nature portfolio

Peer Review File

MOCHA’s advanced statistical modeling of scATAC-seq datae nables functional genomic inference in large human cohorts

Reviewer #1 (Remarks to the Author):

This work provides an interesting and useful tool for analyzing scATAC-seq data. The major suggestions are: 1) the performance evaluation can be statistically improved by considering false positive discoveries and specificity; 2) the DAA/CAA performance evaluation of ArchR and Signac should be also based on the LRM outcome for a more informative comparison; 3) providing comparison/benchmarking for the network analysis and the longitudinal analysis.

Major concerns:
1. Design of the positive control overlaps with the performance comparison. The ground truth was generated using MACS2 on the pseudo-bulk data from the scATAC-seq data and the MOCHA LRM model was trained according to such ground truth data. Then the performance of MOCHA was compared with MACS2. This may cause potential confounding. Could the authors try various approaches of ground truth generation to mitigate confounding?
2. Sensitivity. The authors claimed that, because MOCHA identified more tiles as positive tiles, it is more sensitive. However, sensitivity is not defined as how many samples are predicted as positive, but how many positive samples are missed. Also, sensitivity is only meaningful when specificity is discussed at the same time, since a model that predicts every sample as positive will have 100% sensitivity but is not meaningful. Therefore, the claim of higher sensitivity is not supported by the results, and even MOCHA did show higher sensitivity, whether it was meaningful or not was not analyzed in the context of specificity.
3. Generalizability of the LRM model. Just wonder whether the parameters are shared across all i, j, and t? Or for every sample, cell type, and tile, a unique set of parameters needs to be learned? In the Method section, NK cells from the COVID dataset were mentioned as the training set. Did this mean that the parameters learned from NK cells were used for predicting accessibilities of other datasets and other cell types?
4. The loss function of the LRM model was not provided. Also no justification was provided for the necessity of the parameter \theta^{n}. Often if the loss function is suitable and the two classes in the training set are well-balanced, the "default" threshold, which is 0.5, is sufficient.
5. Unbalanced ground truth for training a logistic model. The positive and negative labels look extremely unbalanced (1.15M vs 750M). Could the authors discuss the impact of unbalanced classes on the model performance?
6. Zero inflation is not considered in the LRM component. The authors justified those technical dropouts in the scATAC-Seq data lead to unreliable results, and a major advantage of MOCHA is to use zero inflation correction to address this issue. The inputs for the prediction of tile accessibility using the LRM model are data with dropouts. However, zero inflations were not considered in LRM. Meanwhile, the inputs of the differential accessibility analysis and the co-accessibility analysis are fragments on tiles that are predicted to be accessible by LRM. Could the authors provide evidence and justifications on why zero inflations (dropouts) are not crucial for LRM, which is directly affected by dropouts, but are crucial for the differential accessibility analysis? Would it make more sense to handle zero inflation in the LRM component instead of in the DAA and CAA components?
7. Zero inflation: could the authors delve in more about the patterns of technical dropouts across cell types and samples to check whether dropouts were random? This would provide biological insights why zero inflation correction contribute to better performance.
8. Runtime (Fig 2h): It looks that MACS2 performs better when sampled cells > 50k. The authors argued that MOCHA performed better in practical cases. However, with the fast growth of the single cell omics data and the increasing need of analyzing samples in the contexts of large data repositories such as HuBMAP, Human Cell Atlas, and NCI Human Tumor Atlas Network, more practice cases will involve over 50k cells. So the performance of MOCHA is not as scalable as the other two methods. Also, could the authors provide insights into why MOCHA performs better in the range of 50 – 50k cells?
9. Potential positive false discoveries. Suppl Fig 3c suggested that MOCHA slightly identified more CTCF and TSS tiles than the other methods and the difference is marginal. However, MOCHA significantly identified more tiles. This raises the concern that whether the tiles MOCHA identified are more likely to be false discoveries.
10. The comparison of the DAA and CAA components in MOCHA with ArchR and Signac is entangled by the effects of the LRM model, since the input of ArchR and Signac are raw data, and the input of MOCHA DAA and CAA is based on the MOCHA LRM and thus involved more tiles. Therefore, the contribution of the zero-inflation correction in MOCHA DAA and CAA as well as the

DAA and the CAA models in MOCHA cannot be clearly evaluated. Could the authors provide the DAA and the CAA of ArchR and Signac with the predicted assessable tiles using LRM?
11. Using the DAT annotation as the ground truth (Fig 2 b), the sensitivities of MOCHA for intragenic and distal regions are lower than ArchR and Signac. The overall sensitivity of the three methods is comparable. Specificity has not been compared, which makes it challenging to know whether the sensitivity is meaningful. Since MOCHA identified significantly more tiles (6,211) than Signac (1,266), but the sensitivity is comparable, likely the specificity of Signac is much better than MOCHA. In summary, according to Fig 2 b, it is likely that Signac significantly outperforms MOCHA.
12. The "Networks of alternatively regulated genes in early SARS-CoV-2 infection" and the "Longitudinal analysis of chromatin accessibility during COVID-19 recovery" sections are interesting but lacking comparisons with other approaches to demonstrate the performance of MOCHA. Could the authors indicate new discoveries that were previously not possible but now made possible by MOCHA? For example, how about comparing with state-of-arts approaches? Or, if the authors do not plan to claim novelty of these two components, how about comparing with LRM vs MACS2/HOMER and MOCHA vs ArchR/Signac to demonstrate what new or different knowledge or regulatory network and longitudinal patterns can be learned with MOCHA's novel upstream models?
13. The authors have recently published a comprehensive platform PALMO on Nature Communications (https://www.nature.com/articles/s41467-023-37432-w) – and congratulations to the authors! Since the datasets and the functionalities of PALMO and MOCHA are partially overlapping, both tools were developed by the same lab (that is, the authors of this manuscript should be aware of PALMO), and PALMO was published 4 months ago before the submission of this manuscript (and was preprinted in Oct 2022), a comprehensive comparison of functionalities and performance between PALMO and MOCHA would help audience to understand the new values of MOCHA in the context of PALMO. Also, PALMO should be mentioned in the Introduction besides in the Result section.

Minor concerns:
1. Improving the readiness of figures. Some visualizations can be further improved and some conventions can be considered. Here are a few examples:
a. Color coding. For example, in Fib 2a, maybe different colors should be used for CD4 CTL TEM and CD8 TEM?
b. Orders. Fig 2 a: if the same order of cell types was used, it could improve readability.
c. Scales. Fig 2a, Fig 2b, etc. – using the same scales would help the comparisons.
d. Numbers: Fig 2b – if "400000" could be visualized as "400,000", it would help. And it is inconsistent that, in the same Fig 2b, one subplot was visualized using "400000", and the other as "3e+5".
2. Method: sample preparation and data preparation info has been provided in the preprint and thus not necessary to be re-described in detail.

Reviewer #2 (Remarks to the Author):

In this paper, the authors developed a statistical approach, MOCHA to identify sample-specific cell-type open chromatin regions using scATAC-seq data. They tested MOCHA to multiple single cell datasets such as COVID19, immunology and Hematopoiesis and demonstrated its outperformances over existing methods for detecting sample-specific chromatin accessibility, differential accessibility in covid, and co-accessibility across samples. Moreover, using detected OCRs, the paper further inferred regulatory networks linking TF binding sites, ligands to TSSs, revealing possible alternative gene regulatory mechanisms and longitudinal dynamics in covid.

Overall, the study was designed with reasonable rationale. Identifying sample-specific activities of chromatin (co-) accessibility at the cell type level is an emerging topic, so MOCHA provides a timely statistical tool. The paper was well organized, and the results were presented logically. However, I still have some major concerns, especially about the rigor of the methodology and evaluation.
1. It is unclear if MOCHA is a general method or specific for COVID/immune study. If former, the paper needs to demonstrate broader applications.

2. The logistic regression model for evaluating accessibility needs further clarification. The authors claimed the usefulness of normalized total counts, lamda_1 and almost downstream analyses seem only use lamda_1. However, how important the max count (lamda_2) contributes to the regression? If also important, it remains elusive that lamda_2 is not used in downstream DAT and CAA analyses. Moreover, the study specific prefactor S was insufficiently described without justification.
3. When comparing with MACS2 and HOMER, the authors should also report functional or disease enrichments (e.g., LDSC) of MOCHA OCRs in addition to CTCF sites and TSSs, like their other sections did.
4. MOCHA randomly selected 50 DATs with two clusters by K-Means. How sensitive would its performance be to those hyperparameters? Also, K-Means is also not robust to outliers. This concern applies to many other parameters. The authors need to justify selecting parameters and provide guidelines for the users, especially biologists.
5. The networks that the paper predicted were not fully gene regulatory networks. They only linked TF binding sites, ligands to TSSs (near promoters) so missed other key regulatory mechanisms such as distant regulatory elements (enhancers from scATAC-seq data), gene expression relationships (e.g., co-expression from many methods for predicting gene regulatory networks like SCENIC).
6. For longitudinal analysis, it is unclear that cofounding factors (e.g., sex, age, etc) were considered for detecting chromatin accessibility dynamics.
7. The significant p-values were reported inconsistently thru the paper, e.g., p-value, adjusted p, FDR.

Reviewer #3 (Remarks to the Author):

The manuscript describes MOCHA, a method primarily for carrying out comparisons of single cell ATAC-seq data between groups of subjects. The method also includes features for identifying alternative transcription-start-site regulation, and transcription factor-gene network construction from longitudinal data. More rigorous ways of comparing scATAC-seq data sets are needed, however, it is not clear that MOCHA is making a substantial contribution for reasons given below.

1. The manuscript correctly observes that in single cell analysis comparisons between treatment and control groups should be done on the level of subject rather than cell, as treating cell level data as replicates would artificially inflate the significance of differences. The analysis proposed is therefore based the aggregation of single cell data into pseudobulk representations for different cell types. Differential analysis is then based on comparisons of sample level pseudobulk aggregates for the cell types. The idea of using pseudobulk to make comparisons of single cell data has been previously evaluated by Junttila et al, for example, who compared 18 methods for the identification of differential expression changes between conditions from multisubject scRNA-seq data. Many of the methods assessed by Juntilla et al could also be used to compare scATAC-seq data.

Junttila, Sini, Johannes Smolander, and Laura L. Elo. "Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data." Briefings in bioinformatics 23.5 (2022)

The authors should carry out more careful survey of single cell studies; relevant studies are not cited in the manuscript and there are likely to be many more.

2. Apart from the above mentioned scRNA-seq study, the MOCHA methodology is closely related to methods for bulk differential ChIP-seq or ATAC-seq peak calling, many of which are based on limma, DESeq2 or EdgeR. The following papers need to be cited. It is of critical importance that the methods described in these papers be included in benchmarking comparisons of differential accessible regions:

Gontarz, Paul, et al. "Comparison of differential accessibility analysis strategies for ATAC-seq data." Scientific reports 10.1 (2020)

Chen, Yang, Shue Chen, and Elissa P. Lei. "DiffChIPL: a differential peak analysis method for high-throughput sequencing data with biological replicates based on limma." Bioinformatics 38.17 (2022)

Stark, Rory, and Gordon Brown. "DiffBind: differential binding analysis of ChIP-Seq peak data." R package version 100.4.3 (2011)

Faux, Thomas, et al. "Differential ATAC-seq and ChIP-seq peak detection using ROTS." NAR Genomics and Bioinformatics 3.3 (2021): lqab059.

Qiu, Xintao, et al. "CoBRA: containerized bioinformatics workflow for reproducible ChIP/ATAC-seq analysis." Genomics, Proteomics and Bioinformatics 19.4 (2021)

3. Benchmarking of differential tiles between conditions is an important aspect of the paper and needs to be done rigorously. In the manuscript numbers of distinct genes and numbers of reactome pathways provide some anecdotal evidence that the method is working. However, a reliable gold standard of true differentially accessible regions is never established. To benchmark rigorously, gold standards need to be constructed to test both sensitivity and specificity. The benchmarking methodology used in the DESeq2 paper could be suitable for this. To test specificity, comparisons can be made of groups each containing a mixture of COVID+ and COVID- subjects, positives found in this analysis would be false positives. For sensitivity, the approach from the DESeq2 paper can be used: "we used experimental reproducibility on independent samples (though from the same dataset) as a proxy. We used a dataset with large numbers of replicates in both of two groups, where we expect that truly differentially expressed genes exist. We repeatedly split this dataset into an evaluation set and a larger verification set, and compared the calls from the evaluation set with the calls from the verification set, which were taken as truth." The authors might also consider the concepts introduced in:

Tian, Luyi, et al. "Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments." Nature methods 16.6 (2019)

Benchmarking in the current manuscript has not been done to an acceptable standard, and comparisons have not been made to the most relevant methods.

4. The dataset generated in this study involved numerous subjects and must have been done in several batches. Although batch effects are well known to impact single cell data such effects are not mentioned at all in the manuscript. It is important to provide the batch information and evaluate the degree to which batch effects could be influencing results. For example, when defining cell types do cells in different batches have similar chromatin accessibility, or do batches also define observed chromatin accessibility. A useful reference is:

Luecken, Malte D., et al. "Benchmarking atlas-level data integration in single-cell genomics." Nature methods 19.1 (2022): 41-50.

5. The MOCHA logistic regression model is used to create a matrix of accessibility on a sample by tile level. The procedure is used to collapse the single cell data into indicators of accessible tiles in given cell types and samples. The approach seems overly complicated in comparison with the DESeq/EdgeR/limma based methods, so comparisons will be important. Only lambda 2, the maximum number of fragments in a tile per cell seems to be truly related to single cell analysis. It is not clear how important this parameter is in the analysis and whether the need for this parameter could be obviated through simple filtering measures. For example, filtering identical fragments from the same cell or constraining the maximum number of fragments per tile per cell to 2. In addition, some description of what this parameter is achieving would be helpful.

6. Fig 2 shows total numbers of open tiles, and there is a threshold parameter that controls this number in MOCHA. The number of open tiles determined by MACS2 and HOMER could also be changed by altering cut-off parameters. In benchmarking it is not enough to define true accessible regions, as one can always get more tiles changing thresholds. Unless some way of showing

specificity is included, this analysis is not meaningful.

7. Line 893: "We used a previously published promoter-capture HiC (pcHiC) resource43 which identified promoter-enhancer regulatory links."

No justification is given for using HiC contacts as a gold standard for co-accessibility. First the manuscript should provide a motivation in terms of causality. The causal relationship between HiC and accessibility is not well understood and it is possible that chromatin accessibility causes HiC contacts rather than the other way round. Second, what HiC measures and its relationship to biology needs to be taken carefully into account. HiC is a protocol that measures, in some sense, proximity between genomic regions. It cannot be assumed that HiC precisely measures all the biologically relevant interactions between regions and only these. Third, the limitations in the specific HiC data will include some inaccuracy, limitations in sequencing depth, suboptimal experimental conditions etc. Overall, the case for HiC as a gold standard is not at all compelling.

8. The sections "Networks of alternatively regulated genes in early SARS-CoV-2 infection." and "Longitudinal analysis of chromatin accessibility during COVID-19 recovery" describe results from the COVID data generated in the project but do not evaluate methodology or make any comparisons with other methods. The manuscript notes that "An in-depth, comprehensive analysis of our COVID19 cohort is beyond the scope of current work and will be presented in a follow-up paper." It might be better to leave the longitudinal and gene network analyses for that paper.

9. Line 719 "we applied MACS2 37 ('-g hs -f BED --nolambda --shift -75 --extsize 150 --broad', '--model -n') to identify accessible peaks in the pseudobulk data, using previously published parameters for identifying peaks in scATAC-seq with the modification to call broad rather than narrow peaks.It is not clear why the --broad MACS2 option was used. Can some advantage be demonstrated? The –shift -75 –extsize 150 also doesn't seem to be well motivated.

10. The abstract doesn't describe the manuscript very well. The method seems to be primarily about analysing differential accessibility in multi-sample studies. The question of "proper" handling of technical dropout with zero-inflated methods, is highly debatable. Are the proposed heuristics proper handling? Identification of alternative transcription-starting-site regulation, and transcription factor–gene network construction from longitudinal scATAC-seq data are weak sections without benchmarking comparisons.

11. A complementary approach to single cell analysis is to carry out differential abundance testing. Comment on the relative strengths and weaknesses of the proposed approach relative to methods such as:

Dann, Emma, et al. "Differential abundance testing on single-cell data using k-nearest neighbor graphs." Nature Biotechnology 40.2 (2022): 245-253.

Reviewer #4 (Remarks to the Author):

This manuscript declared MOCHA, a tool to identify the gene regulatory programs when analyzing the scATAC-seq data. MOCHA exhibits the advantages in detecting differential accessible regions and chromatins than widely used tools including MACS2, HOMER, ArchR, Signac. The author also showed the good performance of MOCHA in the large dataset of COVID19 patients, and constructed ligand-TF-gene networks on alternative TSS regulations, which would be used to identify potential targets for COVID19 or other processes. And MOCHA can be integrated with exiting tools such as ArchR, chromVAR, as a valuable extension for analyzing scATAC-seq data. The following comments or issues need to be considered.
Comment：
1) Line105-107, MOCHA identifies sample- and cell type-specific open chromatin, within samples

from different experiment and batches. How to distinguish and balance the bias from the batch effect using the MOCHA?

2) During tiling the genome, MOCHA splits the genome into 500 bp tiles, all the analyses are based on the tiles. But this strategy has been adopted by the previously published SnapATAC. Then what are the differences and advantages of MOCHA compared to SnapATAC?

3) Follow the comment, does MOCHA eventually split the genome into 500 bp tiles? How about other ranges, such as 1 kb, 1.5 kb, 2 kb, 5 kb, which is better?

4) The authors stated that MOCHA is more sensitive in detecting open chromatin regions than MACS2 and HOMER, and detects more differential chromatin than ArchR and Signac. However, the splitting genome into 500 bp tiles could somehow cause potential bias by differences in data qualities when using MOCHA, which needs to be discussed.

5) Fig 4a-c: what’s the meaning of type I and type II sites in fig 4a? It seems that there is no difference between early infection patients and control donors in fig 4b-c? How to understand and calculate the Accessibility Change in Fig 4b-c?

6) Fig 4f and line 309-310 in page 8: The authors identified 122 ligands. Are these ligands regulated by all the differential motifs as shown in fig 4e?

7) Line 318-320: This reviewer couldn’t find the data demonstrated the regulation in CD16 monocyte.

8) How about the computer requirements to run the MOCHA?

9) page 13, row 504: An error labeling of “1x106 cells”.

Manuscript NCOMMS-23-29942-T
Response to Reviewers

We would like to thank all reviewers for their time and effort in reviewing our manuscript and providing insightful comments and feedback. We have incorporated their suggestions, clarified and improved the language where recommended, and added numerous additional analyses to address their concerns.

We provide a line-by-line response below, with

- the original reviewer comments in *italics*
- our responses in **bold**

For our revised manuscript, we have

- left our original text without modification,
- colored blue the added in-text modifications & revisions, and
- ~~striked through~~ text we have removed.

Reviewer #1 (Remarks to the Author):

This work provides an interesting and useful tool for analyzing scATAC-seq data. The major suggestions are: 1) the performance evaluation can be statistically improved by considering false positive discoveries and specificity; 2) the DAA/CAA performance evaluation of ArchR and Signac should be also based on the LRM outcome for a more informative comparison; 3) providing comparison/benchmarking for the network analysis and the longitudinal analysis.

Thank you! We appreciate the feedback and have provided clarifying comments where requested and several additional analyses to help address the concerns and recommendations provided by the reviewer. For the main suggestions, we have: 1) added simulation studies to address concerns regarding recall/sensitivity, Specificity (1-false positive rate, Supplemental Figure 8), and the F1-score (Harmonic mean of PPV, Recall), 2) clarified that the DAA/CAA performance evaluation of ArchR/Signac was based on the same LRM outcome (Fig 3), and 3) expanded benchmarking for differential analysis, network analysis, and longitudinal analysis (lines 174-175, 278-280, 311-314, 389-390), and Supplemental Figure 11,14).

Major concerns:
1. Design of the positive control overlaps with the performance comparison. The ground truth was generated using MACS2 on the pseudo-bulk data from the scATAC-seq data and the MOCHA LRM model was trained according to such ground truth data. Then the performance of MOCHA was compared with MACS2. This may cause potential confounding. Could the authors try various approaches of ground truth generation to mitigate confounding?

We agree that using peaks called by MACS2 as "ground truth" is not ideal, which we previously acknowledged in Discussion. We made two changes to address this concern.

First, we conducted a simulation study in which we simulated peaks in samples of various cell counts from simulated pseudo-bulked scATAC data and summarized MOCHA, MACS2, and HOMER's ability to detect those known peak locations (lines 200-203, 831-870, Supplementary Fig. 8).

Second, we highlighted review articles that recommend a minimum of 50 million reads for reliable open chromatin detection [1-2] (lines 123). With 15 times the minimum read depth (750 million reads total), our 'ground-truth' training set is highly robust, which is further supported by the agreement between simulated and real performance.

> *[1] Yan, F., Powell, D.R., Curtis, D.J. et al. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol 21, 22 (2020). https://doi.org/10.1186/s13059-020-1929-3*
> *[2] Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. 2015;2015:21.29.1–9.*

2. Sensitivity. The authors claimed that, because MOCHA identified more tiles as positive tiles, it is more sensitive. However, sensitivity is not defined as how many samples are predicted as positive, but how many positive samples are missed. Also, sensitivity is only meaningful when specificity is discussed at the same time, since a model that predicts every sample as positive will have 100% sensitivity but is not meaningful. Therefore, the claim of higher sensitivity is not supported by the results, and even MOCHA did show higher sensitivity, whether it was meaningful or not was not analyzed in the context of specificity.

We agree that our wording was not precise, and we removed the incorrect usage of the word 'more sensitive'. Additionally, we provided a more rigorous comparison of open chromatin identification, including sensitivity, specificity and false positive rates, and F1-score measurements, via our simulation studies in Supplementary Fig. 8 (lines 200-203, 754-762, 831-870) to provide a more comprehensive evaluation of performance.

3. Generalizability of the LRM model. Just wonder whether the parameters are shared across all i, j, and t? Or for every sample, cell type, and tile, a unique set of parameters needs to be learned? In the Method section, NK cells from the COVID dataset were mentioned as the training set. Did this mean that the parameters learned from NK cells were used for predicting accessibilities of other datasets and other cell types?

We agree that this was not clarified and thank the reviewer for pointing this out. The model was trained on NK cells only, across a range of cell numbers. The same parameters were applied to other cell types without further adjustment (Methods). A normalizing pre-factor is calculated and applied

before the model is applied to call open chromatin on data from new studies. We modified our results section to make this point explicit and clarify the generalizability (lines 114-118, 130-132). In addition, we conducted a mouse validation to show generalizability outside human studies (lines 227-232, Supplementary Fig 5).

4. The loss function of the LRM model was not provided. Also no justification was provided for the necessity of the parameter \theta^{n}. Often if the loss function is suitable and the two classes in the training set are well-balanced, the "default" threshold, which is 0.5, is sufficient.

We agree that model training was not sufficiently described, and thank the reviewer for highlighting that. We have modified our method section (see Training of LRM) to explicitly mention the logit link and the default loss function from the GLM logistic regression function in R (lines 735-736) the justification for the threshold parameter $\theta^{(n_a)}$, (lines 127-128, 744-746) that was necessary to address the imbalance in training classes when identifying open chromatin. Except for even performance on sensitivity & specificity, a threshold of 0.5 is not sufficient. Youden index is more appropriate & widely applied.

> **[1] Pepe, M. S. The Statistical Evaluation of Medical Tests for Classification and Prediction. (Oxford Univ. Press, 2003).**

5. Unbalanced ground truth for training a logistic model. The positive and negative labels look extremely unbalanced (1.15M vs 750M). Could the authors discuss the impact of unbalanced classes on the model performance?

We thank the reviewer for identifying an area that warrants more description. 1.15M is the number of open tiles, 4.39 million is the number of inaccessible tiles. In addition, 750 million is the number of total fragments in the NK population. So the imbalance is between 1.15M open tiles and 4.39M inaccessible tiles (not 1.15M vs 750M). Additionally, class imbalances are altered by downsampling depending on whether cells containing fragments in specific peaks are removed or not. This imbalance motivates the use of the cell-count dependent Youden Index. We have modified our Results (lines 127-128) and Methods (lines 744-746) sections to explain the class imbalances we observed, both from open vs closed regions, and the distribution of peak intensities.

6. Zero inflation is not considered in the LRM component. The authors justified those technical dropouts in the scATAC-Seq data lead to unreliable results, and a major advantage of MOCHA is to use zero inflation correction to address this issue. The inputs for the prediction of tile accessibility using the LRM model are data with dropouts. However, zero inflations were not considered in LRM. Meanwhile, the inputs of the differential accessibility analysis and the co-accessibility analysis are fragments on tiles that are predicted to be accessible by LRM. Could the authors provide evidence and justifications on why zero inflations (dropouts) are not crucial *for LRM, which is directly affected by dropouts, but are crucial for the differential accessibility analysis? Would it make more sense to handle zero inflation in the LRM component instead of in the DAA and CAA components?*

We thank the reviewers for their feedback on this point, and the need for greater clarity here. In response, we have extended the discussion section to include when and where zero-inflation statistics are critical (lines 431-433, 437-440). In brief, we make open chromatin calls using our LRM on each tile independently per cell type and sample. The LRM is only used on tiles with reads, which by definition is non-zero (lines 437-440). Zero-inflation is taken into account for downstream methods where we use data from a given tile across samples, where some samples have both zero and non-zero measurements.

7. Zero inflation: could the authors delve in more about the patterns of technical dropouts across cell types and samples to check whether dropouts were random? This would provide biological insights why zero inflation correction contribute to better performance.

We thank the reviewer for their excellent question. In response, we have added additional results, showing that our zero-inflated modeling functions can separate out both biological (Age, Sex) and technical (Cell counts) drop-outs in 3 cell populations: CD4 Naive T Cells, Naive B Cells and CD16 Monocytes (lines 364-366, Supplementary Fig. 18). A more comprehensive analysis on this interesting topic is certainly warranted but beyond the scope of the manuscript.

8. Runtime (Fig 2h): It looks that MACS2 performs better when sampled cells > 50k. The authors argued that MOCHA performed better in practical cases. However, with the fast growth of the single cell omics data and the increasing need of analyzing samples in the contexts of large data repositories such as HuBMAP, Human Cell Atlas, and NCI Human Tumor Atlas Network, more practice cases will involve over 50k cells. So the performance of MOCHA is not as scalable as the other two methods. Also, could the authors provide insights into why MOCHA performs better in the range of 50 – 50k cells?

We thank the authors for this comment and recognize that the text was not clear on this point. MOCHA was designed for identifying open chromatin by individual samples and cell types. In a given sample, most cell populations have < 5k cells. As a result, when calling open tiles in samples and cell types, MOCHA produced faster runtime in all applicable settings. We are not aware of a technology that can sequence 50k cells for a single cell type in a single scATAC sample after quality control. If/when that changes, we recognize that MACS2 would be faster under those settings. In our original benchmarking, we pooled cells across samples to simulate cell abundances beyond our current technical limitations. Since the previous analyses caused confusion and cellular abundances per sample do not typically exceed 5-10k, we modified our runtime analysis by varying the number of samples instead of the number of cells. This updated benchmark reflects what users would likely encounter when scaling to

larger datasets with more and more samples. We have updated Figure 2h and our results section (lines 234-237) to reflect these analyses. We moved the previous graph to Supplementary Fig. 4D.

9. Potential positive false discoveries. Suppl Fig 3c suggested that MOCHA slightly identified more CTCF and TSS tiles than the other methods and the difference is marginal. However, MOCHA significantly identified more tiles. This raises the concern that whether the tiles MOCHA identified are more likely to be false discoveries.

We agree that analyzing real data prevents us from calculating 'sensitivity' and 'specificity' in the absence of ground truth. To complement these analyses, we conducted an additional simulation study that enables us to properly compare false positives/negatives and model performances in the context of simulated ground truth. These simulation studies are found in Supplementary Fig. 8, and provide insights into sensitivity and specificity (1- false positive rates, lines 200-203, 754-762, 831-870). MOCHA had better F1 scores than MACS2 in all tested cases and HOMER when the cell number is 100 or above.

10. The comparison of the DAA and CAA components in MOCHA with ArchR and Signac is entangled by the effects of the LRM model, since the input of ArchR and Signac are raw data, and the input of MOCHA DAA and CAA is based on the MOCHA LRM and thus involved more tiles. Therefore, the contribution of the zero-inflation correction in MOCHA DAA and CAA as well as the DAA and the CAA models in MOCHA cannot be clearly evaluated. Could the authors provide the DAA and the CAA of ArchR and Signac with the predicted accessible tiles using LRM?

We recognize that we did not sufficiently describe the tile set used. All differential methods were benchmarked using the same predicted accessible tiles from the LRM. We updated the results section of the text so as to avoid future confusion for readers (lines 240-243).

11. Using the DAT annotation as the ground truth (Fig 2 b), the sensitivities of MOCHA for intragenic and distal regions are lower than ArchR and Signac. The overall sensitivity of the three methods is comparable. Specificity has not been compared, which makes it challenging to know whether the sensitivity is meaningful. Since MOCHA identified significantly more tiles (6,211) than Signac (1,266), but the sensitivity is comparable, likely the specificity of Signac is much better than MOCHA. In summary, according to Fig 2 b, it is likely that Signac significantly outperforms MOCHA.

We agree that Fig 3b was not clear and have updated it to reflect absolute tile numbers, with relative percentage next to it, and adjusted the figure and legend to better describe the figure and avoid unnecessary confusions. The percentage in Fig 3b reflects relative composition within a set of differential tiles, which is unrelated to a method's sensitivity. If % of promoters increase, then % of distal regions must decrease (and vice-versa), therefore these percentages cannot be used to claim specificity and sensitivity, rather the composition of the tiles.

12. The "Networks of alternatively regulated genes in early SARS-CoV-2 infection" and the "Longitudinal analysis of chromatin accessibility during COVID-19 recovery" sections are interesting but lacking comparisons with other approaches to demonstrate the performance of MOCHA. Could the authors indicate new discoveries that were previously not possible but now made possible by MOCHA? For example, how about comparing with state-of-arts approaches? Or, if the authors do not plan to claim novelty of these two components, how about comparing with LRM vs MACS2/HOMER and MOCHA vs ArchR/Signac to demonstrate what new or different knowledge or regulatory network and longitudinal patterns can be learned with MOCHA's novel upstream models?

We agree that benchmarking novel analytical frameworks is critical and added additional benchmarking when comparable methods exist. For alternatively regulated genes from Figure 4, we have added Supplementary Fig 14 that benchmarks MOCHA's performance with ArchR and Signac (lines 311-314). For Figure 5, we note that longitudinal data analyses with zero-inflation are not currently supported by either ArchR, Signac, or PALMO (lines 388-390, Supplemental Table 1). Instead we had previously benchmarked pseudo-time analyses using ArchR's gene scores with MOCHA's promoter tiles. Here we demonstrated that MOCHA-based results were more biologically informative and aligned better with the expected roles of CD16 monocytes than GeneScore-based ones (lines 357-362).

13. The authors have recently published a comprehensive platform PALMO on Nature Communications (https://www.nature.com/articles/s41467-023-37432-w) – and congratulations to the authors! Since the datasets and the functionalities of PALMO and MOCHA are partially overlapping, both tools were developed by the same lab (that is, the authors of this manuscript should be aware of PALMO), and PALMO was published 4 months ago before the submission of this manuscript (and was preprinted in Oct 2022), a comprehensive comparison of functionalities and performance between PALMO and MOCHA would help audience to understand the new values of MOCHA in the context of PALMO. Also, PALMO should be mentioned in the Introduction besides in the Result section.

Thank you! We agree and have addressed by modifying the introduction (lines 80, and 174-175) and providing a table that contrasts MOCHA and PALMO's functionalities (Supplementary Table 1).

Minor concerns:
1. Improving the readiness of figures. Some visualizations can be further improved and some conventions can be considered. Here are a few examples:
a. Color coding. For example, in Fib 2a, maybe different colors should be used for CD4 CTL TEM and CD8 TEM?
b. Orders. Fig 2 a: if the same order of cell types was used, it could improve readability.
c. Scales. Fig 2a, Fig 2b, etc. – using the same scales would help the comparisons.
d. Numbers: Fig 2b – if "400000" could be visualized as "400,000", it would help. And it is

inconsistent that, in the same Fig 2b, one subplot was visualized using "400000", and the other as "3e+5".

We thank the reviewer for pointing these out and have addressed these edits indicating how each was addressed below.

1a - We added more axes and legend labels to better clarify that the cell types are color-coded by major cell class, so all T-cell populations will share the same colors.
1b - This has been addressed – the orders have been fixed.
1c - This has been addressed – all have the same scales.
1d - This has been addressed – all graphs follow the same format.

2. Method: sample preparation and data preparation info has been provided in the preprint and thus not necessary to be re-described in detail.

We will leave it for Editors to decide whether such information is needed.

Reviewer #2 (Remarks to the Author):

In this paper, the authors developed a statistical approach, MOCHA to identify sample-specific cell-type open chromatin regions using scATAC-seq data. They tested MOCHA to multiple single cell datasets such as COVID19, immunology and Hematopoiesis and demonstrated its outperformances over existing methods for detecting sample-specific chromatin accessibility, differential accessibility in covid, and co-accessibility across samples. Moreover, using detected OCRs, the paper further inferred regulatory networks linking TF binding sites, ligands to TSSs, revealing possible alternative gene regulatory mechanisms and longitudinal dynamics in covid.

Overall, the study was designed with reasonable rationale. Identifying sample-specific activities of chromatin (co-) accessibility at the cell type level is an emerging topic, so MOCHA provides a timely statistical tool. The paper was well organized, and the results were presented logically. However, I still have some major concerns, especially about the rigor of the methodology and evaluation.

Thank you! We appreciate the feedback and conducted a variety of additional analyses to help address the benchmarking concerns and recommendations highlighted by the reviewer.

1. It is unclear if MOCHA is a general method or specific for COVID/immune study. If former, the paper needs to demonstrate broader applications.

We thank the reviewers for pointing this out and agree that we should've further demonstrated the generalization of MOCHA. We have addressed this by benchmarking performance using a murine scATAC atlas (Supplemental Figure 5), which shows MOCHA's ability to generalize across multiple species and 13 tissue types (lines 130-132, 227-232).

Additionally, we show applications of zero-inflation across cell types to study biology (lines 364-366, Supplementary Fig. 18).

2. The logistic regression model for evaluating accessibility needs further clarification. The authors claimed the usefulness of normalized total counts, lamda_1 and almost downstream analyses seem only use lamda_1. However, how important the max count (lamda_2) contributes to the regression? If also important, it remains elusive that lamda_2 is not used in downstream DAT and CAA analyses. Moreover, the study specific prefactor S was insufficiently described without justification.

We agree that this was not clear and thank the reviewer for pointing this out. When training the LRM, we tried creating as many statistically informative features to identify open chromatin, and selected $\lambda^{(1)}$, $\lambda^{(2)}$ to identify open chromatin based on single-cell and pseudo-bulk characteristics. Supplementary Fig. 7 illustrates the statistical significance of the final features in the model, whose coefficients had 95% confidence intervals well outside zero. We added some details to clarify the issue (lines 142-145, 412-414).

Additionally, we revised the text to provide greater clarity on the language around the different model parameters for the LRM, including specifics around the global prefactor S (lines 114-118, 130-132).

3. When comparing with MACS2 and HOMER, the authors should also report functional or disease enrichments (e.g., LDSC) of MOCHA OCRs in addition to CTCF sites and TSSs, like their other sections did.

We thank the reviewer for suggesting another validation strategy, and we conducted an LDSC analysis per the reviewer's recommendations (Supplementary Fig. 9). The LDSC results further support previous results and demonstrate that MOCHA's open chromatin model agrees with existing methods while detecting additional, functional enriched regions (lines 210-213).

4. MOCHA randomly selected 50 DATs with two clusters by K-Means. How sensitive would its performance be to those hyperparameters? Also, K-Means is also not robust to outliers. This concern applies to many other parameters. The authors need to justify selecting parameters and provide guidelines for the users, especially biologists.

We recognize that the context for the K-means clustering results was not clear. This analysis was strictly for benchmarking purposes, to demonstrate that MOCHA tiles on average contained more biological information, and not as a framework for others to use in their interpretive biological analyses. We have addressed the language to clarify (lines 259).

5. The networks that the paper predicted were not fully gene regulatory networks. They only linked TF binding sites, ligands to TSSs (near promoters) so missed other key regulatory mechanisms such as distant regulatory elements (enhancers from scATAC-seq data), gene *expression relationships (e.g., co-expression from many methods for predicting gene regulatory networks like SCENIC).*

We agree the language around gene regulatory regions that MOCHA infers is not clear and have addressed the language to clarify when and how distal regulatory regions were included in our gene regulatory networks (lines 315-317, 1018-1028). Since MOCHA is a scATAC-seq centric tool, gene expression and multi-modal gene regulatory network analyses (e.g., SCENIC) are extremely interesting but outside the scope of this study.

6. For longitudinal analysis, it is unclear that confounding factors (e.g., sex, age, etc) were considered for detecting chromatin accessibility dynamics.

We agree that this was not mentioned in the main text and thank the reviewer for pointing this out. We did adjust for age and sex, and we have revised the main text to make this explicit (lines 353-354, 369).

7. The significant p-values were reported inconsistently thru the paper, e.g., p-value, adjusted p, FDR.

We thank the reviewer for pointing these out and have addressed these edits in the main text and methods to make explicit what correction procedures were used when and where.

Reviewer #3 (Remarks to the Author):

The manuscript describes MOCHA, a method primarily for carrying out comparisons of single cell ATAC-seq data between groups of subjects. The method also includes features for identifying alternative transcription-start-site regulation, and transcription factor-gene network construction from longitudinal data. More rigorous ways of comparing scATAC-seq data sets are needed, however, it is not clear that MOCHA is making a substantial contribution for reasons given below.

Thank you. We appreciate the feedback. We have improved the language around our analyses to address gaps highlighted by the reviewer. Additionally, we conducted a variety of additional analyses to help address the benchmarking concerns and recommendations highlighted by the reviewer.

1. The manuscript correctly observes that in single cell analysis comparisons between treatment and control groups should be done on the level of subject rather than cell, as treating cell level data as replicates would artificially inflate the significance of differences. The analysis proposed is therefore based the aggregation of single cell data into pseudobulk representations for different cell types. Differential analysis is then based on comparisons of sample level pseudobulk aggregates for the cell types. The idea of using pseudobulk to make comparisons of single cell data has been previously evaluated by Junttila et al, for example, who compared 18 methods for the identification of differential expression changes between conditions from *multisubject scRNA-seq data. Many of the methods assessed by Juntilla et al could also be used to compare scATAC-seq data.*

> *Junttila, Sini, Johannes Smolander, and Laura L. Elo. "Benchmarking methods for detecting differential states between conditions from multi-subject single-cell RNA-seq data." Briefings in bioinformatics 23.5 (2022)*

The authors should carry out more careful survey of single cell studies; relevant studies are not cited in the manuscript and there are likely to be many more.

We agree with the reviewer that the idea of pseudo bulking is not new in scRNA. This inspired its use in scATAC for MOCHA. Given the advantages that pseudo bulk has shown by the scRNA community, we apply and cite accordingly when applying to scATAC. While this was already mentioned and cited in our writing, we provided further clarification in the discussion (lines 444-448).

Additionally, we agree that a variety of other methods from single cell and bulk sequencing assays could theoretically be applied to scATAC-seq data. Per the reviewer's recommendations, we added two additional methods into our benchmarking (Supplementary Fig. 11, 12) and revised the corresponding text (Lines 278-280, 438-440). A complete review of all potential differential methods in scATAC is beyond the scope of the current work.

2. Apart from the above mentioned scRNA-seq study, the MOCHA methodology is closely related to methods for bulk differential ChIP-seq or ATAC-seq peak calling, many of which are based on limma, DESeq2 or EdgeR. The following papers need to be cited. It is of critical importance that the methods described in these papers be included in benchmarking comparisons of differential accessible regions:

> *Gontarz, Paul, et al. "Comparison of differential accessibility analysis strategies for ATAC-seq data." Scientific reports 10.1 (2020)*
>
> *Chen, Yang, Shue Chen, and Elissa P. Lei. "DiffChIPL: a differential peak analysis method for high-throughput sequencing data with biological replicates based on limma." Bioinformatics 38.17 (2022)*
>
> *Stark, Rory, and Gordon Brown. "DiffBind: differential binding analysis of ChIP-Seq peak data." R package version 100.4.3 (2011)*
>
> *Faux, Thomas, et al. "Differential ATAC-seq and ChIP-seq peak detection using ROTS." NAR Genomics and Bioinformatics 3.3 (2021): lqab059.*
>
> *Qiu, Xintao, et al. "CoBRA: containerized bioinformatics workflow for reproducible ChIP/ATAC-seq analysis." Genomics, Proteomics and Bioinformatics 19.4 (2021)*

We agree that MOCHA implements pseudo-bulking to make some calculations more statistically robust, leading to some similarities with existing methods for ChIP-Seq or ATAC-seq (lines 438-440). We thank the reviewers for highlighting citations we have missed. We have included these citations (439), and included two of them in our benchmarks (lines 278-280). Additionally, we have clarified the methodological differences between MOCHA and these methods, and shown how these methods cannot be directly applied to scATAC.

3. Benchmarking of differential tiles between conditions is an important aspect of the paper and needs to be done rigorously. In the manuscript numbers of distinct genes and numbers of reactome pathways provide some anecdotal evidence that the method is working. However, a reliable gold standard of true differentially accessible regions is never established. To benchmark rigorously, gold standards need to be constructed to test both sensitivity and specificity. The benchmarking methodology used in the DESeq2 paper could be suitable for this. To test specificity, comparisons can be made of groups each containing a mixture of COVID+ and COVID- subjects, positives found in this analysis would be false positives. For sensitivity, the approach from the DESeq2 paper can be used: "we used experimental reproducibility on independent samples (though from the same dataset) as a proxy. We used a dataset with large numbers of replicates in both of two groups, where we expect that truly differentially expressed genes exist. We repeatedly split this dataset into an evaluation set and a larger verification set, and compared the calls from the evaluation set with the calls from the verification set, which were taken as truth." The authors might also consider the concepts introduced in:

> *Tian, Luyi, et al. "Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments." Nature methods 16.6 (2019)*

Benchmarking in the current manuscript has not been done to an acceptable standard, and comparisons have not been made to the most relevant methods.

We thank the reviewer for pointing this out, and agree that resampling methods provide robust evaluations to assess false positives and false negatives. For this purpose, we applied a leave-one-out approach (Fig 3F), similar to the approach by DESeq2, to assess potential false positives and false negatives across methods. We revised the language to clarify how we applied one of these approaches (lines 260-266).

4. The dataset generated in this study involved numerous subjects and must have been done in several batches. Although batch effects are well known to impact single cell data such effects are not mentioned at all in the manuscript. It is important to provide the batch information and evaluate the degree to which batch effects could be influencing results. For example, when defining cell types do cells in different batches have similar chromatin accessibility, or do batches also define observed chromatin accessibility. A useful reference is:

Luecken, Malte D., et al. "Benchmarking atlas-level data integration in single-cell genomics." Nature methods 19.1 (2022): 41-50.

We agree that batch effect is a frequent technical challenge and agree that we did not clarify how we address it. We provide an additional figure highlighting batch effects in both single cell and pseudo-bulk space. In single cell space, the standard UMAP approach shows minimal batch effects in our Longitudinal COVID19 dataset (supp Fig3B). In pseudo bulk, batch effects are minimized after normalization (Supp Fig3A). Furthermore, we tested for the influence of batch using variance decomposition analysis, revealing that 0.193% (225/116,632) tiles have variance mostly described by batch effects (Suppl Table 3). Therefore, we did not include batch effects as a covariate in our COVID19 analysis. Finally, we discuss how any residual batch effect can be controlled downstream if necessary (lines 106-108, 366-368, 1108-1116).

5. The MOCHA logistic regression model is used to create a matrix of accessibility on a sample by tile level. The procedure is used to collapse the single cell data into indicators of accessible tiles in given cell types and samples. The approach seems overly complicated in comparison with the DESeq/EdgeR/limma based methods, so comparisons will be important. Only lambda 2, the maximum number of fragments in a tile per cell seems to be truly related to single cell analysis. It is not clear how important this parameter is in the analysis and whether the need for this parameter could be obviated through simple filtering measures. For example, filtering identical fragments from the same cell or constraining the maximum number of fragments per tile per cell to 2. In addition, some description of what this parameter is achieving would be helpful.

We agree that having a count matrix data structure such as a gene-by-sample matrix facilitates any downstream analyses using DESeq, EdgeR, or limma. Our tile-sample-accessibility matrix (TSAM) is equivalent to their 'gene-by-sample' matrix, where we replace genes with open tiles. However, unlike RNA-seq, open tiles must be determined from the data. We added some background on the use of TSAM (lines 650-679).

Additionally, we agree that filtering duplicate fragments and other QC methods are critical to ensure robust data analysis. MOCHA is designed to run after QC steps. Duplicate fragments were removed and data QC concerns were addressed during and after running the 10x pipeline (CellRanger), including the removal of low quality cells and doublets. We have revised and clarified the language in Abstract (lines 28-29) and Results (lines 93-94) to make this clearer.

The parameter $\lambda^{(2)}$ is important for the logistic regression model (LRM) of identifying open chromatin. When training the LRM, we tried creating as many statistically informative features to identify open chromatin, and selected $\lambda^{(1)}$, $\lambda^{(2)}$ to identify open chromatin based on single-cell and pseudobulk characteristics. Supplementary Fig. 7 illustrates the statistical significance of the final features in the model, whose coefficients had 95% confidence intervals well outside zero. We added some details to clarify the issue (lines 142-145, 412-414).

6. Fig 2 shows total numbers of open tiles, and there is a threshold parameter that controls this number in MOCHA. The number of open tiles determined by MACS2 and HOMER could also be changed by altering cut-off parameters. In benchmarking it is not enough to define true accessible regions, as one can always get more tiles changing thresholds. Unless some way of showing specificity is included, this analysis is not meaningful.

We agree that thresholding can change the results. To address this valid concern, we added simulation studies to demonstrate the open-chromatin performance for all three methods (Lines 200-203, 754-752, 831- 870, Supplementary Fig. 8) in simulation settings including measurements of specificity and sensitivity across a range of cellular abundances.

7. Line 893: "We used a previously published promoter-capture HiC (pcHiC) resource43 which identified promoter-enhancer regulatory links."

No justification is given for using HiC contacts as a gold standard for co-accessibility. First the manuscript should provide a motivation in terms of causality. The causal relationship between HiC and accessibility is not well understood and it is possible that chromatin accessibility causes HiC contacts rather than the other way round. Second, what HiC measures and its relationship to biology needs to be taken carefully into account. HiC is a protocol that measures, in some sense, proximity between genomic regions. It cannot be assumed that HiC precisely measures all the biologically relevant interactions between regions and only these. Third, the limitations in the specific HiC data will include some inaccuracy, limitations in sequencing depth, suboptimal experimental conditions etc. Overall, the case for HiC as a gold standard is not at all compelling.

We thank the reviewer for pointing out the limitations of HiC, and our phrasing. Our datasets cannot address causality, and so we refrain from commenting on any causal relationships. Instead, we use HiC and scATAC results to enrich for regions that are both open and potentially interacting to benchmark distinct correlation methods. We also agree that HiC is not a gold standard for co-accessibility. Despite limits to its accuracy and precision, HiC has provided valuable information on chromatin interactions and correlations between genomic regions in epigenetic data, including ATAC-seq [1-2].

> **[1] Fortin, JP., Hansen, K.D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 16, 180 (2015). https://doi.org/10.1186/s13059-015-0741-y**
> **[2] Gate, Rachel E., et al. "Genetic determinants of co-accessible chromatin regions in activated T cells across humans." *Nature genetics* 50.8 (2018): 1140-1150.**

In response to this question, we modified the text in the results (lines 282-284) and Methods (lines 946-958) to clarify this benchmarking analysis and removed references to HiC as a 'gold standard'.

8. The sections "Networks of alternatively regulated genes in early SARS-CoV-2 infection." and "Longitudinal analysis of chromatin accessibility during COVID-19 recovery" describe results from the COVID data generated in the project but do not evaluate methodology or make any comparisons with other methods. The manuscript notes that "An in-depth, comprehensive analysis of our COVID19 cohort is beyond the scope of current work and will be presented in a follow-up paper." It might be better to leave the longitudinal and gene network analyses for that paper.

We agree that benchmarking novel analytical frameworks is critical and added additional benchmarking when comparable methods exist. For alternatively regulated genes from Figure 4, we have added Supplementary Fig 14 that benchmarks MOCHA's performance with ArchR and Signac for alternative TSSs analysis (lines 311-314). For Figure 5, we note that longitudinal data analyses with zero-inflation are not currently supported by either ArchR, Signac, or PALMO (lines 174-175, 389-390, Supplemental Table 1). Instead we benchmarked pseudotime analyses using ArchR's gene scores with MOCHA's promoter tiles. Here we demonstrated that MOCHA-based results were more biologically informative and aligned better with the expected roles of CD16 monocytes than GeneScore-based ones (lines 357-362) .

While we appreciate the reviewer's feedback, we respectfully disagree as these two sections highlight the biological utility of MOCHA.

9. Line 719 "we applied MACS2 37 ('-g hs -f BED --nolambda --shift -75 --extsize 150 --broad', '--model -n') to identify accessible peaks in the pseudobulk data, using previously published parameters for identifying peaks in scATAC-seq with the modification to call broad rather than narrow peaks. It is not clear why the --broad MACS2 option was used. Can some advantage be demonstrated? The –shift -75 –extsize 150 also doesn't seem to be well motivated.

We thank the reviewer for pointing this out. Regarding parameter choice, we applied previously published parameters for MACS2-based peak calling from scATAC data (lines 777-778). We recognize that parameters could be further optimized, but on this point, we defer to the authors of other packages that apply MACS2 in their pipelines.

10. The abstract doesn't describe the manuscript very well. The method seems to be primarily about analysing differential accessibility in multi-sample studies. The question of "proper" handling of technical dropout with zero-inflated methods, is highly debatable. Are the proposed heuristics proper handling? Identification of alternative transcription-starting-site regulation, and transcription factor–gene network construction from longitudinal scATAC-seq data are weak sections without benchmarking comparisons.

We thank the reviewer for their feedback and have addressed this by modifying our abstract to explicitly denote the improvements by MOCHA (lines 25-29). Each figure indicates the

different types of analyses available in MOCHA, ranging from identifying open tiles to longitudinal analyses of open chromatin.

11. A complementary approach to single cell analysis is to carry out differential abundance testing. Comment on the relative strengths and weaknesses of the proposed approach relative to methods such as:

Dann, Emma, et al. "Differential abundance testing on single-cell data using k-nearest neighbor graphs." Nature Biotechnology 40.2 (2022): 245-253.

We thank the reviewer for highlighting this paper. Since MOCHA pseudobulks after cell labeling, this type of analysis is unrelated to MOCHA's approach, and therefore not directly comparable. However, in response to this reviewer, we now include abundance statistics in the metadata of MOCHA objects, which can be used by researchers interested in differential abundance and describe how to access them (lines 650-679).

Reviewer #4 (Remarks to the Author):

This manuscript declared MOCHA, a tool to identify the gene regulatory programs when analyzing the scATAC-seq data. MOCHA exhibits the advantages in detecting differential accessible regions and chromatins than widely used tools including MACS2, HOMER, ArchR, Signac. The author also showed the good performance of MOCHA in the large dataset of COVID19 patients, and constructed ligand-TF-gene networks on alternative TSS regulations, which would be used to identify potential targets for COVID19 or other processes. And MOCHA can be integrated with existing tools such as ArchR, chromVAR, as a valuable extension for analyzing scATAC-seq data. The following comments or issues need to be considered.

Thank you! We appreciate the feedback and provide clarifying comments and additional analyses to help address the concerns and recommendations provided by the reviewer.

Comment：

1) Line105-107, MOCHA identifies sample- and cell type-specific open chromatin, within samples from different experiment and batches. How to distinguish and balance the bias from the batch effect using the MOCHA?

We agree that batch effect is a frequent technical challenge and agree that we did not clarify how we address it. We provide an additional figure highlighting batch effects in both single cell and pseudo-bulk space. In single cell space, the standard UMAP approach shows minimal batch effects in our Longitudinal COVID19 dataset (supp Fig3B). In pseudo bulk, batch effects are minimized after normalization (supp Fig3A). Furthermore, we tested for the influence of batch using variance decomposition analysis, revealing that 0.192% (245/127,075) tiles have variance mostly described by batch effects (Suppl Table 3). Therefore, we did not include

batch effects as a covariate in our COVID19 analysis. Finally, we discuss how any residual batch effect can be controlled downstream if necessary (lines 106-108, 366-368, 1108-1116).

2) During tiling the genome, MOCHA splits the genome into 500 bp tiles, all the analyses are based on the tiles. But this strategy has been adopted by the previously published SnapATAC. Then what are the differences and advantages of MOCHA compared to SnapATAC?

We agree that tiling the genome is not a novel strategy, and has been successfully applied by SnapATAC and others. In addition, we recognize that while we did not explicitly benchmark Signac, snapATAC and ArchR for open chromatin identification, these methods implement MACS2 for peak calling, and are thus benchmarked indirectly via our MACS2 peak calls. We have revised the text in Results (lines 181-182) and in Discussion (lines 405-407).

3) Follow the comment, does MOCHA eventually split the genome into 500 bp tiles? How about other ranges, such as 1 kb, 1.5 kb, 2 kb, 5 kb, which is better?

We agree that this was not clarified. MOCHA starts by splitting the genome into 500 bp tiles, and the model is explicitly trained for 500 bp, in-line with the standards set by others (e.g., ArchR, SnapATAC, and Signac). We revised the discussion to provide additional language around this topic (lines 405-407).

We agree that while modeling open regions of distinct sizes may be useful, we selected 500 bps to balance between having large coarse regions and small, extremely sparse regions. Additionally, we provide a code template (see Github code repository) for retraining open chromatin models of other tile sizes.

4) The authors stated that MOCHA is more sensitive in detecting open chromatin regions than MACS2 and HOMER, and detects more differential chromatin than ArchR and Signac. However, the splitting genome into 500 bp tiles could somehow cause potential bias by differences in data qualities when using MOCHA, which needs to be discussed.

We recognize that tiling the genome could cause potential bias for both open chromatin detection and differential accessibility. We revised the text to clarify how we addressed this bias, for both steps. In the discussion (lines 405-407) we added new language discussing the motivation and implications of tiling the genome for open chromatin prediction.

To address potential biases around differentials, we used the same set of predicted accessible regions from the LRM for all methods. We updated the results section of the text so as to avoid future confusion for readers (lines 240-243).

5) Fig 4a-c: what's the meaning of type I and type II sites in fig 4a? It seems that there is no difference between early infection patients and control donors in fig 4b-c? How to understand and calculate the Accessibility Change in Fig 4b-c?

We agree that this was not clear and have revised the text to clarify (lines 303-306). We have provided explicit definitions of type I and type II sites, and clarify how the significant difference in accessibility was visualized.

6) Fig 4f and line 309-310 in page 8: The authors identified 122 ligands. Are these ligands regulated by all the differential motifs as shown in fig 4e?

We agree that this was not clear and we've clarified our text (see Methods: Ligand-motif enrichment analysis, lines 1030-1031) to specify that ligands are regulating transcription factors (and thus motif enrichment), and not vice versa.

7) Line 318-320: This reviewer couldn't find the data demonstrated the regulation in CD16 monocyte.

We thank the reviewer for pointing this out. The full network and analyses leading up to that are all in Source Data Figure 4, including DATs, Motif Enrichment, Ligand-Motif Enrichment, and the full network's nodes and edges.

8) How about the computer requirements to run the MOCHA?

We agree that this was not clarified and have addressed this by revising the text (lines 175-178) including the technical specifications of benchmarks conducted on a laptop.

9) page 13, row 504: An error labeling of "1x106 cells".

Thank you for pointing this out. We've corrected it in the manuscript.

Reviewer #1 (Remarks to the Author):

The authors have thoroughly replied to each concern, with sufficient further analysis and explanations when necessary. I have no further concerns.

Reviewer #1 (Remarks on code availability):

The code is functional and the instruction is sufficient.

Reviewer #2 (Remarks to the Author):

Thank authors for considering my comments. They have addressed many of my previous concerns. I just have two more responses:

1) for gene regulatory networks, I understand the study focuses on chromatin regions but the authors at least can check if their predicted TFs express in the corresponding cell types. If those TFs lowly express, the network links might be false positive;

2) for K-means, the authors claim that it's just for benchmarking. How optimally were the parameters tuned across different methods including K to make a fair benchmarking?

Reviewer #3 (Remarks to the Author):

The revised version of the paper provides some more evidence for the method's performance, however several comments from the previous review have not been adequately addressed. In addition, some of the new material raises further questions.

Major comments

1. When making predictions of accessibility on a new dataset a parameter S is used to scale fragment counts. If there are very few reads in the new dataset there would not be much statistical support for peaks. Wouldn't MOCHA yield results with a high false positive rate in such cases? Can you show mathematically that this is not a problem?

2. A major part of the manuscript is about differential analysis of scATAC-seq data, therefore I think that a review of the available methods for such analysis is a necessary part of the introduction to the study.

3. In the manuscript, much is made of the role of zero-inflation. However, it is debatable whether a model of zero-inflation is more suitable than a negative binomial one. See for example. Sarkar, A., Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNAsequencing analysis. Nat Genet 53, 770–777 (2021). https://doi.org/10.1038/s41588-021-00873-4
A more comprehensive analysis of the performance of available negative binomial models, using credible benchmarking standards is needed.

4. The method for assessing differentially accessible tiles has several drawbacks, the use of k-means, the G index and the number of DATs randomly selected. It does not directly show that the DATs called are true, but that the DATs that are called, yield better G indices when analyzed using k-means clustering. The degree to which these results support the performance in the differential analysis are not well understood. For this reason I previous suggested the authors make more direct assessments of differential analysis. The leave-one-out approach they have added to the revision has little semblance to the method in DESeq2 that was suggested. My comment from the first review is repeated here:
" The benchmarking methodology used in the DESeq2 paper could be suitable for this. To test specificity, comparisons can be made of groups each containing a mixture of COVID+ and COVID- subjects, positives found in this analysis would be false positives. For sensitivity, the approach

from the DESeq2 paper can be used: "we used experimental reproducibility on independent samples (though from the same dataset) as a proxy. We used a dataset with large numbers of replicates in both of two groups, where we expect that truly differentially expressed genes exist. We repeatedly split this dataset into an evaluation set and a larger verification set, and compared the calls from the evaluation set with the calls from the verification set, which were taken as truth."

5. Another issue from the previous review has not been adequately addressed:
"Fig 2 shows total numbers of open tiles, and there is a threshold parameter that controls this number in MOCHA. The number of open tiles determined by MACS2 and HOMER could also be changed by altering cut-off parameters. In benchmarking it is not enough to define true accessible regions, as one can always get more tiles changing thresholds. Unless some way of showing specificity is included, this analysis is not meaningful." The revision does contain additional analysis in Supplementary Fig 8, which is helpful. However, the significant region count shown in Fig 2 has the same problems as before.

6. The simulation study is a promising addition, but it is unclear how MOCHA was applied. It seems that the simulation should include more than one cell type. How are the initial peaks called by MACS2? Which cells is the MOCHA logistic regression trained on? Which other types of cells are the parameters extended to ? How are different cell types included in the simulation ?

7. ATAC-seq analysis is typically carried out using narrow, not broad, peak calling, so shouldn't the standard practice be represented?

Reviewer #3 (Remarks on code availability):

Code is there with some limited documentation. I did not install and run it.

Reviewer #4 (Remarks to the Author):

The authors have addressed all my concerns. I am satisfied with the responses.

Response to Reviewers

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors have thoroughly replied to each concern, with sufficient further analysis and explanations when necessary. I have no further concerns.

Reviewer #1 (Remarks on code availability):
The code is functional and the instruction is sufficient.

Reviewer #4 (Remarks to the Author):

The authors have addressed all my concerns. I am satisfied with the responses.

Reviewer #2 (Remarks to the Author):

Thank authors for considering my comments. They have addressed many of my previous concerns. I just have two more responses:

1) for gene regulatory networks, I understand the study focuses on chromatin regions but the authors at least can check if their predicted TFs express in the corresponding cell types. If those TFs are lowly expressed, the network links might be false positive.

> This is a fair point. Using published RNA-seq datasets (Monaco and Schmiedel Datasets), we found that 9 out of the 10 transcription factors in Figure 5D had high to medium expression in CD16 monocytes (aka non-classical monocytes), relative to other cell types. However, the last TF, SPIB, had very low expression in CD16 monocytes but was highly expressed in macrophages (Archs4 database), which differentiate from CD16 monocytes. We have modified the text (lines 400-402) to reflect this information.

2) for K-means, the authors claim that it's just for benchmarking. How optimally were the parameters tuned across different methods including K to make a fair benchmarking?

For the benchmarking analysis in question, the only changeable parameter is the number of input DATs in clustering. We tested N = 25, 50, 75 and 100 DATs in the benchmarking. We must set K = 2 in the K-means clustering to benchmark whether the DATs can effectively separate COVID+ from COVID– samples in a 2-group comparison. We have added such information in the text (lines 267-272). We also recognize that we previously used K to denote the # of DATs in Supplementary Figures 11-12, which may cause confusion. We have changed it to N.

Reviewer #3 (Remarks to the Author):

The revised version of the paper provides some more evidence for the method's performance, however several comments from the previous review have not been adequately addressed. In addition, some of the new material raises further questions.

Major comments

1. When making predictions of accessibility on a new dataset a parameter S is used to scale fragment counts. If there are very few reads in the new dataset there would not be much statistical support for peaks. Wouldn't MOCHA yield results with a high false positive rate in such cases? Can you show mathematically that this is not a problem?

This is a fair point. To address this concern, we ran additional simulations in which we varied the number of fragments per cell, from 4k fragments/cell down to 1k fragments/cell, so as to simulate performance as the number of reads is lowered. We did not go below 1k fragments per cell since existing scATAC-seq data analysis softwares (e.g., Signac, ArchR, and SnapATAC) would toss out between 50% to 100% of cells in their quality control steps. MOCHA attained the highest F-1 score in 210/240 iterations (87.5%, Supplementary Fig 8B-D). We have revised the text (lines 208-215) to incorporate these new results, and showcased that MOCHA controls FPR in datasets with few reads.

We also note that in our original submission, downsampling on real data demonstrated that MOCHA was able to control FPR at low cell counts (Supplementary Fig 6b-d, model training). Likewise, our simulation showed that MOCHA's false positive rate with low cell counts (and thus low total reads) was smaller than HOMER's and MACS2's FPR (Supplementary Fig 8A).

2. A major part of the manuscript is about differential analysis of scATAC-seq data, therefore I think that a review of the available methods for such analysis is a necessary part of the introduction to the study.

We have revised our introduction (lines 65-68) to broaden our discussion of methods used for scATAC-seq differential analysis. Specifically, we have added four references (Shi et al, 2022; Chen et al, 2019; Nuno et al, 2023; and Stuart et al, 2021) to expand our review of available methods.

3. In the manuscript, much is made of the role of zero-inflation. However, it is debatable whether a model of zero-inflation is more suitable than a negative binomial one. See for example. Sarkar, A., Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nat Genet 53, 770–777 (2021). https://doi.org/10.1038/s41588-021-00873-4
A more comprehensive analysis of the performance of available negative binomial models, using credible benchmarking standards is needed.

We acknowledge that there has been a debate on this topic with respect to scRNA-seq data and are not aware of similar debates in scATAC-seq. There are a number of recently published papers for and against zero-inflated models. Below are several studies using zero-inflated methods in scATAC-seq or multi-omics data.

- Maniatis, Christos, Catalina A. Vallejos, and Guido Sanguinetti. "SCRaPL: A Bayesian hierarchical framework for detecting technical associates in single cell multiomics data." *PLoS computational biology* 18.6 (2022): e1010163.

- Dayu Hu, Ke Liang, Zhibin Dong, Jun Wang, Yawei Zhao, Kunlun He, Effective multi-modal clustering method via skip aggregation network for parallel scRNA-seq and scATAC-seq data, *Briefings in Bioinformatics*, Volume 25, Issue 2, March 2024, bbae102, https://doi.org/10.1093/bib/bbae102
- Chen, Li, et al. "scaDA: A Novel Statistical Method for Differential Analysis of Single-Cell Chromatin Accessibility Sequencing Data." *bioRxiv* (2024): 2024-01

A recent deep-learning based tool for scATAC-seq data implements zero-inflated approaches.

> W. Lan, X. Sun, Q. Chen, J. Ye, X. Zhu and Y. Pan, "scIAC: clustering scATAC-seq data based on Student's t-distribution similarity imputation and denoising autoencoder," *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA, 2022, pp. 206-211, doi: 10.1109/BIBM55620.2022.9995225.

We have modified the text to include this information, including recent preprints investigating the use of zero-inflated models (lines 62, 466-468).

To address the reviewer's concern about zero-inflation in pseudobulk scATAC-seq data, we fit a negative binomial (NB) distribution on pseudobulked accessibility ($\lambda_{(1)}$) at each open tile. We then tested whether $\lambda_{(1)}$ is zero-inflated using an independent R package, DHARMa, which utilizes the model fits from the glmmTMB R package. More specifically, the test compares the observed number of zeros with that expected from a NB distribution: An estimate of >1 means that there are more zeros than expected by a NB model and a p < 0.05 means that the observed and the expected zeros in the data is significantly different. To be comprehensive, we ran the test using two standard parameterizations of the NB family, as implemented in the glmmTMB package: The variance grows either linearly (NB1) or quadratically (NB2) with the mean. Under the first parameterization (NB1), we observed that 41%, 37%, and 38%, respectively, of open tiles in the CD4 Naive T cells, Naive B cells, and CD16 Monocytes were zero-inflated (p < 0.05 & statistic > 1) (Supplementary Figure 18C - left column). The corresponding rates under the NB2 were 46%, 42%, and 44%, respectively (Supplementary Figure 18C - right column). For comparison, underinflation of zeros was not observed (p < 0.05 & statistic < 1, see inserts of Supplementary Figure 18C). These evidences justified the use of zero-inflation methods in analyzing pseudobulk scATAC-seq data. Finally, we want to point out that the zero-inflated methods in MOCHA can also handle data without zero inflation. We have revised the text to include this information (lines 154-155, 466-468, 906-915).

4. The method for assessing differentially accessible tiles has several drawbacks, the use of k-means, the G index and the number of DATs randomly selected. It does not directly show that the DATs called are true, but that the DATs that are called, yield better G indices when analyzed using k-means clustering. The degree to which these results support the performance in the differential analysis are not well understood. For this reason I previously suggested the authors make more direct assessments of differential analysis. The leave-one-out approach they have added to the revision has little semblance to the method in DESeq2 that was suggested. My comment from the first review is repeated here:
" The benchmarking methodology used in the DESeq2 paper could be suitable for this. To test specificity, comparisons can be made of groups each containing a mixture of COVID+ and COVID- subjects, positives found in this analysis would be false positives. For sensitivity, the approach from the DESeq2 paper can be used: "we used experimental reproducibility on independent samples (though from the same

dataset) as a proxy. We used a dataset with large numbers of replicates in both of two groups, where we expect that truly differentially expressed genes exist. We repeatedly split this dataset into an evaluation set and a larger verification set, and compared the calls from the evaluation set with the calls from the verification set, which were taken as truth."

The Reviewer has the correct interpretation on our K-means/G Index evaluation but requests for specific assessments on specificity and sensitivity. For specificity, we conducted the suggested analysis by shuffling the labels of COVID+ and COVID- samples, and then testing for differential accessibility. MOCHA had a 0% false positive rate (FPR) across all 50 random permutations (Supplementary Figure 12B). In comparison, the minimum, median, and maximum FPR for other methods are: DESeq2 (0%, 0%, 3%), DiffChipL (0%, .05%, 11.1%), Signac (0%, 0.43%, 13.5%), and ArchR (0.11%, 1.0%. 41.6%). While all methods had very low median FPRs, DiffChipL, Signac, and ArchR showed some instability in FDR as the corresponding maximum FDR was 11.1%, 13.5%, and 41.6%, respectively. We updated the main text with these results (lines 292 - 295).

For sensitivity, we believe that the suggested assessment is inappropriate for our dataset. Here is the method described in DESeq2: "To obtain an impression of the sensitivity of the algorithms, we considered the Bottomly et al. [16] dataset, which contains **ten and eleven replicates** of two different, **genetically homogeneous mice strains**. This allowed for a split of three vs three for the evaluation set and seven vs eight for the verification set, which were **balanced across the three experimental batches**." (Highlights are marked by us.) In this idealized situation, it is reasonable to expect the same differential genes to be identified from both the evaluation set and the verification set. However, our COVID19X human dataset contains 22 COVID- samples and 17 COVID+ samples and is much more complex than the genetically and epigenetically homogeneous dataset used in DESeq2. The suggested assessment on sensitivity has two major challenges for our data: 1) Human samples are heterogeneous with differences in sex, age, health conditions, disease comorbidities, genetics, life experience, professional and environmental hazard exposure, and many many known and unknown factors. It is not feasible to split the samples into "evaluation" and "verification" subsets and eliminate all biases between them. In other words, differences observed between any two subsets can be real and biological, in addition to technical/methodological artifacts. 2) Splitting the samples will reduce the power for the study and increase the number of false negatives in both subsets. Nevertheless, to address the reviewer's concern on sensitivity assessment, we downsampled the number of subjects from the original n = 39 to n = 30 (a >20% reduction) and measured how sensitive each method was able to detect the original DATs (supplementary Figure 12C). Signac obtained higher recalls with its very conservative (thus very few) DAT calls, while the remaining 4 methods obtained similar recalls. We note that the recall evaluated in this approach was likely a lower bound due to disease/human heterogeneity and sample size reduction, a loss of > 20% of power when the sample size was reduced from n = 39 to n = 30 based on t-test (Supplementary Fig 12D). Overall, MOCHA provided comparable performance on recall, and better performance on FPR than other methods. We have updated our text to reflect these results (lines 295-301).

5. Another issue from the previous review has not been adequately addressed:
"Fig 2 shows total numbers of open tiles, and there is a threshold parameter that controls this number in MOCHA. The number of open tiles determined by MACS2 and HOMER could also be changed by altering cut-off parameters. In benchmarking it is not enough to define true accessible regions, as one can always get more tiles changing thresholds. Unless some way of showing specificity is included, this

analysis is not meaningful." The revision does contain additional analysis in Supplementary Fig 8, which is helpful. However, the significant region count shown in Fig 2 has the same problems as before.

a. We realize our original description did not make it clear that the cell count-dependent thresholds were fixed in all our analyses post the training of the logistic regression models (LRMs) on NK cells in our COVID19 dataset. We have modified the text to clarify the confusion (lines 132-135).
b. We would like to point out that we provided specificity results in Supplementary Fig 6d on validation datasets, which were obtained without changing the threshold and previously described in lines 136-138. We have added a reference to Supplementary Fig 6c,d to make it clear to readers (line 139).
c. Previously we used simulations to compare MOCHA, MACS2, and HOMER on their F1 score, recall, false discovery rate, and the number of total detected open regions as cell counts in the data are decreased (Supplementary Figure 8A). We have expanded the simulation to examine how sequencing depth and the number of "true" open regions impact the performance of the three methods (Supplementary Fig 8b-d). We have revised the text to incorporate these new results (lines 213-215).
d. As mentioned above, we did not adjust MOCHA parameters in open tile evaluation. Likewise, we used the default threshold settings for HOMER and MACS2, as shown in the code repository. Therefore, we believe our comparison on the performance of the three methods is fair.
e. Youden Index is widely used to balance the tradeoff between sensitivity and specificity (see page 80 of Pepe 2003). We mistakenly stated that the Youden Index is used "to balance accessible and inaccessible tiles in real data". As described by Eq (4.5) in Pepe 2003, the receiver operating characteristics (ROC) curve is fully determined by the score density distributions of disease and control samples. Thus the ROC curve and its optimal point do not depend on the numbers of accessible and inaccessible tiles. We have modified the text accordingly (lines 130 - 131).

Pepe, Margaret Sullivan. 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. OUP Oxford.

We believe we have adequately addressed the reviewer's concerns here.

6. The simulation study is a promising addition, but it is unclear how MOCHA was applied. It seems that the simulation should include more than one cell type. How are the initial peaks called by MACS2? Which cells is the MOCHA logistic regression trained on? Which other types of cells are the parameters extended to? How are different cell types included in the simulation?

Question 6A): it is unclear how MOCHA was applied

MOCHA was applied to the simulated data in the same way as to real data, a detail that has been clarified explicitly in the text (lines 210 - 211).

Question 6B): It seems that the simulation should include more than one cell type

We previously simulated 'generic' fixed peaksets as the ground truth for open and closed regions, a detail that has been added in the text (lines 208 - 209). We have broadened our simulation to cover 'multiple' cell types (see response to Question 6E-F below).

Question 6C): How are the initial peaks called by MACS2

We assume this question refers to finding ‘ground truth’ for training MOCHA’s logistic regression models. In these simulations, there is no need for “initial” peak calling by MACS2 since all “ground truth” peaks were predefined in the simulations and MOCHA was not retrained for the analysis.

Question 6D): Which cells is the MOCHA logistic regression trained on?

As previously described, MOCHA logistic regression models were exclusively trained on NK cells in our COVID19 dataset (lines 125 & 136).

Question 6E): Which other types of cells are the parameters extended to?

No MOCHA parameters were changed in analyzing the simulated data. With the exception for S, all other MOCHA parameters were fixed post training, and then applied as is to all cell types and datasets in this study (Figure 2, Supplementary Figure 5). We updated the text to make it clear to readers (lines 133 - 135, 210-211).

Question 6F): How are different cell types included in the simulation?

Previously, different cell types were modeled by varying the location of the open tiles in peaksets, while fixing the number of total open regions (Supplementary Fig 8a). In the new simulations, we have changed the number of fragments per cell (to model variation in sequencing depth and/or cell type, i.e., ploidy) and both the number and locations of open tiles to model cell type-related variations (Supplementary Fig 8b-d). We have modified Results (lines 208-215), Discussion (lines 445-447), and Methods (lines 887-900) to discuss the additional simulations and better describe the analyses.

7. ATAC-seq analysis is typically carried out using narrow, not broad, peak calling, so shouldn’t the standard practice be represented?

We recognize that many tools identify narrow peaks to answer specific biological questions (e.g., nucleosome positioning, insertion cut sites, etc..). These tools provide valuable insight for those specific questions and do not generally overlap with MOCHA’s functionalities [1-2]. Since MOCHA was designed to analyze pseudo-bulk open chromatin, we identified broad peaks, following previously published best-practices. On this topic, the literature states that tools which “stitch nearby narrow peaks to form broad peaks such as MACS2, HOMER, and SICER/epic2 are also thought to provide more meaningful results [3].” We have revised Discussion and added these references to highlight the different use-cases (lines 452-454).

[1] Gong, W., Dsouza, N. & Garry, D.J. SeATAC: a tool for exploring the chromatin landscape and the role of pioneer factors. *Genome Biol* 24, 125 (2023). https://doi.org/10.1186/s13059-023-02954-5

[2] Xu B, Li X, Gao X, Jia Y, Liu J, Li F, Zhang Z. DeNOPA: decoding nucleosome positions sensitively with sparse ATAC-seq data. Brief Bioinform. 2022 Jan 17;23(1):bbab469. doi: 10.1093/bib/bbab469. PMID: 34875002.

[3] Yan, F., Powell, D.R., Curtis, D.J. *et al.* From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* 21, 22 (2020). https://doi.org/10.1186/s13059-020-1929-3

Reviewer #3 (Remarks on code availability):

Code is there with some limited documentation. I did not install and run it.

All scripts for this manuscript are deposited on GitHub (https://github.com/aifimmunology/MOCHA_Manuscript/). The MOCHA website (https://aifimmunology.github.io/MOCHA/) contains Documentation, Vignettes, and Instructions on how to install and run MOCHA. The link to the website is also available on MOCHA's github repository.

Reviewer #3 (Remarks to the Author):

The revised manuscript is substantially improved from the last revision. The benchmarking based on the COVID data, in particular, is a useful addition. There are however a couple of points that could be made clearer.

1. In the simulation (Supp Fig 8) the question of different cell types is not clear. It seems that different cell types are represented through the inclusion of different numbers of fragments per cell, where the fragment positions are drawn from the same underlying distribution. If this is the case, this would not represent different cell types but different measurements of the same cell population. Some clarification or adjustment of the simulation is needed here.
2. The discussion of zero-inflation might be clearer if the distinction between accessibility models and measurement models is made (along the lines suggested by Sarkar and Stephens (2020)), and the reason for including zero-inflation. In other words, the biological phenomenon and the measurement thereof are two separate issues and there are different approaches to modeling the combination. What is the adopted in the manuscript?