach needs further consideration.

Throughout the manuscript it is unclear what definition is being used to define a differentially accessible region. How such regions are defined should be clearly stated early in the manuscript. What p-value cutoff is used, what multiple testing correction method is used, and what minimum fold change and minimum accessibility is used is required information that should be added.

Most of the issues relate to experiment 5 (fold change filtering). Here the authors have used a method of ranking cells based on fold change and the lower X% of cells. This is not an approach that would be used in practice, and would provide poor results in cases where all the peaks have low fold change or high fold change. Different cutoff values should be chosen instead (for example, 1-fold. 2-fold, 3-fold). This was stated in the original review for the registered report.

How fold changes are computed needs to be more thoroughly described. Different softwares used different implementations. The authors should provide the formula used in the methods section, and ensure that this is indeed what is being used in the software running the tests. Differences in, for example, the magnitude of the pseudocount or when the pseudocount is added can have a large impact on the fold change values returned. Log2 fold change should also be reported as this is more interpretable (rather than natural log).

In general, more methodological details are required. For example, how is FDR computed, how is TF-IDF computed. The authors should state the formula and functions (with parameters used), and software versions. Notably there have been significant bugs in past versions of Seurat regarding the calculation of fold changes with scATAC data, so it is essential that versions are stated and the authors are using up-to-date software implementations.

Figure 5 is hard to interpret in several ways. First, the X-axis states “# of cells” but a percentage is given along the axis. Based on the description in the methods, I believe this should be “% peaks removed” or something similar. Second, the legend states “Effect size (Cohen’s d) of increasingly stringent log-fold change filtering on the AUCC between single-cell and bulk ATAC-seq DA.” It’s unclear what is actually being shown. The text also discusses the comparison of fold change filtering vs random filtering, but the figure compares fold change filtering to no filtering. The authors should rethink how to present the results from experiment 5 in a clear and compelling way. The most useful information to readers is: how does the false positive rate change for different fold-change filters, how does the concordance between bulk vs single-cell

and RNA vs ATAC change with different fold-change filters, how does all of this compare to random removal of the same number of peaks.

In the normalization section (experiment 6) it's stated that "We found that the number of false discoveries was consistently reduced after applying TP10K or TP(median) normalization (Fig. 6i-k and Supplementary Fig. 17b-d). However, this improvement in the control of false discoveries came at the expense of biological accuracy (Fig. 6h)." – this does not seem to be the case for TF-IDF and the t-test. Better panel labeling is also needed for figure 6. Several panels appear with the same axes and title, and the reader has to carefully look at the legend to work out what it's showing. Expressing all the results as Cohen's D here decreases the interpretability of the results. It is better to report, for example, the false positive rate itself (figure panel 6i, etc).

Other points:
- "analytical workflows implemented in different laboratories bear little resemblance to one another" – this is inaccurate. The workflows, while having differences, do fundamentally follow similar strategies of quantifying peaks, normalizing the data, defining groups of cells, and performing a statistical test.
- In figure 8 (scalability) the number of tests being performed should be stated
- In the scalability section, the differential testing functions are incorrectly attributed to the Signac package (they are in fact part of Seurat). Same is true for several other functions.

Reviewer #2 (Remarks to the Author):

This registered report is well written and data analysis is well presented. Results indicate the advantage of standard bulkRNA-seq approaches (DEseq and EdgeR) in the overall problem. While results support this, it cannot be excluded that these conclusions arise from the fact DE analysis used to generate gold standards in most of the evolution scenarios.

This was discussed in previous rounds of review and is indicated in the limitation section. Interesting contributions to the literature include the analysis of best practices in scATAC-seq data, which indicates potential benefits of binarization, effects of normalization and peak calling. A standing critical point is the fact that majority of results do not use best practices (mostly adopted in the scATAC-seq literature and confirmed by some of the results) in their benchmarking. Authors needs to address this and discuss these points and describe potential limitations more explicitly.

Major points:

1. AUCC reported in Fig. 2 are overall quite low (maximum of 0.32) and indicate that most methods perform quite bad in the DA problem. Another explanation for this would be that this evaluation is sub-optimal, i.e. DA peaks in single cell data do not match the ones in bulk due to lower quality of either bulk or scATAC-seq experiments. This should be discussed in the manuscript.

2. Usually single cell ATAC-seq peaks are derived from the bulking the single cell data by

considering either all or clustered cells. As my request, authors include an analysis on using the first evaluation (Fig. 2). Interestingly, reported results have significantly higher AUCCs (maximum of 0.42 and 25% improvement for all methods). This supports a higher quality of scATAC-seq data, as previously discussed in my revisions. Also, in contrary to the manuscript statement, the ranking of methods is distinct for panel Supp. 1b and Fig. 2b.

3. Still on the same topic using promoters lead to a much higher AUCC (0.76) and very distinct method rankings. The reason besides this is not discussed in the manuscript.

4. Given points 1, 2 and 3, authors should be more critical regarding their overall conclusion for this analysis. In one hand, methods could be ranked be considering all distinct scenarios. On the other hand, authors need to highlight results based on scenarios implementing best practices in of scATAC-seq.

5. Points 1-4 potentially also impact the multimodal data analysis (Fig. 2d & e).

6. Results from the evaluation of Fig. 3 are stronger than Fig. 2, as some approaches can control for false positives (no false positive peaks). A question raised is how best practices (peak calling from scATAC-seq) impact this evaluation needs also to be addressed here. Indeed, it is shown in Fig. 6 that best practices (binarization) do impact of the performance (and ranking) of methods. Peak calling strategies in Fig. 3 are unfortunately not evaluated in this context. Again, it is unclear how their adoption would change rankings reported here. As in point 4. authors should have an overall ranking but also a ranking based on the adoption of the best practices.

7. The analysis of the impact of the FC is sound, but overall difficult to follow. As a (minor) suggestion authors could consider shortening it.

8. The best practice analysis is an interesting aspect of this study. Authors should mention here the peak calling strategy (see point 2). The binarization aspect is also relevant, as this potentially contradict the previous literature. On the other hand, authors should consider the fact that previous work supporting the use of count data focus on the cell type identification problem and not DA analysis.

9. A major critical point of the study is the fact it does not consider the combination of DA methods with best practices (scATAC-seq peak calling; binarization, type of normalization). Indeed, in their literature research the authors only evaluated the use of DA methods, but not how the data was previously treated (binarization, normalization). As expected, these steps will impact each other, as shown in the manuscript. On the other hand, some of the best practices are already used by default by some of the evaluated methods. For example, snapatac is based on both scATAC-seq peak calling and binarization. Both these practices have a large impact in snapatac DA method results. Indeed, an evaluation considering the best practices (or best approach per DA method) will give a more realistic evaluation and relevant results for practitioners of scATAC-seq analysis.

10. Alternatively, authors need to clearly indicate limitation of their study, whenever conclusions are based on non-best practices.

Minor

Authors should make clear which strategy is being used for each evaluation (how data are normalized or not, how peaks are called). A matrix in the supplement would be extremely helpful.

Page 20, “under the hypothesis that bulk ATAC would yield higher quality set of peaks”. This could be revised, as the evaluation do not support this.

Reviewer #3 (Remarks to the Author):

In this manuscript, the authors address an exciting, yet problematic analysis applied to single-cell chromatin accessibility profiling data: detecting and identifying differential active elements. While these types of analyses in bulk-derived data have well-understood best practices, their behavior in single cell data has not been thoroughly characterized. As the authors importantly state, it is still unresolved to what extent data from individual cells is actually quantitative. Here, the authors perform exhaustive comparisons of differential analysis methods applied to single-cell ATAC-seq data in an attempt to define a best-practice.

Overall, I find that this manuscript would be useful for the field, however, in its current state, its salient point (ie., defining a best practice) gets lost in the details/technicalities. In addition, a major shortcoming (and missed teaching opportunity) is that the authors do not discuss why certain methods work better, and importantly, when their use is appropriate given experimental design choices/realities.

Major points:

1) Section "Experiment 2" comparing DA in sc-ATAC and sc-RNA (mulitome analysis) doesn't seem to add anything outright useful to the manuscript. One problem is that I am not sure how one defines a good concordance between the two modalities. The connectivity between regulatory elements (accessible peaks) and genes is largely unknown and no current method seems to do be particularly successful (for example, distance and the popular ABC model seem to perform roughly similarly). I understand the hypothesis that if a gene expression is changed than it is likely driven by changes in the cis-regulatory elements within some "regulatory domain", however, the definition of these domains are arbitrary (as is the approach taken in this manuscript). As the main goal of this paper is evaluating which methods are best for detecting differentially accessible peaks, the analyses in this section seem to be a distraction.

2) The authors apply/evalute all of the methods but do not comment on their statistical underpinnings (pros and cons for different applications/use cases). For example, the non-psuedobulking approaches are destined to perform poorly because they make assumptions of about background distribution of the raw count signal that a largely not reflected in the data (for example the data is overdispersed with respect to the binomial distribution). A discussion of

statistical methods and assumptions behind these different approaches will help the average user/reader build an intuition about the general applicability and appropriateness of different methods for their particular application.

3) Related to my comment above, it seems that the non-psuedobulk methods are commonly used because single-cell studies rarely perform biological or technical replicates (despite their weaker performance). Given the cost associated (and 10X genomics monopoly on the market) with single-cell studies, I doubt this will change anytime soon. As such, I think that some more attention/focus should be placed on how to perform DA analyses when no replicates are available or are limited (or demonstrate the importance of replicates).

4) Generally, I find the figures overly technical and it is hard to parse the immediate conclusion(s). Personally speaking, I don't think it is necessary to show all method contrasts in each figure. In addition (or alternatively) I might group the methods by some of their commonalities (psuedobulk, etc.)

Minor points:

1) The text could be substantially condensed for readability.

2) Figure 5: it appears from my reading of the text that the x-axis should be "% percent filtered peaks" and not "# of cells"

Response to reviewers

We thank the reviewers again for their thoughtful feedback on the Stage 2 submission on our Registered Report, "**Best practices for differential accessibility analysis in single-cell epigenomics**," and we are pleased to respond to the additional points raised in this third round of review.

In responding to the points raised in this round of review, we were mindful of several salient aspects of the Nature publishing group policies relating to the Registered Reports format:

1. First, no preregistered analyses can be removed altogether from the manuscript ("The outcome of all registered analyses must be reported in the manuscript"). Therefore, we were unable to accommodate reviewer requests to remove of any of our preregistered analyses.
2. Second, additional analyses that were not included in the registered submission are permissible but must be clearly indicated as such and documented in a separate section of the manuscript ("Additional analyses that were not included in the registered submission [...] are admissible but must be clearly justified in the text, appropriately caveated, and described in a separate section of the Results titled 'Exploratory analyses'"). Therefore, we have addressed requests to perform additional, non-preregistered experiments by presenting these in a newly added "Exploratory analyses" section at the end of the Results.
3. Third, the Introduction of the manuscript cannot be altered at this stage ("Apart from minor stylistic revisions, the Introduction should not be altered from the approved Stage 1 submission").

We have highlighted all the changes to the text in the accompanying manuscript, but also included the new text in the following responses for convenience. Throughout this document, reviewer comments are shown in **blue**, with our own response in **black**, and changes to manuscript text or figure captions in **red**.

Reviewer #1 (Remarks to the Author):

In their registered report, Teo et al. present a benchmarking study of differential accessibility methods for single-cell ATAC-seq analysis. They focus on 8 experiments using real and simulated datasets, aiming to assess the accuracy, false positive rate, bias, scalability, and influence of data normalization on DA identification. While these experiments are comprehensive, some of the presentation of results and methodological approach needs further consideration.

Thank you again for your comments on the Stage 2 submission of our manuscript.

Throughout the manuscript it is unclear what definition is being used to define a differentially accessible region. How such regions are defined should be clearly stated early in the manuscript. What p-value cutoff is used, what multiple testing correction method is used, and what minimum fold change and minimum accessibility is used is required information that should be added.

This information is provided in the Methods. To avoid reproducing large sections of the Methods in this response, we briefly summarize our answers to the questions raised by the reviewer below, but have confirmed that all of this information can indeed be found in the manuscript and revised the manuscript to address instances where this information could have been presented more clearly. We also wish to reiterate that all of these analyses were preregistered in our Stage 1 submission, and we have adhered precisely to the plan that was previously seen and approved by the three reviewers.

- In Experiments 1 and 2, the use of the area under the concordance curve (AUCC) as the evaluation metric requires only a ranked list of peaks. Therefore, peaks were ranked by p-values assigned by each DA method, and no p-value cutoffs, multiple testing correction, or fold change cutoff were applied.
    - For Experiment 1, no minimum accessibility cutoff was applied for the experiments in the main text, and a cutoff of accessibility (non-zero counts) in ≥5% of cells was applied in **Supplementary Fig. 1c**.
    - For Experiment 2, no minimum accessibility cutoff was applied for the experiments in the main text, and a cutoff of accessibility (non-zero counts) in ≥1% of cells was applied in **Supplementary Fig. 2d**.
- In Experiment 3, a threshold of a 5% false discovery rate was used to define DA peaks. No minimum fold change or minimum accessibility cutoffs were applied.
- In Experiment 4, peaks were ranked by p-values assigned by each DA method and lists of top-ranked peaks of equal size were compared (n = 1,000 in the main text, 500 or 5,000 in **Supplementary Fig. 5**). No p-value cutoffs, multiple testing correction, minimum fold change or minimum accessibility cutoffs were applied.
- In Experiment 5, the AUCC was calculated as in Experiments 1 and 2, but after applying a series of increasingly stringent fold change thresholds (see also our discussion below in response to the following point).
- In Experiment 6, the AUCC was calculated as in Experiment 1 and the number of false discoveries was calculated as in Experiment 3.
- In Experiment 7, the AUCC was calculated as in Experiments 1 and 2.

Most of the issues relate to experiment 5 (fold change filtering). Here the authors have used a method of ranking cells based on fold change and the lower X% of cells. This is not an approach that would be used in practice, and would provide poor results in cases where all the peaks have low fold change or high fold change. Different cutoff values should be chosen instead (for example, 1-fold. 2-fold, 3-fold). This was stated in the original review for the registered report.

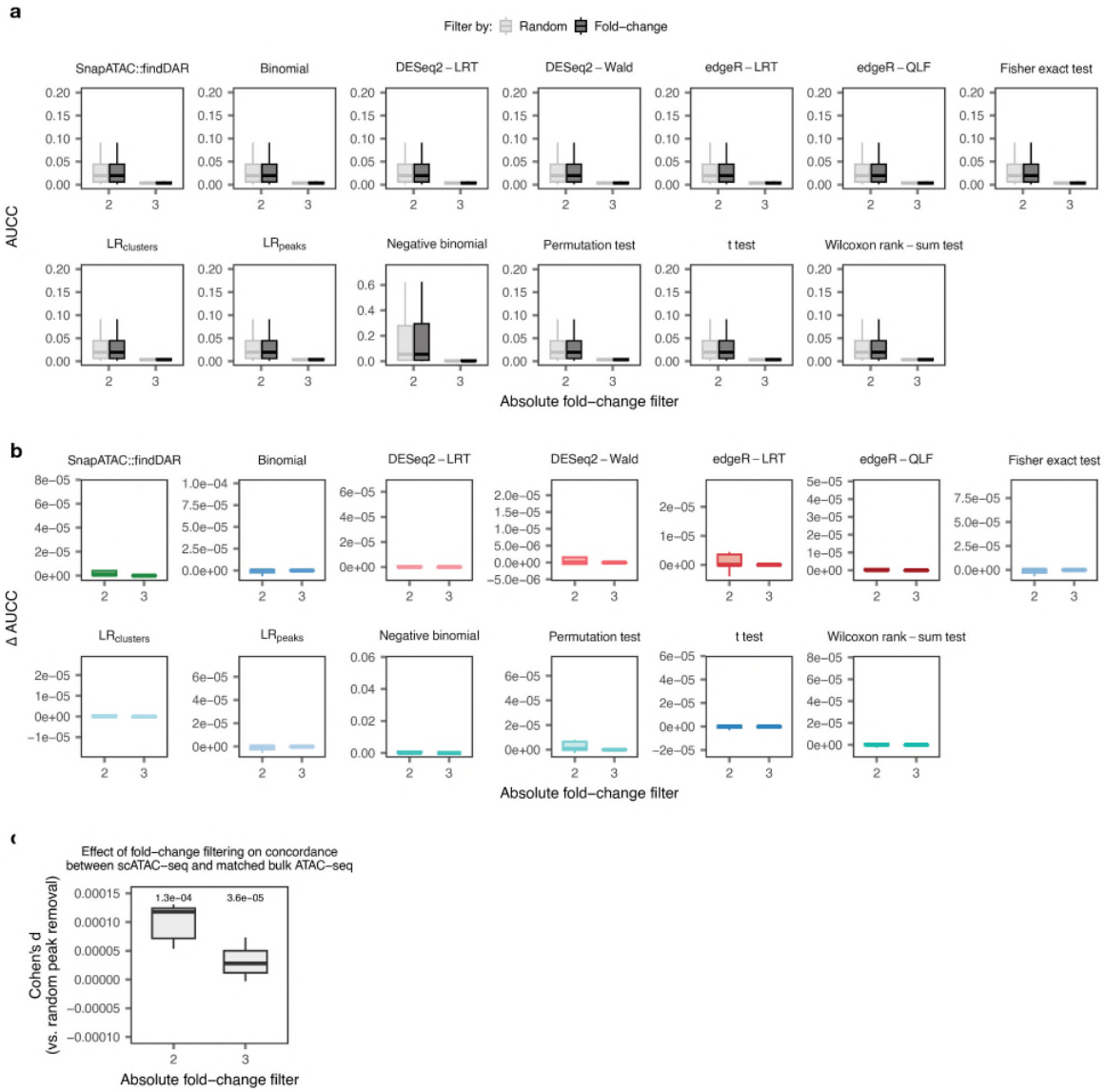
We were very surprised to read this comment. In response to the reviewer's comments on our Stage 1 submission, we had clearly articulated exactly how we planned to address the comments raised by the reviewer in their initial review of our proposed analyses. Our responses were read by the reviewer, who replied: "The authors have fully addressed all of my comments on the proposal. I appreciate the thorough and thoughtful revision, and look forward to reading the results of the study." In this Stage 2 submission, we have conformed exactly to the plan that was articulated in this proposal. We have carefully compared the Stage 2 submission and the accepted Stage 1 manuscript to identify any discrepancy, and can identify no deviation from the analyses that we had proposed.

It is unfortunate that the reviewer has now developed the opinion that the analyses proposed in our Registered Report should be altered. It seems to us that revising the prespecified analysis methodology at this point is both against the spirit of the Registered Report format and, more pointedly, explicitly against journal policy. Specifically, this policy stipulates that the Stage 1 submission must contain "[...] a description of experimental procedures in sufficient detail to allow another researcher to repeat the methodology exactly, without requiring further information", and that "[...] these procedures must be adhered to exactly in the subsequent experiments." Furthermore, journal policy is that "[...] the outcome of all registered analyses must be reported in the manuscript." This precludes the removal of our prespecified analyses at this juncture.

Nonetheless, we understand that it is permitted to add new, *post hoc* experiments or analyses at this stage, so long as they are clearly indicated as such in the text and documented in a section labeled "Exploratory analysis." Therefore, in an effort to address the reviewer's comment as best as possible, we have now performed the additional requested analyses. We filtered peaks on the basis of 2-fold or 3-fold change thresholds as requested by the reviewer (1-fold change is equivalent to no filtering) and repeated each of the experiments shown in **Fig. 5** and **Supplementary Figs. 9-11**. The results of these experiments were consistent with the prespecified analyses presented in the main text: that is, we found that log-fold change filtering based on a fixed threshold did not consistently improve, and frequently decreased, the accuracy of DA analysis.

We incorporated these results in the newly-added "Exploratory analyses" section at the end of the Results, and in **Supplementary Figs. 30-32**, each of which is reproduced below:

Second, we observed that filtering peaks by log-fold change between conditions did not consistently improve, and frequently decreased, the accuracy of DA analysis (**Fig. 5**). Our preregistered analysis involved removing equal proportions of peaks within each dataset on the basis of a percentile threshold. The question arose as to whether the imposition of a fixed fold-change threshold would affect our conclusions. Therefore, we repeated these analyses when filtering peaks with less than a two- or three-fold change between conditions. In these experiments, we again observed that log-fold change filtering did not improve the concordance of DA analyses of matched single-cell and bulk ATAC-seq (**Supplementary Fig. 30**), or between scATAC-seq and scRNA-seq from the same cells (**Supplementary Fig. 31**), and often substantially increased the number of false discoveries (**Supplementary Fig. 32**), all of which are consistent with the results of our initial, prespecified analyses.

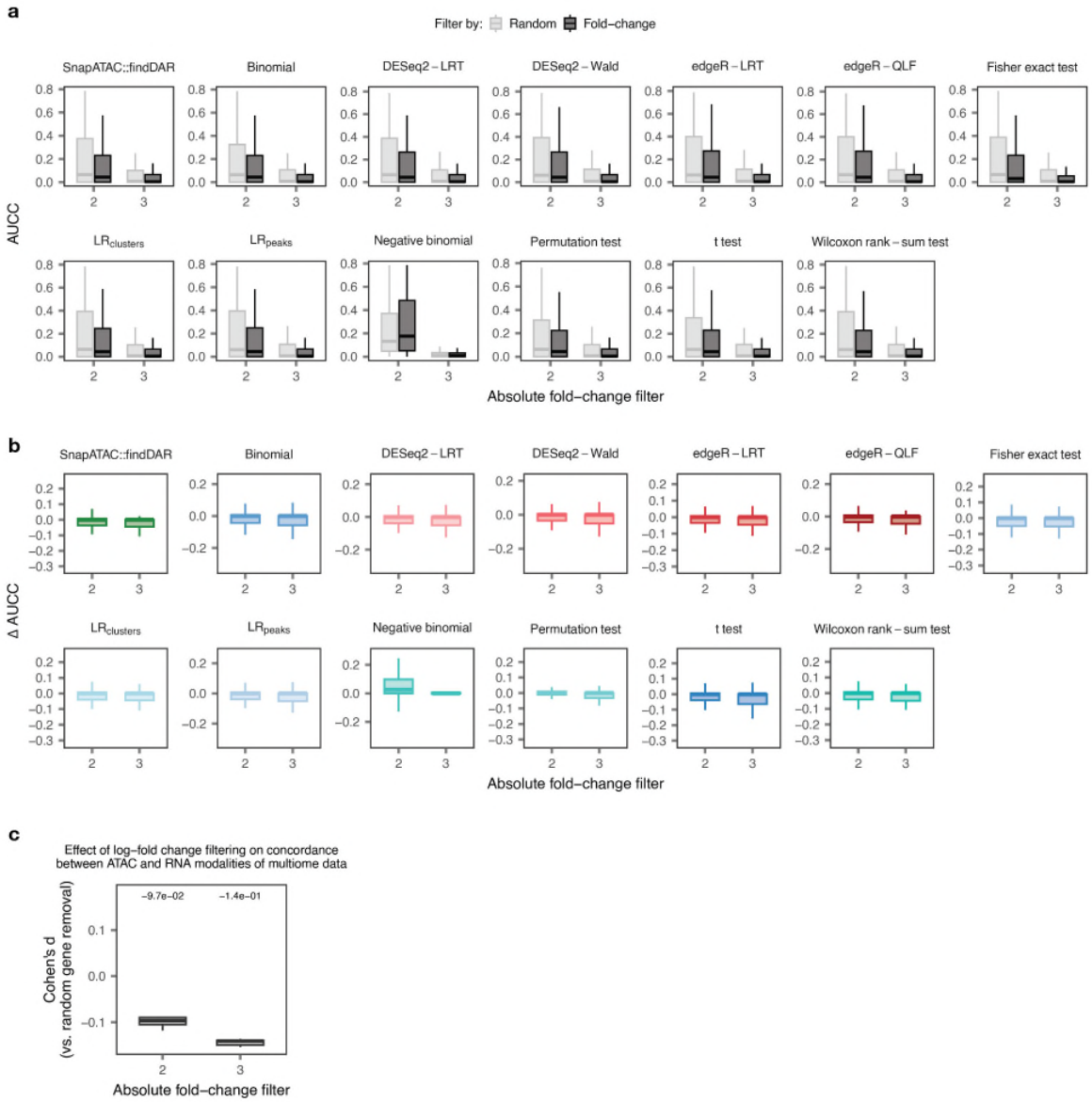


Supplementary Fig. 30 | Exploratory analysis: impact of log-fold change filtering on single-cell DA analysis with fixed thresholds, matched bulk ATAC-seq.

a, Area under the concordance curve (AUCC) for single-cell DA methods in Experiment 1, using matching bulk ATAC-seq as a reference, after filtering peaks with less than a two- or three-fold change between conditions or an equivalent number of randomly selected peaks.

b, As in **a,** but showing the difference in the AUCC (ΔAUCC) between filtering by fold change and filtering randomly selected peaks.

c, Effect size (Cohen's *d*) of increasingly stringent fold change filtering on the AUCC, relative to the removal of an equivalent number of peaks selected at random. Inset text shows the median Cohen's *d*.

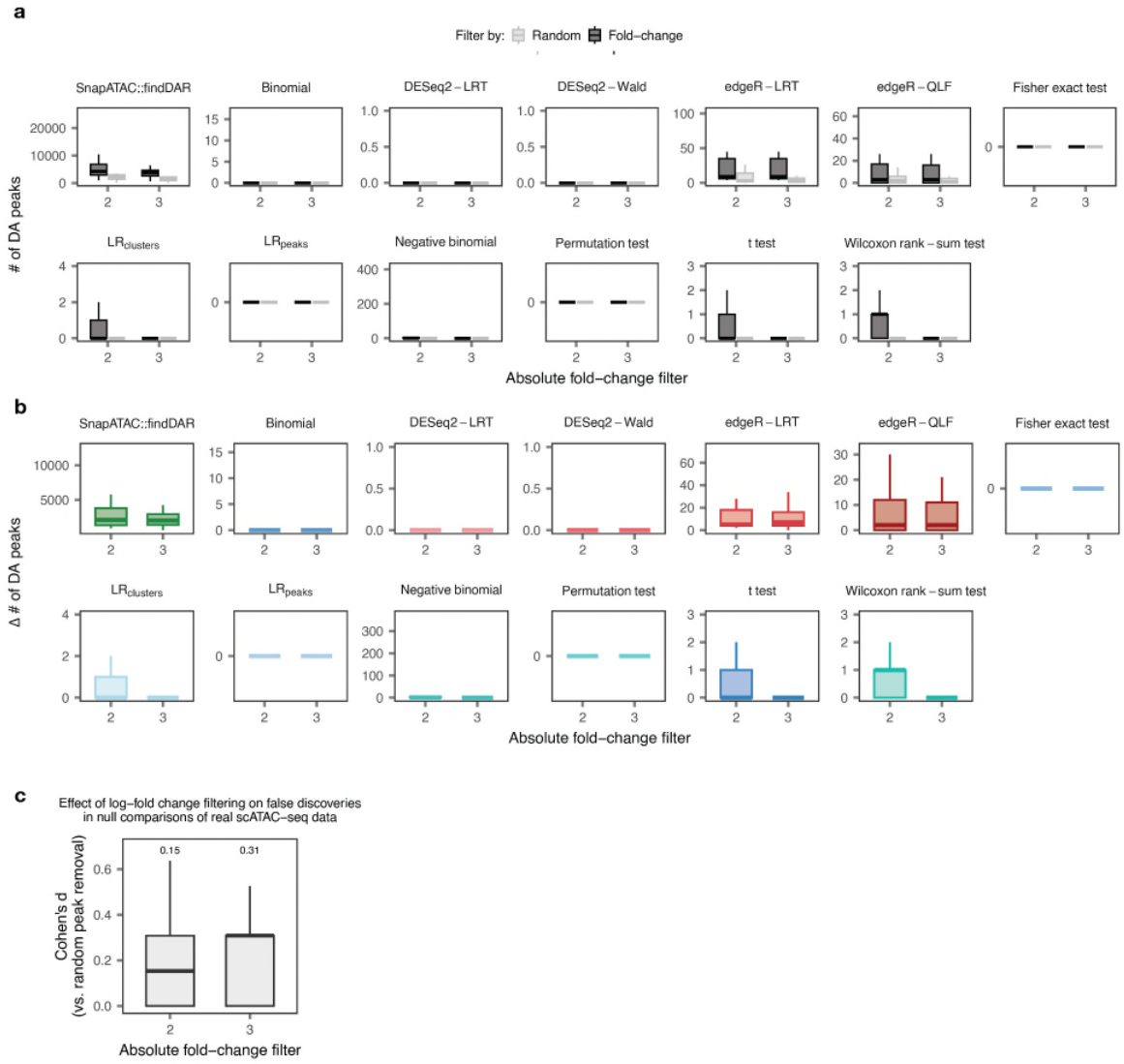


Supplementary Fig. 31 | Exploratory analysis: impact of log-fold change filtering on single-cell DA analysis with fixed thresholds, multiome.

a, Area under the concordance curve (AUCC) for single-cell DA methods in Experiment 2, using matching snRNA-seq as a reference, after filtering regions with less than a two- or three-fold change between conditions or an equivalent number of randomly selected regions.

b, As in **a,** but showing the difference in the AUCC (ΔAUCC) between filtering by fold change and filtering randomly selected regions.

c, Effect size (Cohen's *d*) of increasingly stringent fold change filtering on the AUCC, relative to the removal of an equivalent number of regions selected at random. Inset text shows the median Cohen's *d*.



Supplementary Fig. 32 | Exploratory analysis: impact of log-fold change filtering on single-cell DA analysis with fixed thresholds, false discoveries.

a, Number of DA peaks detected between randomly assigned replicates at 5% FDR within each cell type in random comparisons of published scATAC-seq data, after filtering peaks with less than a two- or three-fold change between conditions or an equivalent number of randomly selected peaks.

b, As in **a,** but showing the difference in the number of DA peaks between filtering by fold change and filtering randomly selected peaks.

c, Effect size (Cohen's *d*) of increasingly stringent fold change filtering on the number of false discoveries in null comparisons of published scATAC-seq data, relative to the removal of an equivalent number of regions selected at random. Inset text shows the median Cohen's *d*.

How fold changes are computed needs to be more thoroughly described. Different softwares used different implementations. The authors should provide the formula used in the methods section, and ensure that this is indeed what is being used in the software running the tests. Differences in, for example, the magnitude of the pseudocount or when the pseudocount is added can have a large impact on the fold change values returned. Log2 fold change should also be reported as this is more interpretable (rather than natural log).

We certainly agree that fold-change calculations differ between software implementations, and that this is an important consideration insofar as it has the potential to introduce a confounding variable between methods. For exactly this reason, we stated in our Stage 1 submission:

> [...] Moreover, log-fold change estimates differ across software implementations, most notably between DA methods that operate on individual cells versus on so-called pseudobulks. For this analysis, we will standardize log-fold change calculation for all DA methods using the code implemented in Signac.

We have conformed to this plan by using Signac to calculate log-fold changes (or, more precisely, using functions from Seurat; see also our response to the reviewer's related point below) in every analysis presented in the manuscript. This is also re-stated in the Methods section of the Stage 2 submission. We now additionally, and as requested in a separate point, state the precise versions of Signac/Seurat used in the manuscript (see also our response to the comment immediately below). The relevant section of the Methods now reads as follows:

> Impact of log-fold change filtering
> To study the impact of log-fold change filtering, we again made use of the matching bulk ATAC-seq and single-cell multi-omic datasets employed in Experiments 1 and 2, respectively. We binned peaks (genes) according to their deciles of absolute $\log_2$-fold change in the scATAC-seq data, and then removed the bottom 10% to 90% of peaks (genes) with the lowest log-fold change in each dataset. This procedure was repeated after removing an equal number of peaks from the dataset at random. We then computed the difference in the AUCC when filtering by log-fold change versus at random, and summarized the difference across DA methods using Cohen's d. Because log-fold change estimates differ across software implementations, we standardized log-fold change calculation for all DA methods using the code implemented in the Seurat function FoldChange[35]:
>
> $$\widetilde{X_1} = \{\widetilde{X_{1,g_1}}, \ldots, \widetilde{X_{1,g_m}}\}, \quad \widetilde{X_{1,g_j}} = \frac{1}{n_1}\sum_{i}^{n_1}(x_{g_j} + 1)$$
>
> $$\log FC = \log\left(e^{\widetilde{X_1}} - 1\right) - \log\left(e^{\widetilde{X_2}} - 1\right)$$
>
> Separately, we evaluated the effect of log-fold change filtering on the number of false discoveries in single-cell DA analysis. For this purpose, we repeated the random comparisons of published scATAC-seq data in Experiment 3, and compared the number of false discoveries returned by each DA method after log-fold change filtering versus random peak filtering.

Because we had indicated in our Stage 1 submission that we would report log-fold changes, we believe that journal policy surrounding Registered Reports would preclude us from altering the base of the logarithm can be altered at this point; however, we are certainly happy to make this minor change throughout the manuscript at the discretion of the editor.

In general, more methodological details are required. For example, how is FDR computed, how is TF-IDF computed. The authors should state the formula and functions (with parameters used), and software

versions. Notably there have been significant bugs in past versions of Seurat regarding the calculation of fold changes with scATAC data, so it is essential that versions are stated and the authors are using up-to-date software implementations.

We have extensively revised the Methods sections to respond to all of these points. Specifically, the revised Methods now provides the version numbers of all packages as well as specific functions that were called. To avoid reproducing the majority of the Methods section here, we refer the reviewer to the revised manuscript for the complete list of revisions. To respond to the specific points raised in this comment, FDR was computed using the base R function 'p.adjust' (with method = 'BH'), whereas TF-IDF normalization was performed using the 'RunTFIDF' function in Signac (version 1.15.1). We also take the opportunity here to highlight that we have made all of the analysis code used in this study publicly available via GitHub.

Figure 5 is hard to interpret in several ways. First, the X-axis states "# of cells" but a percentage is given along the axis. Based on the description in the methods, I believe this should be "% peaks removed" or something similar. Second, the legend states "Effect size (Cohen's d) of increasingly stringent log-fold change filtering on the AUCC between single-cell and bulk ATAC-seq DA." It's unclear what is actually being shown. The text also discusses the comparison of fold change filtering vs random filtering, but the figure compares fold change filtering to no filtering. The authors should rethink how to present the results from experiment 5 in a clear and compelling way. The most useful information to readers is: how does the false positive rate change for different fold-change filters, how does the concordance between bulk vs single-cell and RNA vs ATAC change with different fold-change filters, how does all of this compare to random removal of the same number of peaks.

The reviewer is correct to point out that the x-axis title was incorrect in the original figure. We revised the x-axis titles to "% of peaks (genes) removed." Moreover, we agree that all of the questions the reviewer raised—"how does the false positive rate change for different fold-change filters, how does the concordance between bulk vs single-cell and RNA vs ATAC change with different fold-change filters, how does all of this compare to random removal of the same number of peaks"—are interesting and important questions. Indeed, these are exactly the questions that the experiment in question was designed to address. Our presentation of the data in **Fig. 5** may have obscured this. As shown in the accompanying supplementary figures (**Supplementary Figs. 9a, 10a, 11a**), and described in the Methods and the Stage 1 submission, we have compared the AUCC (or number of false discoveries) when filtering peaks based on their log-fold change, versus removal of an equal number of randomly selected peaks. Unfortunately, the y-axis was incorrectly labelled "Cohen's d (vs. no filter)" whereas this should have read "Cohen's d (vs. random peak removal)." We corrected the y-axis title, the figure legend and additionally revised the titles of each figure panel. The new figure is reproduced below:

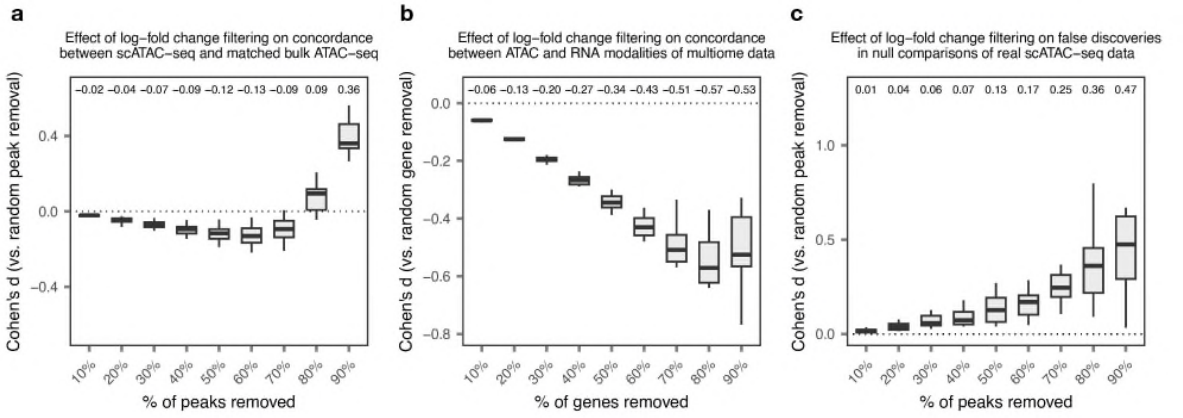


Fig. 5 | Impact of log-fold change filtering on single-cell DA analysis.
a, Effect size (Cohen's *d*) of increasingly stringent log-fold change filtering on the AUCC between single-cell and bulk ATAC-seq DA, relative to the removal of an equivalent number of peaks selected at random. Inset text shows the median Cohen's *d.*
b, As in **a,** but for the AUCC between the ATAC and RNA modalities of single-cell multi-omics data.
c, As in **a,** but for the number of false discoveries in null comparisons of published scATAC-seq data.

In the normalization section (experiment 6) it's stated that "We found that the number of false discoveries was consistently reduced after applying TP10K or TP(median) normalization (Fig. 6i-k and Supplementary Fig. 17b-d). However, this improvement in the control of false discoveries came at the expense of biological accuracy (Fig. 6h)." – this does not seem to be the case for TF-IDF and the t-test. Better panel labeling is also needed for figure 6. Several panels appear with the same axes and title, and the reader has to carefully look at the legend to work out what it's showing. Expressing all the results as Cohen's D here decreases the interpretability of the results. It is better to report, for example, the false positive rate itself (figure panel 6i, etc).

We respond to each point raised by the reviewer in order:

1. Our data shows that TP10K or TP(median) normalization both reduce the number of false discoveries, but this has either no effect or a negative effect on the biological accuracy, depending on the statistical test that is used downstream of normalization. With respect to the specific combination of TF-IDF with the t-test, this produces the opposite pattern: the number of false discoveries is increased, but so is the biological accuracy. We agree that we can flesh out our presentation of this data. The revised paragraph now reads as follows:

> We then asked whether specific approaches to normalization could mitigate the appearance of false discoveries. We found that the number of false discoveries was consistently reduced after applying TP10K or TP(median) normalization (**Fig. 6i-k** and **Supplementary Fig. 17b-d**). However, these two approaches did not improve, and in some cases decreased, the concordance between single-cell and bulk ATAC-seq data (**Fig. 6h**). On the other hand, the combination of TF-IDF normalization with the t-test, which increased the concordance between single-cell and bulk ATAC-seq data, also increased the number of false discoveries.

2. Thank you for pointing this out. We have revised the titles of each panel in **Fig. 6b-d** and **i-j** to remove the redundant axes/titles and clearly indicate which set of analyses are shown in each panel.

3. We agree that it is important to show the metrics that are the input to the calculation of Cohen's d (e.g., the number of false discoveries). These metrics are shown in the numerous supplementary figures that accompany **Fig. 6** (i.e., **Supplementary Figs. 14-26**). Unfortunately, the sheer volume of data in these figures does not permit us to include them in the main text, given that it occupies 8 supplementary figures, with most of these being a full page in size. We have tried to show the full distributions of these metrics in the main text where possible (e.g. **Fig. 6a-g, o-p**) but this is particularly difficult in the analyses specifically related to normalization, and therefore we have instead elected to summarize the effects of six normalization methods across three statistical tests in the main text in **Fig. 6h-n**.

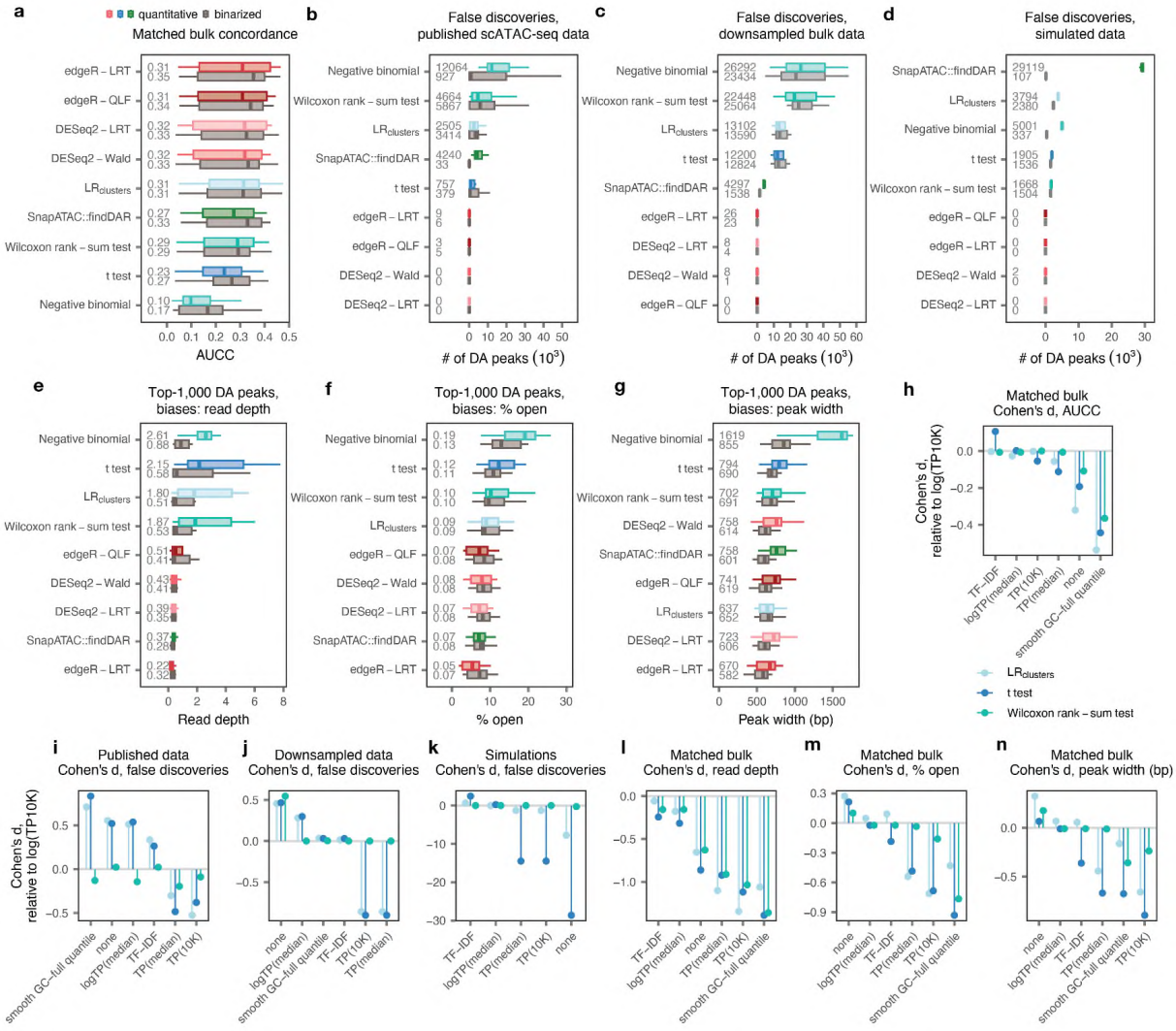


Fig. 6 | Best practices for scATAC-seq analysis.
a, Area under the concordance curve (AUCC) for single-cell DA methods using matching bulk ATAC-seq as a reference, before and after binarizing scATAC-seq data. Inset text shows the median AUCC.
b, Number of DA peaks detected between randomly assigned replicates at 5% FDR within each cell type in random comparisons of published scATAC-seq data, before and after binarization. Inset text shows the median number of DA peaks.
c, As in **b,** but showing the number of DA peaks detected at 5% FDR in DA analysis of downsampled bulk ATAC-seq libraries without biological differences between experimental conditions.

d, As in **b,** but showing the number of DA peaks detected at 5% FDR in model-based simulations of scATAC-seq data without biological differences between experimental conditions.
e, Mean read depth of the top-1,000 DA peaks identified by each single-cell DA method in published scATAC-seq datasets, before and after binarization. Inset text shows the median across comparisons.
f, As in **e,** but showing the proportion of cells in which these peaks are open.
g, As in **e,** but showing the width of each peak.
h, Effect size (Cohen's *d*) of different approaches to normalization of scATAC-seq data on the AUCC between single-cell and bulk ATAC-seq DA, relative to log-TP10K normalization.
i, As in **h,** but showing the number of DA peaks detected between randomly assigned replicates at 5% FDR within each cell type in random comparisons of published scATAC-seq data.
j, As in **h,** but showing the number of DA peaks detected at 5% FDR in DA analysis of downsampled bulk ATAC-seq libraries without biological differences between experimental conditions.
k, As in **h,** but showing the number of DA peaks detected at 5% FDR in model-based simulations of scATAC-seq data without biological differences between experimental conditions.
l, As in **h,** but showing the mean read depth of the top-1,000 DA peaks identified by each single-cell DA method in published scATAC-seq datasets.
m, As in **l,** but showing the proportion of cells in which these peaks are open.
n, As in **l,** but showing the width of each peak.
[...]

Other points:
- "analytical workflows implemented in different laboratories bear little resemblance to one another" – this is inaccurate. The workflows, while having differences, do fundamentally follow similar strategies of quantifying peaks, normalizing the data, defining groups of cells, and performing a statistical test.

This sentence is found within the Introduction, which we understand per journal policy cannot be revised at this point in the Registered Report review process ("Please note that following AIP, the Introduction section cannot be altered apart from correction of factual errors, typographic errors and altering of tense from future to past"). With that said, we agree with the reviewer that there are fundamental similarities in addition to important differences between workflows, and so we would be happy to make this change at the discretion of the editor.

- In figure 8 (scalability) the number of tests being performed should be stated

We have implemented this suggestion, and the revised figure legend now reads as follows:

Fig. 8 | Scalability of single-cell DA methods
a, Wall clock time required by each DA method to execute each comparison in Experiment 1 (*n* = 16). Inset text shows the median runtime in minutes.
b, Peak memory usage of each DA method while executing each comparison in Experiment 1 (*n* = 16). Inset text shows the median peak memory usage in GB.
c, As in **a,** but for each comparison in Experiment 2 (*n* = 306).
d, As in **b,** but for each comparison in Experiment 2 (*n* = 306).
e, Wall clock time required by alternative implementations of three DA methods (t-test, Wilcoxon rank-sum test, and negative binomial regression). Inset text shows the median runtime in minutes. ***, $p < 10^{-15}$, paired t-test.
f, As in **e,** but showing peak memory usage by each alternative implementation. Inset text shows the median peak memory usage in GB.

- In the scalability section, the differential testing functions are incorrectly attributed to the Signac package (they are in fact part of Seurat). Same is true for several other functions.

We have made edits throughout the Methods section to clarify the origin of functions implemented in Signac (e.g., RunTFIDF, CallPeaks) versus functions implemented in Seurat (e.g., FindMarkers, FoldChange).

Reviewer #2 (Remarks to the Author):

This registered report is well written and data analysis is well presented. Results indicate the advantage of standard bulkRNA-seq approaches (DEseq and EdgeR) in the overall problem. While results support this, it cannot be excluded that these conclusions arise from the fact DE analysis used to generate gold standards in most of the evolution scenarios.

This was discussed in previous rounds of review and is indicated in the limitation section. Interesting contributions to the literature include the analysis of best practices in scATAC-seq data, which indicates potential benefits of binarization, effects of normalization and peak calling. A standing critical point is the fact that majority of results do not use best practices (mostly adopted in the scATAC-seq literature and confirmed by some of the results) in their benchmarking. Authors needs to address this and discuss these points and describe potential limitations more explicitly.

Thank you again for your constructive comments and efforts to improve our manuscript.

Major points:

1. AUCC reported in Fig. 2 are overall quite low (maximum of 0.32) and indicate that most methods perform quite bad in the DA problem. Another explanation for this would be that this evaluation is sub-optimal, i.e. DA peaks in single cell data do not match the ones in bulk due to lower quality of either bulk or scATAC-seq experiments. This should be discussed in the manuscript.

Thank you for this comment. We agree that an AUCC of 0.32 superficially seems like a low value. The two possibilities for this observation suggested by the reviewer—namely, that the observed AUCCs may reflect poor performance of all DA methods, and/or technical differences between bulk and single-cell ATAC-seq—are both plausible. A third possibility is that this reflects a property of the metric itself, i.e., that two independent realizations of the same experiment would yield an AUCC substantially below 1.0 solely due to the kind of experimental variability that is characteristic of any high-throughput sequencing technology. Our data does not allow us to distinguish between these explanations, and it seems very plausible that all three may play a role. We additionally agree that it would strengthen the manuscript to highlight these considerations, and have done so in a paragraph added to the Discussion:

> In comparisons of scATAC-seq and matched bulk ATAC-seq data, we generally observed that the AUCC was less than 0.5, with no DA method achieving an AUCC greater than 0.52 in any primary or secondary analysis (**Supplementary Fig. 1c**). This observation could reflect the presence of technical differences between bulk and single-cell ATAC-seq, a universally poor performance of DA methods in either single-cell or bulk ATAC-seq data, or the properties of the AUCC metric itself. With respect to the final possibility, it is noteworthy that Soneson *et al.*[55] previously observed that, in the context of scRNA-seq, applying two different DE methods to the same dataset often yielded an AUCC substantially lower than 0.5. Our data do not allow us to distinguish between these potential mechanisms, and plausibly reflect a combination of all three possibilities.

2. Usually single cell ATAC-seq peaks are derived from the bulking the single cell data by considering either all or clustered cells. As my request, authors include an analysis on using the first evaluation (Fig. 2). Interestingly, reported results have significantly higher AUCCs (maximum of 0.42 and 25% improvement for all methods). This supports a higher quality of scATAC-seq data, as previously discussed in my revisions. Also, in contrary to the manuscript statement, the ranking of methods is distinct for panel Supp. 1b and Fig. 2b.

We have revised the sentence in question to emphasize that although the rankings of DA methods are broadly similar between **Fig. 2b** and **Supplementary Fig. 1d**, they are not identical:

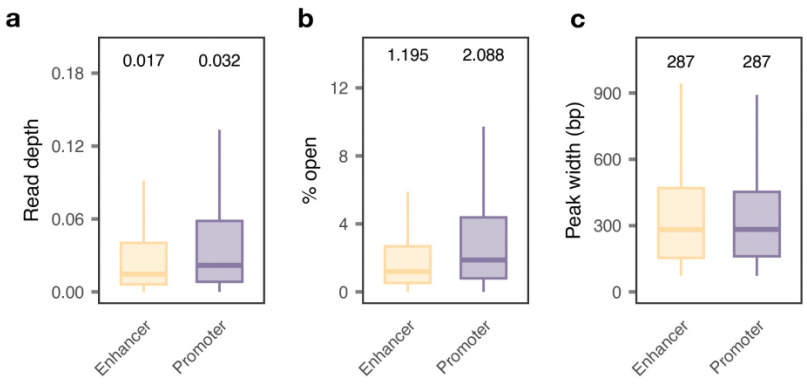
> [...] In all of the above analyses, we had called peaks in the matched bulk ATAC-seq data, but widely used software packages such as Signac[35], ArchR[33], and SnapATAC[34] instead call peaks in 'pseudobulk' samples created by pooling the single-cell data. We found that this procedure improved concordance between single-cell and bulk ATAC overall, but that the relative performance of single-cell DA methods remained similar although not identical (**Supplementary Fig. 1d**). [...]

3. Still on the same topic using promoters lead to a much higher AUCC (0.76) and very distinct method rankings. The reason besides this is not discussed in the manuscript.

We agree that this is an interesting difference. Our data do not allow us to establish a definitive causal relationship, but we observe that peaks in promoter regions differ in several important technical respects from those in enhancer regions: they tend to be open in a greater number of cells and supported by a greater number of reads. These observations are congruent with observations made in both bulk and single-cell ATAC-seq data, as noted in the Methods, and both would be expected *a priori* to improve the accuracy of DA analysis by increasing the signal-to-noise ratio.

Because this analysis was not preregistered, we have incorporated these results in the new "Exploratory analyses" section at the end of the Results, which is reproduced below along with the newly added **Supplementary Fig. 29**:

> Exploratory analyses
> During the peer review of this manuscript, the reviewers suggested additional analyses that had not been preregistered. In this section, we present the results of these *post hoc*, unregistered analyses.
> First, we found that the concordance between single-cell and bulk ATAC-seq data was substantially higher for peaks located in promoter regions than for those located in enhancers (**Supplementary Fig. 1f**). Whereas our data does not allow us to establish a definitive causal explanation for this phenomenon, we found that peaks in promoter regions tended to be open in a greater number of cells, and supported by a greater number of reads, as compared to promoter regions (**Supplementary Fig. 29**). These observations are congruent with observations made in both bulk and single-cell ATAC-seq data[2,32] and likely explain, at least in part, the greater biological accuracy of DA analysis for peaks in promoter regions.



Supplementary Fig. 29 | Exploratory analysis: technical characteristics of promoter vs. enhancer regions.

4. Given points 1, 2 and 3, authors should be more critical regarding their overall conclusion for this analysis. In one hand, methods could be ranked be considering all distinct scenarios. On the other hand, authors need to highlight results based on scenarios implementing best practices in of scATAC-seq.

Thank you for this suggestion. To address this, as well as related comments (points 5, 6, and 7), we have produced three additional summary figures that are included in the revised manuscript.

In the first, we have ranked DA methods according to their performance in every scenario in Experiments 1 and 2, including all of the sensitivity analyses shown in **Supplementary Figs. 1-3**, and produced a more detailed summary figure that is now included as **Supplementary Fig. 33**, and reproduced below. This figure more clearly exposes the degree of variability between DA methods across specific aspects of the design of these experiments highlighted by the reviewer, e.g., the approach used to call peaks. It also highlights that some DA methods consistently achieve very good or very poor performance: for instance, DESeq2-LRT ranks among the top tercile of DA methods in all but two scenarios, whereas negative binomial regression, the binomial test, and Fisher's exact test never rank among the top tercile in any scenario.

Supplementary Fig. 33 | Summary of DA method performance across Experiments 1 and 2, including primary and all sensitivity analyses.
Methods were grouped into terciles of high, low, or intermediate performance on the basis of their quantitative performance on each task, as in **Fig. 9**, and ranked by their average performance across all primary and secondary analyses shown in **Fig. 2** and **Supplementary Figs. 1-3**.

Second, the reviewer specifically highlighted the need to evaluate DA methods “based on scenarios implementing best practices in of scATAC-seq.” We understood this comment to refer to our observation that some DA methods performed better when specific binarization and/or normalization strategies were

employed; for instance, SnapATAC::findDAR demonstrated much better performance when binarizing counts than when aggregating quantitative accessibility measurements across cells. We agree that, broadly speaking, we might have designed our prespecified analyses differently had we known the outcomes of the best-practices experiments in advance. That we did not reflects the nature of preregistration and the Registered Report format more specifically. Nonetheless, to address this point, we produced two additional summary figure panels that specifically show the performance of DA methods in the 'best practices' experiments (Experiment 6). In the first of these (**Supplementary Fig. 34a**), we compared DA methods across each of the individual binarization and normalization scenarios shown in **Fig. 6** and **Supplementary Figs. 12-26**. In the second of these (**Supplementary Fig. 34b**), we selected only the single top-performing scenario for each DA method.

Supplementary Fig. 34 | Summary of DA method performance across Experiment 6, including all binarization and normalization strategies.
a, Methods were grouped into terciles of high, low, or intermediate performance on the basis of their quantitative performance on each task, as in **Fig. 9**, separately across each of the individual binarization and normalization scenarios evaluated in the study.
b, As in **a,** but showing only the single top-performing scenario for each DA method.

We present these new, non-prespecified analyses in the "Exploratory analyses" section of the Results:

> Third, the design of Experiments 1, 2, and 3 was preregistered before the results of Experiment 6 became available. In view of the observation that the performance of some DA methods benefited from

binarization and/or alternative approaches to normalization, the question arose as to how these preprocessing strategies would affect the summary of DA methods shown in **Fig. 9**. A similar question arose with respect to the observation that the concordance between single-cell and matched bulk ATAC-seq was generally higher when calling peaks in performance in 'pseudobulks' created from single-cell data. To explore these points, we produced additional summary figures that compared DA methods across all primary and secondary analyses in Experiments 1 and 2 (**Supplementary Fig. 33**) and across all binarization and normalization scenarios in Experiment 6 (**Supplementary Fig. 34**). These analyses corroborated the observation that some DA methods consistently achieved very good or very poor performance: for instance, in Experiments 1 and 2, DESeq2-LRT ranked among the top tercile of DA methods in all but two analyses, whereas negative binomial regression, the binomial test, and Fisher's exact test never ranked among the top tercile in any analysis. Considering all binarization and normalization approaches yielded rankings of DA methods that were broadly consistent with those shown in **Fig. 9**, but with some important differences: notably, SnapATAC::findDAR emerged as a top-performing DA method with respect to concordance with matched bulk ATAC-seq data, albeit not control of the false discovery rate, when applied to binarized data.

5. Points 1-4 potentially also impact the multimodal data analysis (Fig. 2d & e).
6. Results from the evaluation of Fig. 3 are stronger than Fig. 2, as some approaches can control for false positives (no false positive peaks). A question raised is how best practices (peak calling from scATAC-seq) impact this evaluation needs also to be addressed here. Indeed, it is shown in Fig. 6 that best practices (binarization) do impact of the performance (and ranking) of methods. Peak calling strategies in Fig. 3 are unfortunately not evaluated in this context. Again, it is unclear how their adoption would change rankings reported here. As in point 4. authors should have an overall ranking but also a ranking based on the adoption of the best practices.

We have addressed both of these points in the response immediately above.

7. The analysis of the impact of the FC is sound, but overall difficult to follow. As a (minor) suggestion authors could consider shortening it.

We have condensed this section, primarily by removing introductory material, as shown below:

Experiment 5: Impact of log-fold change filtering

~~In addition to the statistical significance of DA estimated by any given method, the difference in the number of reads overlapping a particular genomic region can be quantified by measures of effect size, most commonly the log-fold change.~~ Differential analyses of ATAC-seq data may discard differentially accessible regions with a log-fold change below an arbitrary threshold, on the grounds that regions with small fold changes are unlikely to be biologically relevant[44]. Similar practices are ubiquitous in the analysis of other sequencing-based technologies, such as RNA-seq, where the choice of log-fold change threshold may alter the biological interpretation of a given experiment[63]. These observations motivated an empirical assessment of the impact of log-fold change filtering on single-cell DA analysis. ~~Specifically, we repeated several of the analyses described in Experiments 1, 2, and 3 after imposing more or less stringent thresholds on the minimum log-fold change.~~

We recognized that the quantitative measures of performance evaluated in these experiments (AUCC or number of false discoveries) are sensitive to the total number of peaks being tested. This introduces a potential confounding factor, in that simply filtering peaks at random would also be expected to increase the AUCC and decrease the number of false discoveries. To account for this effect, for each log-fold change threshold, we tested the effect of removing an equal number of peaks from the dataset at random. We then used this data to determine whether filtering by log-fold change increased the AUCC or decreased the number of false discoveries to a degree greater than random.

We first examined the concordance of DA between scATAC-seq and matching bulk ATAC-seq. As expected, increasingly stringent log-fold change filters increased concordance for all DA methods. However, the same effect was also seen when removing equivalent numbers of random peaks (**Supplementary Fig. 9**). When controlling for ~~the effect of~~ random peak filtering, we observed that the concordance initially decreased

when filtering up to 70% of peaks based on log-fold change, and then increased when filtering 80% or more of peaks. To summarize these trends, we visualized the effect size of log-fold change filtering across all DA methods, compared to random peak filtering (**Fig. 5a**). ~~Because the magnitude of biologically relevant log-fold changes will expectantly vary across datasets depending on the biological system under study, this finding suggests it is difficult *a priori* to establish a log-fold change threshold that can increase the biological accuracy of DA analysis without risking a decrease in accuracy.~~

We next examined the concordance of DA between the ATAC and RNA modalities of single-cell multi-omics data. This analysis recapitulated the risks of log-fold change filtering. In this experiment, we found that log-fold change filtering consistently decreased concordance, relative to random filtering, when removing up to 90% of genes (**Fig. 5b** and **Supplementary Fig. 10**).

Last, we tested the impact of log-fold change filtering on the appearance of false discoveries. Relative to random filtering, we consistently observed more false discoveries across all DA methods when filtering by log-fold change (**Fig. 5c** and **Supplementary Fig. 11**).

8. The best practice analysis is an interesting aspect of this study. Authors should mention here the peak calling strategy (see point 2). The binarization aspect is also relevant, as this potentially contradict the previous literature. On the other hand, authors should consider the fact that previous work supporting the use of count data focus on the cell type identification problem and not DA analysis.

We revised the relevant paragraph of the Discussion to reiterate our findings with respect to peak calling, as shown below:

> Beyond comparisons of individual methods for single-cell DA analysis, we sought to use DA as a litmus test to identify best practices for the analysis of single-cell epigenomics data more generally. In some cases, these analyses supported clear recommendations. For instance, we found that filtering peaks by their log-fold change between conditions did not consistently improve DA analysis, relative to removing an equal number of peaks chosen at random. Moreover, in comparisons of approaches to the normalization of scATAC-seq data, we did not identify any method that consistently improved performance relative to log-TP10K normalization. We additionally observed a higher degree of concordance between DA analyses of matched bulk and single-cell ATAC-seq datasets when analyzing peaks called in 'pseudobulk' samples created by pooling the single-cell data, which supports the widespread application of this approach the analysis of scATAC-seq data. In other cases, our results were more ambiguous. [...]

We also are in full agreement with the reviewer's point that previous work focused on cell type identification, rather than DA analysis, and have revised the Discussion to explicitly highlight this discordance and the possibility that binarization may be helpful for some facets of scATAC-seq data analysis but not others:

> [...] Notably, we found that binarization reduced the biases of single-cell DA methods towards peaks accessible in a greater number of cells and to broader peaks, and generally increased the concordance between DA analyses of single-cell and matched bulk ATAC-seq data. Conversely, we found that the binarization did not always decrease (and sometimes considerably increased) the number of false discoveries. The improved biological accuracy of DA analysis when using binarized data as input is at odds with the argument that binarization unnecessarily discards quantitative information embedded within scATAC-seq data[31,32]. One possible interpretation of our data that could reconcile this apparent paradox is that binarization generally reduces the biases of DA methods towards more accessible and broader peaks, even for DA methods that are less affected by these biases in the first place, and that mitigating these biases can outweigh the potential negative effects of binarization, at least in the setting of DA analysis. Moreover, previous work primarily studied the impact of binarization on cell type identification within scATAC-seq data, and the data presented in this study does not formally exclude the possibility that binarization may be helpful for some facets of scATAC-seq data analysis but not others. Future work will be necessary to more firmly establish the mechanisms by which binarization may improve some aspects of scATAC-seq data analysis.

9. A major critical point of the study is the fact it does not consider the combination of DA methods with best practices (scATAC-seq peak calling; binarization, type of normalization). Indeed, in their literature research the authors only evaluated the use of DA methods, but not how the data was previously treated (binarization, normalization). As expected, these steps will impact each other, as shown in the manuscript. On the other hand, some of the best practices are already used by default by some of the evaluated methods. For example, snapatac is based on both scATAC-seq peak calling and binarization. Both these practices have a large impact in snapatac DA method results. Indeed, an evaluation considering the best practices (or best approach per DA method) will give a more realistic evaluation and relevant results for practitioners of scATAC-seq analysis.

First, it is not correct that our literature review did not evaluate how scATAC-seq data was preprocessed. **Fig. 1d** shows that over half of published studies included in our literature review (53%) binarized scATAC-seq data before analysis.

Second, we agree that our data underscore the impact of data preprocessing strategies on the biological accuracy and FDR control of DA analysis. As discussed in our response to point 4, above, the fact that our primary analyses are not based on the results of experiments that had not yet been performed is an unavoidable consequence of preregistration. Nonetheless, and at the request of the reviewer, we have included new summary figures that specifically highlight the interaction between data preprocessing and DA method performance in our data. Moreover, we specifically highlight the impact of binarization on SnapATAC's findDAR method at multiple points in the Results section of the manuscript.

10. Alternatively, authors need to clearly indicate limitation of their study, whenever conclusions are based on non-best practices.

With our responses to points 4, 5, 6 and 7 above, and the modifications to the Results and Discussion and the addition of **Supplementary Figs. 33-34**, we believe that we have addressed this point.

Minor

Authors should make clear which strategy is being used for each evaluation (how data are normalized or not, how peaks are called). A matrix in the supplement would be extremely helpful.

Thank you for this suggestion. This information was provided in the Methods section, but we agree that it would also be helpful to readers to have the option to access the same information in tabular format. We have added an additional supplementary table, **Supplementary Table 3**, in the revised manuscript that clearly indicates the peak-calling strategy, normalization approach, and binarization for each main text and supplementary figure panel in the manuscript. This table is cited in the Methods as follows:

> The specific combination of input features (e.g., peaks called from bulk ATAC-seq data versus from 'pseudobulks' of single-cell data), normalization approaches, and binarization that are shown in every figure or supplementary figure panel in the manuscript is provided in **Supplementary Table 3**.

Page 20, "under the hypothesis that bulk ATAC would yield higher quality set of peaks". This could be revised, as the evaluation do not support this.

We revised the sentence in question to indicate that this was our hypothesis at the time that we formulated our preregistered analyses (i.e., at the time of the Stage 1 submission):

[...] In our primary analysis, we obtained such a peak set by calling peaks in the matched bulk ATAC-seq data, under the prespecified hypothesis that analyzing bulk data would yield a higher-quality set of peak definitions. [...]

Reviewer #3 (Remarks to the Author):

In this manuscript, the authors address an exciting, yet problematic analysis applied to single-cell chromatin accessibility profiling data: detecting and identifying differential active elements. While these types of analyses in bulk-derived data have well-understood best practices, their behavior in single cell data has not been thoroughly characterized. As the authors importantly state, it is still unresolved to what extent data from individual cells is actually quantitative. Here, the authors perform exhaustive comparisons of differential analysis methods applied to single-cell ATAC-seq data in an attempt to define a best-practice.

Overall, I find that this manuscript would be useful for the field, however, in its current state, its salient point (ie., defining a best practice) gets lost in the details/technicalities. In addition, a major shortcoming (and missed teaching opportunity) is that the authors do not discuss why certain methods work better, and importantly, when their use is appropriate given experimental design choices/realities.

Thank you again for your careful review of our manuscript and your constructive feedback.

Major points:

1) Section "Experiment 2" comparing DA in sc-ATAC and sc-RNA (mulitome analysis) doesn't seem to add anything outright useful to the manuscript. One problem is that I am not sure how one defines a good concordance between the two modalities. The connectivity between regulatory elements (accessible peaks) and genes is largely unknown and no current method seems to do be particularly successful (for example, distance and the popular ABC model seem to perform roughly similarly). I understand the hypothesis that if a gene expression is changed than it is likely driven by changes in the cis-regulatory elements within some "regulatory domain", however, the definition of these domains are arbitrary (as is the approach taken in this manuscript). As the main goal of this paper is evaluating which methods are best for detecting differentially accessible peaks, the analyses in this section seem to be a distraction.

We agree that comparing DA in scATAC-seq data with DE in scRNA-seq data from the same cells does not constitute a ‘gold standard’ benchmark, and specifically that there are a number of mechanisms by which changes in promoter accessibility may not be mirrored by changes in gene expression (and vice-versa). We did acknowledge this in the “Limitations” section of our Discussion, writing:

> […] our comparison to multiome data was based on the premise that genes that are DE across biological conditions will also tend to have DA promoters, but exceptions to this assumption are known, notably during differentiation[68]. The majority of the comparisons that we propose entail the comparison of two fully differentiated cell types, mitigating this concern to some degree.

Notwithstanding these limitations, which are acknowledged in the text, we disagree that this analysis does not add anything useful to the manuscript. In fact, the conclusions from this analysis broadly corroborate those from our comparison of matched single-cell and bulk ATAC-seq data, which provides an additional measure of confidence in these analyses.

More pointedly, removal of preregistered analyses would both directly contradict journal policy for the Registered Reports format and also undermine the primary goal of preregistration, which is to prevent the selective publication of results.

Therefore, to address this comment, we revised the Discussion to more clearly address additional limitations of this analysis. The paragraph in question now reads as follows:

> Our study also has a number of limitations. First, our comparisons to parallel bulk ATAC-seq or multiome data leveraged high-throughput assays carried out on matching samples (or even cells) under identical experimental conditions, but a limitation of these analyses is the requirement of the use of statistical methods for DA and DE analysis in these parallel data sets. We carried out sensitivity analyses to test the impact of leaving out any given bulk DA or DE method, which supported the robustness of our conclusions. However, we cannot formally exclude the possibility that these comparisons favor single-cell DA methods that yield output more similar to that of other methods in general, for technical rather than biological reasons. Second, our comparison to multiome data was based on the premise that genes that are DE across biological conditions will also tend to have DA promoters, but exceptions to this assumption are known, notably during differentiation[68]. The majority of the comparisons that we propose entail the comparison of two fully differentiated cell types, mitigating this concern to some degree. Third, our approach to quantifying gene-level accessibility involved aggregating reads around the transcription start site, which required the specification of an arbitrary window within which reads were considered to be associated with promoter accessibility. Fourth, the relative paucity of datasets with paired bulk ATAC-seq data from identical biological conditions means that the datasets to be used in Experiment 1 comparisons involve comparisons both between and within cell types. However, these are comparisons with different underlying biological motivations, and in particular, the effect sizes of comparisons between cell types are likely to be larger than those within cell types and across conditions.

2) The authors apply/evalute all of the methods but do not comment on their statistical underpinnings (pros and cons for different applications/use cases). For example, the non-psuedobulking approaches are destined to perform poorly because they make assumptions of about background distribution of the raw count signal that a largely not reflected in the data (for example the data is overdispersed with respect to the binomial distribution). A discussion of statistical methods and assumptions behind these different approaches will help the average user/reader build an intuition about the general applicability and appropriateness of different methods for their particular application.

We agree with the reviewer that some of the methods compared in this study make questionable statistical assumptions about the distribution of scATAC-seq data. However, we do not believe that our data support the conclusion that this is the primary factor that determines the biological accuracy or control of the false discovery rate in DA analysis. Instead, our data strongly suggest that controlling for biological and technical variation across replicates is the single most important factor that determines the accuracy of DA analysis in scATAC-seq data. Approaches that explicitly account for variation across replicates were consistently the top-performing DA methods in terms of their ability to mirror DA analyses of matched bulk ATAC-seq data or DE analyses of scRNA-seq data from the same cells (**Figs. 1** and **2**). On the other hand, approaches that do not account for variation across replicates spuriously identify thousands of differentially accessible peaks in null comparisons of published scATAC-seq data and in simulated datasets (**Fig. 3**). Moreover, we show that even with suboptimal statistical approaches, the appearance of false discoveries can be dramatically reduced in experiments with a sufficiently high number of replicates (**Supplementary Fig. 4b**). Perhaps most critically in view of this specific comment, the secondary analysis presented in **Supplementary Fig. 4e** demonstrates that applying the exact same statistical model with or without a fixed effect term for each replicate dramatically alters the control of the false discovery rate. Importantly, this analysis compares two models that make identical assumptions about the underlying distribution of the count data, but which lead to very different conclusions.

To address this comment, we revised the first paragraph of the Discussion to explicitly highlight this point:

Discussion

The increasingly broad adoption of technologies to interrogate the epigenome at single-cell resolution has exposed a lack of consensus on how to analyse the resulting data. In this study, we formulated and implemented a series of preregistered analyses that aimed to systematically compare statistical methods for DA analysis of epigenomics data. We carried out an extensive review of the literature to identify all of the DA methods that had been applied in published single-cell studies, and established an epistemological framework that would enable a comparison of these methods with respect to their biological accuracy and their propensity to produce false discoveries. These analyses established that statistical methods that aggregated individual cells to form 'pseudobulks' generally yielded DA results that were better-aligned with orthogonal measurements of the same biological systems (i.e., matching bulk ATAC-seq or matching RNA-seq from the same cells) than methods that treated individual cells as the biologically relevant unit of observation. Our data further suggested that these differences in performance could be attributed, at least in part, to the appearance of false discoveries in DA analysis of individual cells. In null comparisons of published scATAC-seq data, and in a series of simulation studies, we found that the most widely used DA methods identified thousands of DA peaks in the absence of any underlying biological differences. These false discoveries preferentially arose from wider peaks and peaks accessible in a greater number of cells, mirroring the tendency for false discoveries to arise from highly-expressed genes in single-cell data[49], and were abrogated both by pseudobulk DA methods and by mixed-effects models that incorporated replicate as a random effect. Remarkably, models that made identical assumptions about the underlying distribution of the count data varied markedly in their ability to avoid false discoveries, depending on whether and how they controlled for biological and technical variation across replicates. Together, these findings suggest that single-cell DA methods must account for technical or biological variation between replicates in order to enable accurate DA analysis and avoid a proliferation of false discoveries.

3) Related to my comment above, it seems that the non-psuedobulk methods are commonly used because single-cell studies rarely perform biological or technical replicates (despite their weaker performance). Given the cost associated (and 10X genomics monopoly on the market) with single-cell studies, I doubt this will change anytime soon. As such, I think that some more attention/focus should be placed on how to perform DA analyses when no replicates are available or are limited (or demonstrate the importance of replicates).

As discussed above, our data strongly suggest that controlling for biological and technical variation across replicates is the single most important factor that determines the accuracy of DA analysis in scATAC-seq data. Approaches that explicitly account for variation across replicates were consistently the top-performing DA methods in terms of their ability to mirror DA analyses of matched bulk ATAC-seq data or DE analyses of scRNA-seq data from the same cells (**Figs. 1** and **2**). On the other hand, approaches that do not account for variation across replicates spuriously identify thousands of differentially accessible peaks in null comparisons of published scATAC-seq data and in simulated datasets (**Fig. 3**). The secondary analysis presented in **Supplementary Fig. 4e** demonstrates that applying the exact same statistical model with or without a fixed effect term for each replicate dramatically alters the control of the false discovery rate. Moreover, we found that even with suboptimal statistical approaches, the appearance of false discoveries can be dramatically reduced in experiments with a sufficiently high number of replicates (**Supplementary Fig. 4b**).

All of these results collectively underscore that accounting for variation across replicates is a critical consideration in single-cell DA analysis. Conversely, we do not believe that our data supports any conclusions about how to analyze datasets with a single replicate. From a broader perspective, it is questionable whether it is ever appropriate to draw biological conclusions from a single sample: this kind of experimental design makes it impossible to distinguish biologically relevant effects from biological or technical variation.

On the other hand, we believe our work supports strong conclusions on the issue of how to perform DA analyses when replicates are *limited.* For instance, all of the datasets that we analyzed Experiment 1 profiled between n = 2 and 4 replicates per condition, and all of our simulations use n = 3 per condition by default (with additional sensitivity analyses of more or fewer replicates in the supplement). We regret that this point was not clear and have addressed this by now providing the number of replicates per condition for every comparison in the paper in the new **Supplementary Table 2**. We reference this supplementary table in the Results section, which provides an opportunity to highlight the number of replicates in these studies:

> Experiment 1: Evaluating single-cell DA methods with matched bulk data
> We first used our assembled compendium of matched bulk and single-cell ATAC datasets to evaluate the biological accuracy of each single-cell DA method, using the bulk data as a reference. Our survey of the literature identified five studies in which matching single-cell and bulk epigenomics data were collected from the same populations of purified cells and sequenced within the same laboratory (**Fig. 2a**). These studies collected between two to four scATAC-seq libraries per condition (**Supplementary Table 2a**). We performed DA analysis of both the bulk and single-cell ATAC-seq datasets, [...]
>
> Experiment 2: Evaluating single-cell DA methods with single-cell multi-omics
> We next identified four studies that employed multi-omic assays to quantify both gene expression and chromatin accessibility across tens of thousands of nuclei, each profiling between two and 13 replicates (**Supplementary Table 2b**), and leveraged these datasets to repeat the comparisons of single-cell DA methods described in Experiment 1, now using gene expression as the reference (**Fig. 2c**). [...]

Additionally, we have revised the Discussion to place a greater degree of emphasis on the implications of our data for experimental design of single-cell epigenomics studies, in the newly added paragraph below:

> More broadly, our data emphasizes the importance of sound experimental design for single-cell epigenomics studies. Even suboptimal statistical methods for DA analysis, which produced thousands of false discoveries in comparisons involving small sample sizes, could achieve good control of the false discovery rate in comparisons of ten samples per experimental condition. Conversely, increasing the number of cells profiled from a small number of replicates only exacerbated the appearance of false discoveries. Moreover, we observed that the number of false discoveries correlated with the intensity of batch effects between replicates, as expected. Together, these observations highlight the importance of (i) minimizing batch effects during data collection, (ii) profiling a sufficient number of samples, and (iii) using appropriate statistical approaches to analyze the resulting data.

4) Generally, I find the figures overly technical and it is hard to parse the immediate conclusion(s). Personally speaking, I don't think it is necessary to show all method contrasts in each figure. In addition (or alternatively) I might group the methods by some of their commonalities (psuedobulk, etc.)

We regret that the reviewer felt that the figures were overly technical or hard to parse. We believe that this is, to some extent, unavoidable: the sheer volume of data generated in our preregistered analyses is substantial, as reflected in the 28 supplementary figures that accompany the main-text figures. More pointedly, journal policy precludes the removal of any prespecified analyses altogether at this point. A compromise might involve showing only a subset of the methods in the main text and the rest in the supplementary figures, as the reviewer suggests; however, we suspect that this may be confusing rather than clarifying, as it would require the reader to compare results across two separate figures for every experiment presented in the main text. Last, comments by the other reviewers on our Stage 2 submission made the opposite request, i.e., that we show *more* data in the main-text figures. Therefore, we have limited changes to the figures to specific points that were felt to be unclear or require revision.

With respect to the second point raised in this comment, we did indeed group methods by salient commonalities: methods that aggregate cells from each replicate to form ‘pseudobulks’ (edgeR and DESeq2) are shown in red, whereas methods that aggregate cells across replicates (SnapATAC::findDAR) are shown in green and methods that do not aggregate information across cells are shown in blue. We revised the legend of **Fig. 2** to clarify this decision:

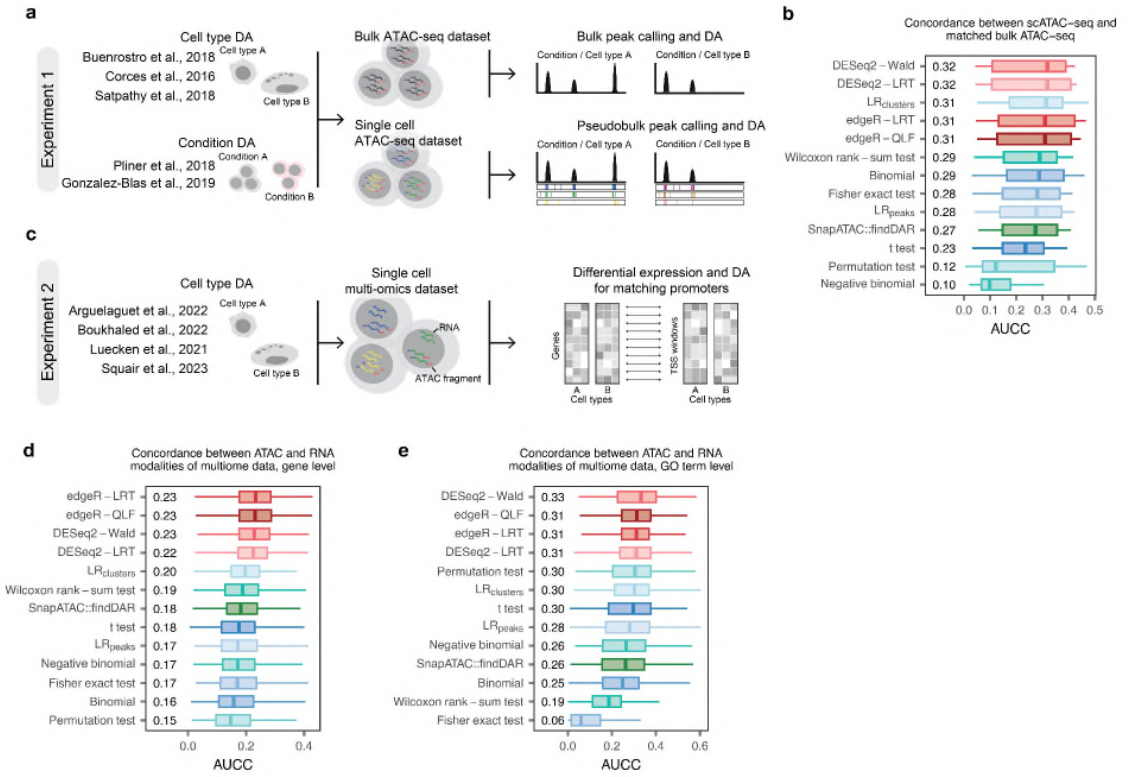


Fig. 2 | Evaluating single-cell DA methods with matched bulk data and single-cell multi-omics.
a, Design of Experiment 1. DA analysis was performed between cell types or conditions for single-cell datasets with matching bulk ATAC-seq as a reference. Peaks were called either in the bulk ATAC-seq data (primary analysis) or in pseudobulk single-cell data (sensitivity) analysis).
b, Area under the concordance curve (AUCC) for single-cell DA methods in Experiment 1, using matching bulk ATAC-seq as a reference. Inset text shows the median AUCC. Methods that aggregate counts within replicates to form ‘pseudobulks’ are shown in shades of red; one method that aggregates counts across replicates is shown in green; and methods that do not aggregate information across cells are shown in blue.
c, Design of Experiment 2. DA analysis was performed between cell types using matched scRNA-seq data from the same cell as a reference, comparing DA of genomic intervals around the TSS to gene-level differential expression.
d, Area under the concordance curve (AUCC) for single-cell DA methods in Experiment 2, using matching snRNA-seq as a reference. Inset text shows the median AUCC.
e, As in **d,** but showing concordance at the level of GO terms enriched among DA peaks.

Minor points:

1) The text could be substantially condensed for readability.

Thank you for this suggestion. We agree that readability is certainly an important consideration; however, the other reviewers have commented that we were insufficiently nuanced in our presentation of some results and therefore suggested lengthening the Results and Discussion in numerous places. Given the journal policy that we cannot remove any prespecified analyses at this point, it is difficult to see how the manuscript can be substantially shortened. We did specifically condense our description of the log-fold change analyses in Experiment 5, which now reads as follows:

> Experiment 5: Impact of log-fold change filtering
> ~~In addition to the statistical significance of DA estimated by any given method, the difference in the number of reads overlapping a particular genomic region can be quantified by measures of effect size, most commonly the log-fold change.~~ Differential analyses of ATAC-seq data may discard differentially accessible regions with a log-fold change below an arbitrary threshold, on the grounds that regions with small fold changes are unlikely to be biologically relevant[44]. Similar practices are ubiquitous in the analysis of other sequencing-based technologies, such as RNA-seq, where the choice of log-fold change threshold may alter the biological interpretation of a given experiment[63]. These observations motivated an empirical assessment of the impact of log-fold change filtering on single-cell DA analysis. ~~Specifically, we repeated several of the analyses described in Experiments 1, 2, and 3 after imposing more or less stringent thresholds on the minimum log-fold change.~~
>
> We recognized that the quantitative measures of performance evaluated in these experiments (AUCC or number of false discoveries) are sensitive to the total number of peaks being tested. This introduces a potential confounding factor, in that simply filtering peaks at random would also be expected to increase the AUCC and decrease the number of false discoveries. To account for this effect, for each log-fold change threshold, we tested the effect of removing an equal number of peaks from the dataset at random. We then used this data to determine whether filtering by log-fold change increased the AUCC or decreased the number of false discoveries to a degree greater than random.
>
> We first examined the concordance of DA between scATAC-seq and matching bulk ATAC-seq. As expected, increasingly stringent log-fold change filters increased concordance for all DA methods. However, the same effect was also seen when removing equivalent numbers of random peaks (**Supplementary Fig. 9**). When controlling for ~~the effect of~~ random peak filtering, we observed that the concordance initially decreased when filtering up to 70% of peaks based on log-fold change, and then increased when filtering 80% or more of peaks. To summarize these trends, we visualized the effect size of log-fold change filtering across all DA methods, compared to random peak filtering (**Fig. 5a**). ~~Because the magnitude of biologically relevant log-fold changes will expectantly vary across datasets depending on the biological system under study, this finding suggests it is difficult *a priori* to establish a log-fold change threshold that can increase the biological accuracy of DA analysis without risking a decrease in accuracy.~~
>
> We next examined the concordance of DA between the ATAC and RNA modalities of single-cell multi-omics data. This analysis recapitulated the risks of log-fold change filtering. In this experiment, we found that log-fold change filtering consistently decreased concordance, relative to random filtering, when removing up to 90% of genes (**Fig. 5b** and **Supplementary Fig. 10**).
>
> Last, we tested the impact of log-fold change filtering on the appearance of false discoveries. Relative to random filtering, we consistently observed more false discoveries across all DA methods when filtering by log-fold change (**Fig. 5c** and **Supplementary Fig. 11**).

2) Figure 5: it appears from my reading of the text that the x-axis should be "% percent filtered peaks" and not "# of cells"

This has been corrected in the revised figure:

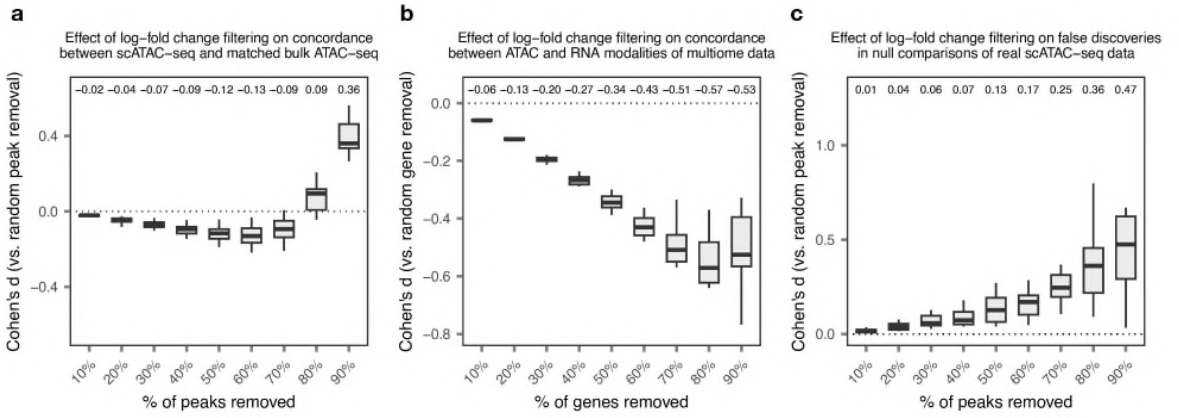


Fig. 5 | Impact of log-fold change filtering on single-cell DA analysis.
a, Effect size (Cohen's *d*) of increasingly stringent log-fold change filtering on the AUCC between single-cell and bulk ATAC-seq DA, relative to the removal of an equivalent number of peaks selected at random. Inset text shows the median Cohen's *d.*
b, As in **a,** but for the AUCC between the ATAC and RNA modalities of single-cell multi-omics data.
c, As in **a,** but for the number of false discoveries in null comparisons of published scATAC-seq data.

Reviewer #1 (Remarks to the Author):

The authors have addressed my comments to the extent that is allowed within the limitations of the registered report format

Reviewer #2 (Remarks to the Author):

The manuscript addresses a relevant problem of differential accessibility analysis in scATAC-seq data. During this rather extensive review rounds, several interesting insights raised from this complex manuscript. Unfortunately, due to the format of the paper as a pre-registered report, some of these did not make to the main figures or were note integral to the main benchmarking analysis.

These include:

- The fact that pseudo-bulk peak calling on scATAC-seq is beneficial (Supp. Fig. 1d).
- The impact of mixed models in controlling false positive discovery (Supp. Fig. 4e), as an alternative to pseudo-bulking approaches.
- The final rankings of methods considering distinct normalization and binarization strategies (Supp. Fig. 34)

Format issues apart, authors have addressed most of my requests, but some issues remain open.

Specific points

Authors should expand the discussion (in the explorative analysis) regarding the relation between read depth (Fig. 4a), detection performance in peak promoters/enhancers (Supp. Fig. 1F) and distinct number of reads in enhancers and promoters (Supplementary Fig. 29). It is likely that researchers doing scATAC-seq are more interested in characterizing enhancers than promoters. It is unclear if this can be inferred from summary analysis reported in Fig. 9 and Supp. Fig. 34.

Quantiles reported in Fig. 9 (Experimental Bias) do not apparently reflect results shown in Fig. 4. Please clarify this.

Reviewer #3 (Remarks to the Author):

The authors have satisfactorily addressed my concerns and I have no substantial concerns about the technical merit of this registered report.

My only standing (and non-technical) comment is that, in its current form, the primary figures in

the article are too dense. To be clear, I think the thoroughness of this work is extraordinary and nothing should be removed from the overall manuscript. However, I think incorporating some simple visual cues into the figures, removing redundancy of the text and moving some aspects to the supplement will help the average reader digest the general conclusions at higher-level before diving into the details.

Reviewer #1 (Remarks to the Author)

The authors have addressed my comments to the extent that is allowed within the limitations of the registered report format

We thank the reviewer again for their careful review of the manuscript and constructive feedback.

Reviewer #2 (Remarks to the Author)

The manuscript addresses a relevant problem of differential accessibility analysis in scATAC-seq data. During this rather extensive review rounds, several interesting insights raised from this complex manuscript. Unfortunately, due to the format of the paper as a pre-registered report, some of these did not make to the main figures or were note integral to the main benchmarking analysis.

These include:

- The fact that pseudo-bulk peak calling on scATAC-seq is beneficial (Supp. Fig. 1d).
- The impact of mixed models in controlling false positive discovery (Supp. Fig. 4e), as an alternative to pseudo-bulking approaches.
- The final rankings of methods considering distinct normalization and binarization strategies (Supp. Fig. 34)

Format issues apart, authors have addressed most of my requests, but some issues remain open.

We thank the reviewer again for their careful review of the manuscript and constructive feedback.

Specific points

Authors should expand the discussion (in the explorative analysis) regarding the relation between read depth (Fig. 4a), detection performance in peak promoters/enhancers (Supp. Fig. 1F) and distinct number of reads in enhancers and promoters (Supplementary Fig. 29). It is likely that researchers doing scATAC-seq are more interested in characterizing enhancers than promoters. It is unclear if this can be inferred from summary analysis reported in Fig. 9 and Supp. Fig. 34.

Our data shows that promoters are typically supported by more reads than enhancers; that DA analysis is generally more accurate for promoters; and that the accuracy of DA analysis continues to benefit from deeper sequencing up to at least 10,000 fragments per cell. If we have correctly understood this comment, the reviewer is suggesting that these observations raise the possibility that a higher sequencing depth would specifically improve DA analysis of enhancers. This possibility was not experimentally tested in our study; however, we agree this is a logical hypothesis that is worthy of discussion. We now raise this possibility in the “Exploratory analyses” section, as requested:

> First, we found that the concordance between single-cell and bulk ATAC-seq data was substantially higher for peaks located in promoter regions than for those located in enhancers (**Supplementary Fig. 1f**). Whereas our data does not allow us to establish a definitive causal explanation for this phenomenon, we found that peaks in promoter regions tended to be open in a greater number of cells, and supported by a greater number of reads, as compared to promoter regions (**Supplementary Fig. 29**). These observations are congruent with observations made in both bulk and single-cell ATAC-seq data[2,32] and likely explain, at least in

part, the greater biological accuracy of DA analysis for peaks in promoter regions. When considered in combination with the observation that the accuracy of DA analysis continues to benefit with increased sequencing depth up to at least 10,000 fragments per cell (**Fig. 7a**), these observations raise the possibility that increased sequencing depth would specifically benefit DA analysis of enhancers, although this possibility was not directly tested in the present study.

Quantiles reported in Fig. 9 (Experimental Bias) do not apparently reflect results shown in Fig. 4. Please clarify this.

We revised the Methods to clarify this point:

Biases of single-cell DA methods
[...]
For the summary plot shown in **Fig. 9**, DA methods were binned into terciles according to the mean rank of their top-1,000 DA peaks across all three properties (i.e., mean read depth, percentage of cells in which the peaks are accessible, mean width in base pairs).

Reviewer #3 (Remarks to the Author):

The authors have satisfactorily addressed my concerns and I have no substantial concerns about the technical merit of this registered report.

My only standing (and non-technical) comment is that, in its current form, the primary figures in the article are too dense. To be clear, I think the thoroughness of this work is extraordinary and nothing should be removed from the overall manuscript. However, I think incorporating some simple visual cues into the figures, removing redundancy of the text and moving some aspects to the supplement will help the average reader digest the general conclusions at higher-level before diving into the details.

We thank the reviewer again for their careful review of the manuscript. We agree that the manuscript presents a large volume of data. However, it is challenging to square this request with the remaining suggestions from reviewer 2 to the opposite effect, i.e., to move figures from the supplement into the main text and continue to expand the text. We feel the manuscript in its present form strikes a reasonable balance between comprehensiveness and accessibility. We also note that the main text figures now include textual cues that directly indicate the hypothesis being tested in each panel, which we think has improved the clarity, particularly for the very large, multi-panel figures (e.g. **Fig. 6**).