nature portfolio

Peer Review File

Depth-corrected multi-factor dissection of chromatin accessibility for scATAC-seq data with PACS

Corresponding Author: Professor Junhyong Kim

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This manuscript presents PACS, a statistical model designed for conducting hypotheis testing on single-cell ATAC-seq data. PACS incorporates a cumulative logit regression model to explicitly account for read capture efficiency and various technical or biological factors. The authors shown that PACS effectively mitigates technical and undesirable influences, such as sequencing depth and batch effects, and differences arising from distinct sample origins. When compared with existing methods, PACS demonstrates superior performance in controlling type I error rates and achieving greater statistical power. The utility of PACS is further illustrated through its application in identifying spatially variable regions and time-dependent open chromatin regions. Though this work is potentially interesting, there are several issues need to be addressed.

Major Comments:

- Several regression based methods for identifying differentially accessible regions (DARs) have been proposed before. For example, Zhang et al. (2021, Cell) developed logistic regression based method for DAR finding. The decision to employ a cumulative logit regression model over more conventional models like the Poisson or negative binomial regression models for read count data necessitates clarification. Are there particular advantages or reasons for this choice?

- A central aspect of PACS is its ability to adjust for confounding factors while focusing on variables of interest. The effectiveness of PACS in this regard is crucial. Nonetheless, the UMAP in Figure 3b suggests potential residual batch effects even after applying PACS. It would be valuable to understand the extent to which PACS can ameliorate batch effects, especially in comparison to methods like Harmony or PeakVI.

- The integration of biological replicate information is pivotal for ensuring the reliability and reproducibility of results during DAR finding. Methods based on bulk or pseudo-bulk data naturally incorporate this aspect when replicates are available. It raises the question of whether PACS can also leverage biological replicate data during hypothesis testing.

Minor Comments:

- The section titled "PACS identifies kidney cell type-specific regulatory motifs and allows for direct batch correction" (Line 277) implies motif analysis, which does not appear to be discussed in the manuscript.

- The statement "We first examined the marginal effect of brain regions on chromatin accessibility, holding other factors constant (Methods)" (Line 332) deserves further explanation. It should be presented in a manner that is more accessible to biologists.

(Remarks on code availability)
The code comes with a README file that details the installation instructions and some examples. I have successfully installed the package and tested some demo.

Reviewer #2

(Remarks to the Author)
In the manuscript, the author proposed a statistical method, mcCLR (missing-corrected cumulative logit regression model), to detect the differential accessible regions (DAR) from single-cell ATAC-seq data. Undoubtedly, the DAR analysis is a critical step for biological discovery from scATAC data. From simulation analysis, the authors showed mcCLR has improved power with good false positive controls, compared to existing methods. Later, they also demonstrated the application of mcCLR in three experimental datasets where ground truth is generally lacking. Overall, the method shows good promise while it is a bit surprising that such multi-factor analysis is not well supported in the existing tools. Here are a few concerns about the manuscript and some comments for the authors to consider.

Major:
1. The primary contribution of this work is the mcCLR, which is a variant of Ordinal Logistic Regression. It seems the term “cumulative logit regression” is less commonly used than “Ordinal Logistic Regression (OLR)”. I would suggest the author either change it to the more commonly used term or at least mention the alternative term in the results. Also, citing a few examples of OLR in biomedical research (e.g., PMID: 1468011, PMID: 9429194) would help the readers to understand it better.
2. It is surprising that only one tool supports generalized linear regression models here (Seurat via logistic regression). For scRNA-seq data, there are multiple GLM methods supporting multi-factor analysis. Since mcCLR directly models the scATAC read counts, I strongly suggest the authors add one scRNAseq-based method for comparison. Particular example is the negative binomial regression with library size (e.g., via DE-seq or edgeR), which has a library size factor, reflecting the function of q_c, and negative binomial supports dispersion. It looks like snapATAC already used edgeR but seems not flexible; the author may directly use edgeR.
3. When comparing cell type label transfer with Naïve Bayes, the article says “Naïve Bayes does not assume missing data; thus, it ignores the cell-specific capturing probability.” Does this refer to q_c in Eq.1? It looks similar to a sequencing depth normalization (e.g., total counts or proportion of non-zero regions), which the naïve Bayes model may be able to consider as an additional variable. Also, it would be better to add more details about the naïve Bayes model here.
4. In batch-effect correction, the strategy of removing batch-effect peaks is very different compared to other methods that preserve all feature sets. Simply removing batch correlated peaks will likely introduce the risk of over-correction, since batch and condition may partly overlap, for example, the datasets you used in Fig. 5. There, you didn’t delete all donor-associated regions; instead you regressed out their effects. Please consider further comparison of the strategies of batch effect removal and add more discussion about the choices for readers.
5. For Fig. 4, the concept of cell type-specific region difference is quite interesting. While penal (e) shows a global pattern, can the author highlight some examples, e.g., via boxplot?
6. For Fig. 5 (time-dependent immune responses after stimulation), the time label is on the sample level, so essentially one can perform such analysis on the pseudo-bulk level, which can mitigate the imbalance of cell counts between samples. Instead of sample-level time, maybe the author should consider showing an example with a cell-level time label, e.g., pseudotime (also as a more continuous variable).
7. In all three experimental datasets, the main argument about DAR detection (only compared to snapATAC) is detecting more DARs. However, more DAR didn’t mean better overall performance, as it may also have higher false positives. While there is no ground truth, the author may consider comparing the downstream interpretation (e.g., enriched TF or gene sets) of the DAR detected by mcCLR and the DAR detected by other method(s).

Minor:
- Consider adding more discussion about the limitations of this method, e.g., how to handle the imbalance of cell counts between conditions, donors, or time points.
- From Eq. 1 and Methods line 576, it looks like T is region-specific and may not be large. Please plot the distribution of T, e.g., in the supplementary figure.
- Line 606, the difference about p_m and p_{cm} is unclear.
- Eq. 6 (line 636) is wrong. The normalization term as the denominator is missing.
- Eq. 7 is a bit unclear, especially the use of the pi_{ct}.

(Remarks on code availability)
I didn't manage to test the package but the repository looks well documented, with vignettes and examples.

Reviewer #3

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have adequately addressed my comments. Therefore, I have no further comments.

(Remarks on code availability)
The code comes with a README file that details the installation instructions and some examples. I have successfully installed the package and tested some demo.

Reviewer #2

(Remarks to the Author)
I thank the authors for the detailed responses in my comments and resolved the majority of the concerns. However, as responded in my minor point 2, the author clarified that the non-zero read counts are 99.87% are 1 and the rest are only 2 counts. This makes the need for the proposed ordinal logistic regression very minimal. Actually, the proposed model is handling data mostly in a way of logistic regression.

Therefore, I would strongly suggest further comparing to logistic regression as a baseline, where the data will be treated as zeros and non-zero and a common choice for normalization is taking the proportion of non-zero peaks for each cell as a covariate. Similarly, the authors can use the L1 or L2 penalty same as their proposed method. I won't be surprised if the logistic regression works very similarly to the proposed PACS. If that's the case, the authors may need to adapt the claim of their technical contribution and the editors may need to assess if introducing logistic regression for differential accessibility detection has sufficient novelty.

(Remarks on code availability)
I had a very quick review of the code repository. The package seems well organised but the analysis notebooks are relatively limited. I suggest releasing all analysis notebooks/scripts for reproducibility.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This manuscript presents PACS, a statistical model designed for conducting hypothesis testing on single-cell ATAC-seq data. PACS incorporates a cumulative logit regression model to explicitly account for read capture efficiency and various technical or biological factors. The authors shown that PACS effectively mitigates technical and undesirable influences, such as sequencing depth and batch effects, and differences arising from distinct sample origins. When compared with existing methods, PACS demonstrates superior performance in controlling type I error rates and achieving greater statistical power. The utility of PACS is further illustrated through its application in identifying spatially variable regions and time-dependent open chromatin regions. Though this work is potentially interesting, there are several issues need to be addressed.

We thank the reviewer for the summary, and we have addressed the concerns in our point-by-point response below.

Major Comments:

- Several regression based methods for identifying differentially accessible regions (DARs) have been proposed before. For example, Zhang et al. (2021, Cell) developed logistic regression based method for DAR finding. The decision to employ a cumulative logit regression model over more conventional models like the Poisson or negative binomial regression models for read count data necessitates clarification. Are there particular advantages or reasons for this choice?

Thank you for the point. Seurat/Signac and Zhang et al. (2021 Cell) both implemented logistic regression-based models for DAR identification, but with different model specification. To be precise, Seurat used the following model:

$$\text{logit}\big(\text{P}(CT_c = ct_1)\big) = \alpha + \beta_0 D_c + \beta_1 Z_{cm} + \sum_{j=2}^{J} \beta_j F_{cj} \quad \text{(Add. Eq.1)}$$

where CT represents cell type label of a cell, D_c represents the total number of reads of a cell, Z_{cm} represents the observed read count in peak m of cell c, and F_{cj} represents the other biological factors to be controlled for.

In Zhang et al. (2021 Cell), the authors used the following model:

$$\text{logit}\big(\text{P}(Z_{cm} = 1)\big) = \alpha + \beta_0 D_c + \beta_1 CT_c + \sum_{j=2}^{J} \beta_j F_{cj} \quad \text{(Add. Eq.2)}$$

The major difference is that in Zhang et al., the observed read count becomes the random component (dependent variable) while the cell type label becomes part of the systematic component (a covariate) in the regression. In our opinion, (Add. Eq2) is more logically consistent with the biological inference problem; i.e., whether the cell type status governs peak-reads due to differential chromatin state. Despite the difference, both models correct for sequencing depth by incorporating total read count (D_c) as an explanatory variable in regression.

In addition to these existing models, negative binomial regression has been used for detecting differentially expressed genes (DEGs) in RNA-seq data, and edgeR is one frequently used package with carefully crafted models for RNA-seq data. In edgeR, the sequencing depth is normalized by library size and trimmed mean of M-values (TMM) in the regression model.

The three methods summarized above are all based on the generalized linear model (GLM) framework. The first key difference between PACS and these existing models is that in PACS, we derived a missing-corrected GLM model, mcCLR, that **generalizes the GLM framework** to further incorporate sample-specific (i.e., cell-specific) incomplete capturing of data. We reason that to explicitly model the data capturing process is more desirable than regressing out total read count or normalizing by size factor, for the following reasons:

1. Under the null hypothesis that a peak is not DAR, Z_{cm} is orthogonal with CT_c. However, Z_{cm} is correlated with D_c (higher sequencing depth, higher probability of observing accessible chromatin) and CT_c is also correlated with D_c if the DAR set is not empty. The existence of these two correlations causes collider bias in the model, resulting in inflated type 1 error rate as shown in the evaluations (**Fig. 2a-i**). In PACS, D_c is not a covariate in the regression model, therefore it is not subject to the collider bias.
2. In edgeR model, the validity of TMM-based correction relies on the assumption that between **any pair of single-cells**, the majority of peaks (or, at least more than half of the peaks) are not differential. In single cell data where data are extremely sparse, this assumption no longer holds, and the estimation of cell-specific normalizing factor will be unreliable.
3. The data generation process for ATAC-seq and RNA-seq is different. In the standard ATAC-seq experiment, two adjacent Tn5 tagmentation events are required for generating one ATAC fragment, and the PCR amplification step further requires the fragment to have one forward and one reverse primer (see a comprehensive review by Adey, *Genome Res.* 2021). Assuming the Tn5 tagmentation event to be a Poisson process, we have shown that the resulting ATAC-seq count will follow a size-selected signed Poisson distribution, and the negative binomial distribution is not accurate (See our work Miao and Kim, *Nat. Methods.*, 2024). Our model estimates separate parameters for the cumulative ordinal counts rather than incorrectly assuming iid counts, which better controls bias as evidenced by our control of Type I error.

In addition, Seurat/Signac by design cannot be applied to study the effect of other covariates on the accessibility, as shown in (Add. Eq.1). More discussions of Seurat model is summarized in the **Supplementary Note** of the manuscript.

To compare the performance of different methods more quantitatively, in the revised manuscript, we have included Zhang et al., (2021, Cell) and edgeR in the comparison (**Fig. 2a-i**). As Zhang et al. later developed snapATAC2 with the same DAR framework, we label the method as snapATAC2.

PACS remains the most powerful method among those that can control type 1 error under the specified significance level.

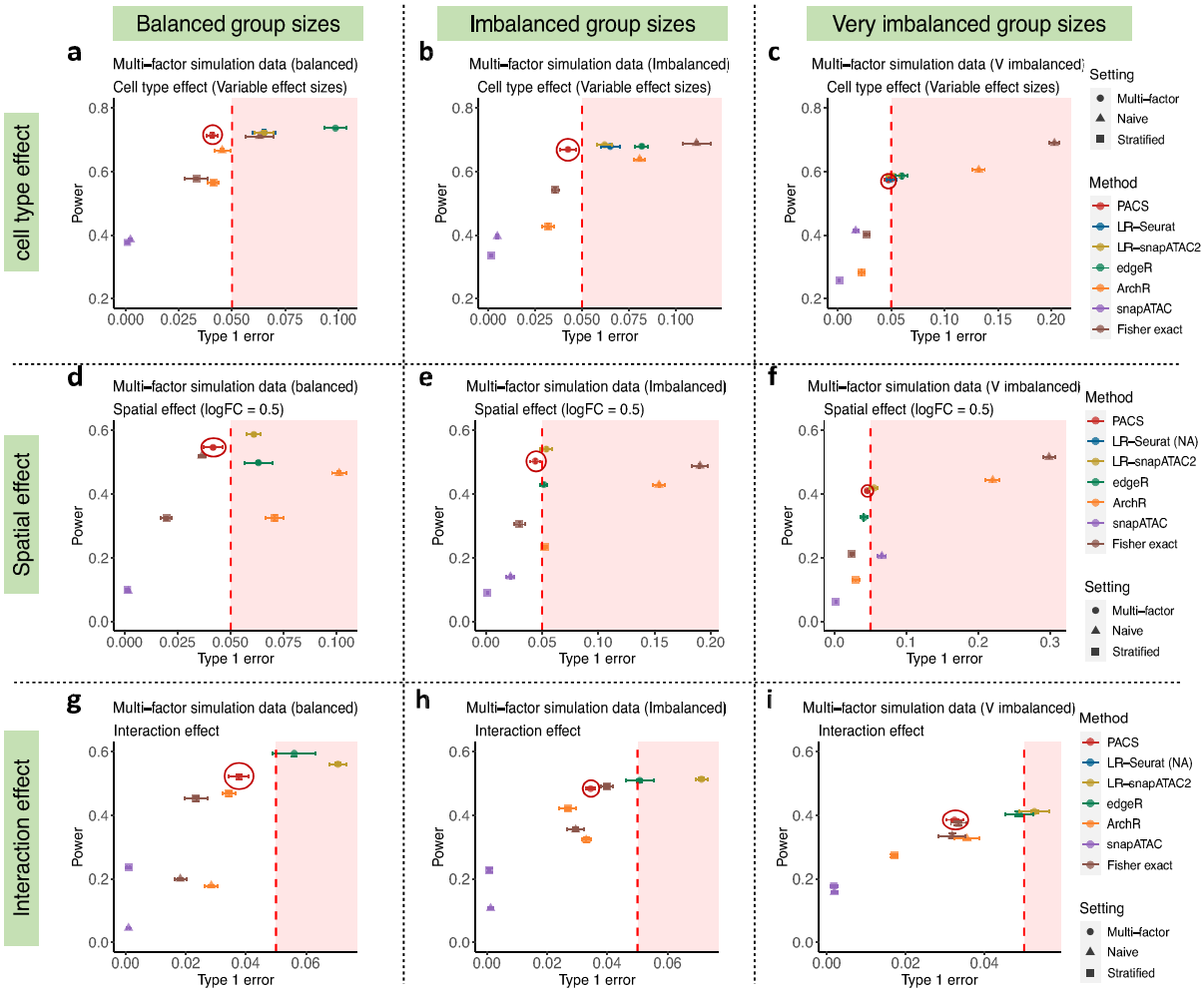


Figure 2. Compound hypothesis testing with PACS is sensitive and specific.

a-i. Type I error and power of different methods evaluated with two-factor simulation data. All tests are two-sided. The best-performing method is indicated by a circle with matched color. Note that Seurat does not support testing factors other than cell type effect. LR: logistic regression.

- A central aspect of PACS is its ability to adjust for confounding factors while focusing on variables of interest. The effectiveness of PACS in this regard is crucial. Nonetheless, the UMAP in Figure 3b suggests potential residual batch effects even after applying PACS. It would be valuable to understand the extent to which PACS can ameliorate batch effects, especially in comparison to methods like Harmony or PeakVI.

Thank you for the points. We first note that existing methods, such as Harmony or PeakVI, only correct the batch effect at the latent space level, but the batch effect is not corrected in the original data matrix. For downstream analyses such as motif enrichment (e.g., using chromVAR or HOMER), peaks with strong batch effect can still bias the results. In contrast, PACS offers an alternative approach by identifying features with significant batch effect. Users can then choose to filter out these peaks (recommended for motif enrichment or trajectory inference) or include batch as a covariate while retaining these peaks (recommended for differential tests with PACS). In brief, PACS provides complimentary functionality that is not available with existing methods like Harmony or PeakVI.

We acknowledge that the UMAP in Figure 3b and our analyses of batch mixing score in PCA space suggest incomplete removal of batch effect. This outcome is partly due to our stringent criteria for defining "peaks with significant batch effect", where only peaks with adjusted P value ≤ 0.05 were removed. Consequently, a substantial number of peaks with moderate batch effect were retained. As a comparison, we examined the effect of using an unadjusted P value of 0.05 as the cutoff, which resulted in a better batch mixing score of 0.417 (**Supplementary Figure 6**). However, adopting a looser criterion for removing batch effect peaks may inadvertently eliminate peaks with both biological effect (e.g., cell type effect) and batch effect, which is undesirable for downstream analyses.

As noted by other studies (e.g., Tran, H.T.N. et al., *Genome Biol.*, 2020; Luecken et al., *Nat. Method.*, 2021; and Li et al., *BMC Bioinfor.*, 2022), there is always a tradeoff between batch correction and loss of biological effects. The decision must be made with careful considerations. In our study, we aimed to correct batch effect only for clustering major cell types and visualization. Using an adjusted P value of 0.05 improved clustering and yielded reasonably mixed cells in low-dimensional spaces (PCA and UMAP), demonstrating utility for these tasks.

We emphasize that for tasks like determining the set of DAR, we do not exclude any peak but rather keep all peaks and regress out the batch effect to obtain an unbiased estimation of true cell type effect. For tasks like **clustering or visualizing a 2-D embedding** where the goal is to have a global overview of the data structure, information loss is both acceptable and inevitable, the approach of excluding significant batch effect peaks can be used, provided that the tradeoffs discussed above are understood.

The analyses presented in Fig. 3b serve as a proof-of-principle that excluding peaks is a valid option for **clustering** at the major cell type level or **visualizing** these major cell types in a 2-D embedding. Importantly, downstream analyses like DAR are still based on the raw count matrix.

- The integration of biological replicate information is pivotal for ensuring the reliability and reproducibility of results during DAR finding. Methods based on bulk or pseudo-bulk data naturally incorporate this aspect when replicates are available. It raises the question of whether PACS can also leverage biological replicate data during hypothesis testing.

Thank you for the insightful comment. PACS is a flexible method and can incorporate biological replicates as covariate in the model, as shown in our real data analyses in the brain.

Theoretically, if the dataset contains multiple samples, mcCLR model for the test can be specified as

$\text{logit}(\text{P}(Y_{cm} \geq t)) = \alpha^{(t)} + \beta_1 \text{CT}_c + \beta_2 \text{Sample}_c + \sum_{j=3}^{J} \beta_j F_{cj}$, where $\text{P}(Z_{cm} \geq t) = \text{P}(Y_{cm} \geq t) q_c;\ t \in \{1,2,\dots,T\}$

Here, q_c is the capturing probability for a cell c, $\text{P}(Y_{cm} \geq t)$ is the sampling probability of cells with accessibility level greater than or equal to t, $\alpha^{(t)}$ is the intercept term in the t^{th} cumulative logit, β_1 is the coefficient for cell type (CT), β_2 is the coefficient for sample label, and F_{cj} represent other factors (covariates) to be included in the model, and β_j is the coefficient for the j^{th} factor.

Practically, PACS accepts a “formula” input, which is consistent with all existing regression packages in Rscript. Therefore, users can specify the sample label as a covariate by using the formula. Below is an example code block demonstrating how to include the "sample" label as a covariate to be adjusted by PACS:

Code block 1. Example codes for running PACs while adjusting sample label

```
p_vals <- pacs_test_sparse(
  covariate_meta.data = meta.data,
  formula_full = ~ factor(cell_type) + factor(sample),
  formula_null = ~ factor(sample),
  pic_matrix = data_mat,
  cap_rates = capturing_probability
)
```

Minor Comments:

- The section titled "PACS identifies kidney cell type-specific regulatory motifs and allows for direct batch correction" (Line 277) implies motif analysis, which does not appear to be discussed in the manuscript.

Thank you, we have corrected the term “motif” to “genes”, as we focus on the enriched genes to evaluate the performance of PACS.

- The statement "We first examined the marginal effect of brain regions on chromatin accessibility, holding other factors constant (Methods)" (Line 332) deserves further explanation. It should be presented in a manner that is more accessible to biologists.

Thank you for raising the question regarding clarity of our description. In response to your comment, we have revised the sentence in question to better explain “marginal effect” for a broader biological audience. The original sentence has been rephrased to:

“We first examined how different brain regions affect chromatin accessibility across all cell types. To accurately capture this relationship, we incorporated other influential factors, such as donor variability, as covariates in our regression model. This methodological approach allows us to assess the specific effects of brain regions while accounting for potential confounding factors (**Methods**).”

Reviewer #1 (Remarks on code availability):

The code comes with a README file that details the installation instructions and some examples. I have successfully installed the package and tested some demo.

Thank you, during the revision process, we have further updated our GitHub page and added more interactive vignettes for future users.

Reviewer #2 (Remarks to the Author):

In the manuscript, the author proposed a statistical method, mcCLR (missing-corrected cumulative logit regression model), to detect the differential accessible regions (DAR) from single-cell ATAC-seq data. Undoubtedly, the DAR analysis is a critical step for biological discovery from scATAC data. From simulation analysis, the authors showed mcCLR has improved power with good false positive controls, compared to existing methods. Later, they

also demonstrated the application of mcCLR in three experimental datasets where ground truth is generally lacking. Overall, the method shows good promise while it is a bit surprising that such multi-factor analysis is not well supported in the existing tools. Here are a few concerns about the manuscript and some comments for the authors to consider.

Major:

1. The primary contribution of this work is the mcCLR, which is a variant of Ordinal Logistic Regression. It seems the term "cumulative logit regression" is less commonly used than "Ordinal Logistic Regression (OLR)". I would suggest the author either change it to the more commonly used term or at least mention the alternative term in the results. Also, citing a few examples of OLR in biomedical research (e.g., PMID: 1468011, PMID: 9429194) would help the readers to understand it better.

Thank you for the comment. We have included the term "ordinal logistic regression" in the manuscript text, and cited the two references as well as a statistical book (Agresti, *Categorical data analysis*, Wiley 2013) for readers to better understand the context and the model.

2. It is surprising that only one tool supports generalized linear regression models here (Seurat via logistic regression). For scRNA-seq data, there are multiple GLM methods supporting multi-factor analysis. Since mcCLR directly models the scATAC read counts, I strongly suggest the authors add one scRNAseq-based method for comparison. Particular example is the negative binomial regression with library size (e.g., via DE-seq or edgeR), which has a library size factor, reflecting the function of q_c, and negative binomial supports dispersion. It looks like snapATAC already used edgeR but seems not flexible; the author may directly use edgeR.

Thank you. We have included the edgeR method in our comparison. As shown in **Fig. 2a-i**, edgeR fails to control type 1 error in most cases and PACS remains the most powerful method among those that can control type 1 error. We note that, although the q_c parameter in PACS helps account for sequencing depth disparity among cells, it has several advantages compared with using the library size factor or using trimmed mean M-value (TMM) in a regression model. Briefly, our probability-based model avoids the collider bias in regression model, and for single cell data, the assumption behind TMM is invalid. Please see our response to the first comment from reviewer #1 for a more detailed discussion.

3. When comparing cell type label transfer with Naïve Bayes, the article says "Naïve Bayes does not assume missing data; thus, it ignores the cell-specific capturing probability." Does this refer to q_c in Eq.1? It looks similar to a sequencing depth normalization (e.g., total counts or proportion of non-zero regions), which the naïve Bayes model may be able to consider as an additional variable. Also, it would be better to add more details about the naïve Bayes model here.

The data missing part is the q_c of our Eq.1, which represent the cell-specific capturing probability of peaks. The standard Naïve Bayes model only takes binary input and cannot take quantities such as sequencing depth as a variable in the model, so it cannot achieve depth normalization. The reason that we compare our method against Naïve Bayes method is that the Naïve Bayes model also relies on the Bayes rule to infer cell type, and it has been widely used in supervised learning and in predicting cell type labels for scRNA-seq data (e.g., Iyer et al., *J. Clin. Med.*, 2020).

The "Naïve" part of the Naïve Bayes method refers to the assumption of the features being conditionally independent of every other feature given the label. The Naïve Bayes model estimates within each group, the probability of count 1 for each feature (i.e., P(Y_cm = 1)) from reference data. After parameter estimation (or, the "learning" step), the model applies the Bayes rule to estimate the probability of each label for new datasets. Specifically, in our context, it assumes that the accessibility of each peak is independent of the accessibility of other peaks, given the cell type. Using the same notation as in the methods section , the parameter estimation procedure for Naïve Bayes method is

$$p_{C_g m}^{(1)} = \frac{\sum_{c \in C_g} I(z_{cm} \geq 1)}{|C_g|} \quad \text{(Eq. 7)}$$

Where $|C_g|$ represents the group size of (i.e., number of cells in) cell type g. With a new set of observations $Z'_{C' \times M}$, the probability of their corresponding cell type labels, $h(Z'_{c*})$ can be estimated using the following equation

$$\begin{aligned} \mathrm{P}(h(Z'_{c*}) = g | Z'_{c*}) &\propto \mathrm{P}(Z'_{c*} | h(Z'_{c*}) = g)\mathrm{P}(h(Z'_{c*}) = g) \\ &\propto \mathrm{P}(h(Z'_{c*}) = g) \prod_{m=1}^{M} \left(p_{C_g m}^{(1)} \right)^{Z'_{cm}} \left(1 - p_{C_g m}^{(1)} \right)^{1 - Z'_{cm}} \end{aligned} \quad \text{(Eq. 8)}$$

We have included these additional details in the **Methods** section of the manuscript.

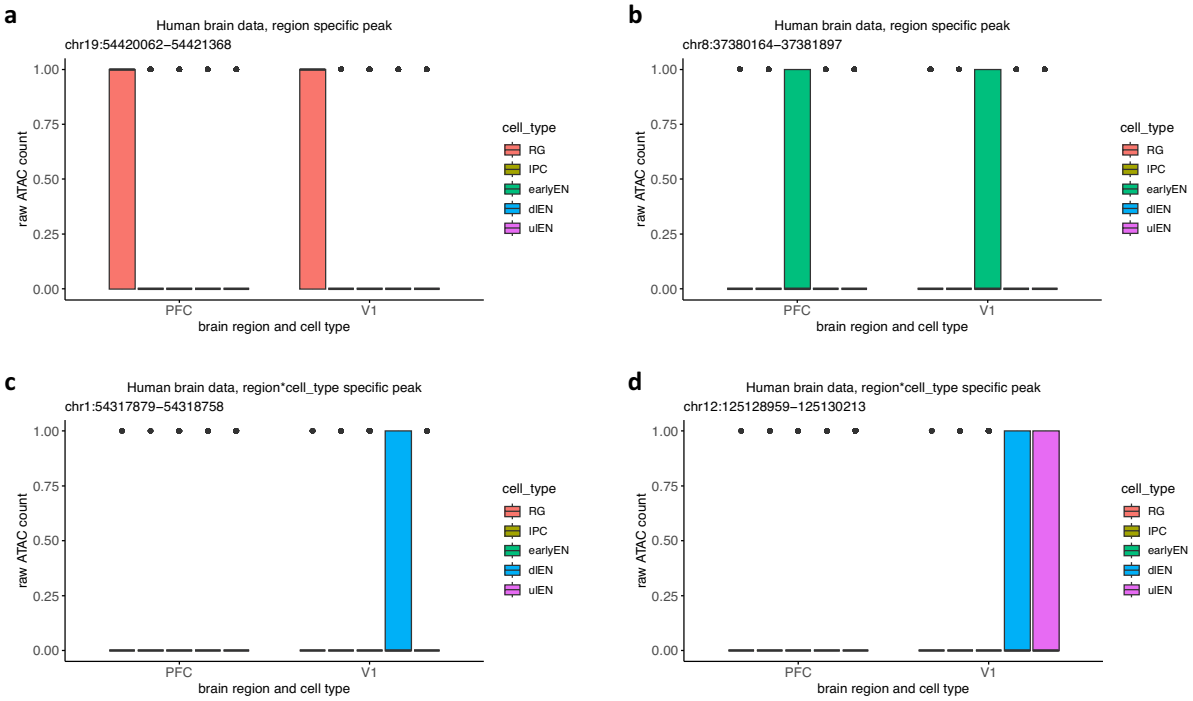
4. In batch-effect correction, the strategy of removing batch-effect peaks is very different compared to other methods that preserve all feature sets. Simply removing batch correlated peaks will likely introduce the risk of over-correction, since batch and condition may partly overlap, for example, the datasets you used in Fig. 5. There, you didn't delete all donor-associated regions; instead you regressed out their effects. Please consider further comparison of the strategies of batch effect removal and add more discussion about the choices for readers.

Thank you and we agree with the points raised here. Indeed, the strategy of excluding significant batch effect peaks should only be used for tasks where a coarse-grained global overview is needed, such as clustering major cell types or visualizing these coarse-grained cell types in a 2-D embedding. For tasks such as identifying motifs associated with DARs, it cannot be done by batch correction in latent space. Thus, our approach provides a complementary approach that can be applied when there are feature level inference problems.

As our response to the comments by reviewer #1 to a related question, there exists a trade-off between batch effect correction and biological effect preservation in choosing significance level for significant batch effect peak exclusion. (Please see our response to the second comment of reviewer #1). We have included the discussions in the manuscript.

5. For Fig. 4, the concept of cell type-specific region difference is quite interesting. While penal (e) shows a global pattern, can the author highlight some examples, e.g., via boxplot?

Thank you. We have included **Supplementary Fig. 9** to show examples of significant cell type-specific DARs (a-b) and significant cell type and brain region interaction DARs (c-d). The significant interaction effect suggests that the cell type-specificity of chromatin accessibility is dependent on the specific brain region. For example, the chromatin region '*chr1:54317879-54318758*' is deep layer excitatory neuron specific only in the V1 brain region. We note that the brain data were processed with snapATAC so the data were binarized count, explaining the configuration of the boxplot.



Supplementary Fig. 9.
a-d. Boxplot showing the raw ATAC count for some significant cell type-specific DARs (a-b) or significant cell type and brain region interaction DARs (c-d). Center line (the thicker line) in box plot represents median and the lower and upper hinges correspond to the first and third quartiles.

6. For Fig. 5 (time-dependent immune responses after stimulation), the time label is on the sample level, so essentially one can perform such analysis on the pseudo-bulk level, which can mitigate the imbalance of cell counts between samples. Instead of sample-level time, maybe the author should consider showing an example with a cell-level time label, e.g., pseudotime (also as a more continuous variable).

Thank you, as a regression-based framework, our model is flexible and can incorporate the cell-level pseudo-time information. We note that with sample-level information, both pseudo-bulk approach and our approach is valid, but the two analyses view the data differently. In pseudo-bulk analysis, the aggregated read counts in one cell type of a single individual are viewed as a single "sample"; whereas in PACS, each single cell is viewed as a single sample. So, the two approaches have different null (and alternative) hypotheses. Despite the difference, PACS provides a more general framework for two reasons. (1) As noted by the reviewer, PACS can incorporate cell-level information such as pseudo-time in the model; and (2) When there is only one replicate in each time point, pseudo-bulk type of analyses will not be valid.

For the temporal PBMC treatment dataset, we decided to apply PACS to test the real time (0h, 1h, and 6) instead of pseudo-time because the analysis with real time provide more interpretable results and is not subject to the issue of data double dipping. In the original publication, the authors constructed the pseudo-time by taking the 50 nearest neighbors for each cell, and computed the average time point of these neighbors as the pseudo-time. We reason that such a pseudo-time label is highly concordant with the real time. Moreover, as pointed out in (Neufeld et al., *Biostatistics* 2022), inference after latent variable estimation (here, pseudo-time being the latent variable) cause the issue of inflated type 1 error due to double dipping of the data. Such an issue of double dipping will occur in testing the pseudo-time effect, but not the real time effect.

In sum, for datasets with real time labels, we recommend applying the PACS to test the effect of real time, and PACS is flexible and can incorporate cell-level input such as pseudo-time.

7. In all three experimental datasets, the main argument about DAR detection (only compared to snapATAC) is detecting more DARs. However, more DAR didn't mean better overall performance, as it may also have higher false positives. While there is no ground truth, the author may consider comparing the downstream interpretation (e.g., enriched TF or gene sets) of the DAR detected by mcCLR and the DAR detected by other method(s).

Thank you and this is a critical point raised by the reviewer. In the model evaluation/comparison part, we used label permutation methods to evaluate the false positive rate (type I error rate) under the significance level of 0.05. Properly controlling the Type I error, guards against false positives and our assessments of power (more DAR) follows only after correct Type I error control. The results suggest that for both simulation data (Fig. 2a-i) and real data (Supplementary Fig. 5a-f), PACS can control false positive rate below 0.05 across all settings/datasets. Thus, the identified DARs by PACS are unlikely due to false positives at the specified Type I error.

In addition, the original publication of the kidney and brain dataset used snapATAC to conduct DAR analysis, and confounding variables (such as sample-specific batch effect) are not considered in the model and thus are not adjusted. As shown in our evaluation section, snapATAC relies on a pre-specified heuristic variation parameter for edgeR, resulting in substantial power loss. Moreover, the unadjusted covariates may increase false discoveries in the test. Taken together, it is expected that snapATAC shows lower power and through comprehensive evaluation, we are confident that PACS has a reliable false positive rate control.

To further support our identified peaks, we conducted TF enrichment analysis for the DAR set identified by PACS but not by snapATAC. The top enriched motifs are highly concordant with the output using the full set of peaks, such as *MEF2B*, *MEF2D* in the V1 region, and *MEIS1* in the PFC region, and shared *NeuroG2* and *Olig2* enrichment at different genomic regions (**Supplementary Tables 13-14**). The concordance of enriched motifs between the full DAR set and the PACS-specific set shows that additional peaks identified by PACS is consistent with but increases the information set found in the PACS-snapATAC common set and supports the reliability of the PACS method.

Minor:
- Consider adding more discussion about the limitations of this method, e.g., how to handle the imbalance of cell counts between conditions, donors, or time points.

We thank the reviewer for the suggestion. We note that the issue of sample imbalance of cell counts are issues of data, but not a limitation of the PACS method per se. Across all existing methods (including edgeR as mentioned by the reviewer), having imbalanced cell counts will result in decreased power (as illustrated in our comparison from simulation data). When the cell counts are extremely imbalanced, ArchR suffer from the most power decrease, as ArchR relies on using random down sampling, which will be governed by the smallest group. For example, if the cell counts in two groups are 200 and 1800, at least 1600 cells from the second group will be removed for the test, resulting in considerable unused information. However, as a regression-based method, all cells contribute to the differential test of PACS, so it is not as restricted by imbalanced cell counts as ArchR. We have added this and additional discussion of the limitation

of our method including effects of how much quantitative information is in the single cell data (e.g., due to amount of sequencing).

- From Eq. 1 and Methods line 576, it looks like T is region-specific and may not be large. Please plot the distribution of T, e.g., in the supplementary figure.

Thank you. T represents for a given chromatin region (peak), the largest integer with substantial occurrence over cells. T is region-specific and its distribution is associated with overall sequencing depth of the data. For most peaks, the frequencies of different reads decrease exponentially, so indeed T is not large. Among all the datasets we evaluated, there are fewer than 50 peaks with $T \geq 3$. Our model allows for larger T for model flexibility considerations. In the kidney data, T = 1 for 300339 peaks, and T = 2 for 416 peaks.

- Line 606, the difference about p_m and p_{cm} is unclear.

Thank you, here, we applied the coordinate descent algorithm for each cell type group separately, and we omit the subscript C_f and superscript (1) for $p_{C_f m}^{(1)}$. We included some clarifying descriptions in the methods section.

- Eq. 6 (line 636) is wrong. The normalization term as the denominator is missing.

Thank you for pointing this out, we misplaced “proportional to” sign with an “equality” sign in the equation. We have corrected the notation in the revised manuscript.

- Eq. 7 is a bit unclear, especially the use of the pi_{ct}.

Thank you. Eq. 7 shows the log likelihood function of our model, which serves as a loss function for parameter estimation. The latent probability of having t fragments in cell c is $P(y_c \geq t) - P(y_c \geq t+1)$, which we define as π_{ct}. To incorporate the cell-specific data capturing probability q_c, we define $\tilde{\pi}_{ct}$ to represent the probability of observing t fragments in cell c and is defined by $\tilde{\pi}_{ct} = \pi_{ct} q_c$.

Reviewer #2 (Remarks on code availability):

I didn't manage to test the package but the repository looks well documented, with vignettes and examples.

Thank you and we have tested our package internally, and we were able to run the method with using different computers and in Google Colab.

Reviewer #3 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thank you for co-reviewing our manuscript with the other reviewer, we hope we have satisfactorily addressed your comments.

Response to Reviewers' comments

Reviewer #1 (Remarks to the Author):
The authors have adequately addressed my comments. Therefore, I have no further comments.

Reviewer #1 (Remarks on code availability):
The code comes with a README file that details the installation instructions and some examples. I have successfully installed the package and tested some demo.

We thank the reviewer with their constructive feedback. We are pleased to hear that our revisions have addressed your comments satisfactorily.

Reviewer #2 (Remarks to the Author):

I thank the authors for the detailed responses in my comments and resolved the majority of the concerns. However, as responded in my minor point 2, the author clarified that the non-zero read counts are 99.87% are 1 and the rest are only 2 counts. This makes the need for the proposed ordinal logistic regression very minimal. Actually, the proposed model is handling data mostly in a way of logistic regression.
Therefore, I would strongly suggest further comparing to logistic regression as a baseline, where the data will be treated as zeros and non-zero and a common choice for normalization is taking the proportion of non-zero peaks for each cell as a covariate. Similarly, the authors can use the L1 or L2 penalty same as their proposed method. I won't be surprised if the logistic regression works very similarly to the proposed PACS. If that's the case, the authors may need to adapt the claim of their technical contribution and the editors may need to assess if introducing logistic regression for differential accessibility detection has sufficient novelty.

We thank the reviewer for this important comment. Based on our consultation with our handling editor, we have surveyed more datasets for appropriateness of our method and carried out additional analyses as we describe below.

We first note that our proposed Firth regularization is designed to handle the issue of separation due to sparse scATAC-seq data. The suggested L1 penalty, also known as LASSO, is a common method for handling high-dimensional data to prevent overfitting. Although LASSO can shrink the coefficient towards zero, its effect is highly dependent on the hyperparameter, λ, and there is a scarcity of efficient approach to hyperparameter tuning for addressing the issue of separation. In fact, depending on the choice of λ, the LASSO method will result in varying type 1 error and power values. Given these limitations, we believe our proposed Firth regularization is a more appropriate model choice for sparse scATAC-seq data.

Next, the comment raises two subsequent questions: (1) What are the advantages in adopting the cumulative (or ordinal) logistic framework over the logistic (i.e., binary) framework? and (2) What are the advantages in introducing the probability model of data capturing over the conventional approach of including the total read count as a covariate in the regression? Our responses to these two questions are as follows.

(1) Cumulative vs binary model

We clarify that in the kidney adult data analyzed in our study, 86.93% counts are one and the rest are two or higher. However, since the full cumulative model calculation can be computationally involved, we have a QC threshold wherein only the peaks whose non-binary counts constitute 25% or more of the total counts are considered for the full model. Otherwise, we simplify to the binary case. With such a QC threshold, cumulative model might be applied to a smaller proportion of the peaks when the dataset has relatively low sequencing coverage. We note that the kidney dataset referenced by the reviewer was collected in 2018, and since then, advancements in sequencing technologies have significantly improved the quality of scATAC-seq data, as suggested by a higher proportion of counts that are greater than or equal to two in a compilation of datasets we examined (now included as **Appendix Table 1**).

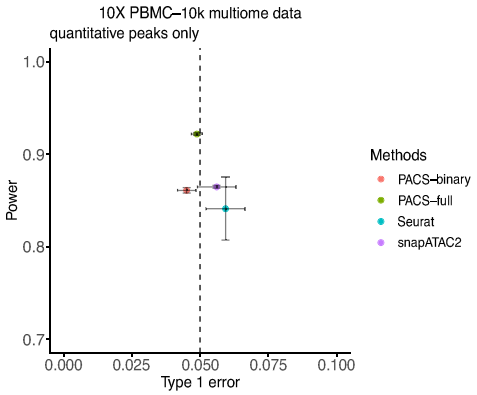
Appendix Table 1. Proportion of quantitative information in relation with technology and sequencing depth.

Technology and tissue	N cells after QC	N peaks after QC	median in-peak fragment counts per cell	proportion of counts greater than one over non-zero counts	Reference

10X mouse kidney adult	14458	251562	4458	13.07%	Miao et al., Nat. Comm. 2021
10X mouse kidney P0	9286	270946	7818.5	16.40%	Miao et al., Nat. Comm. 2021
10X PBMC-5k snATAC-seq	4623	135377	11984	26.52%	https://www.10xgenomics.com/datasets
10X PBMC-10k multiome	11234	139729	12155	26.57%	https://www.10xgenomics.com/datasets
10X BMMC multiome (donor 2 institute 1)	6740	116490	7213.5	21.05%	https://www.10xgenomics.com/datasets
dsc-ATAC-seq mouse brain (mouse 2 channel 1)	3098	454047	12990	36.11%	Lareau et al., Nat. Biotech. 2019
dsc-ATAC-seq mouse brain (mouse 2 channel 2)	4011	449657	9316	25.89321%	Lareau et al., Nat. Biotech. 2019

In fact, we originally considered the simpler binary treatment but developed the full model because the simpler model does not fully capture the nuances of possible data. We implemented the current PACS model, which handles **both near-binary data and more quantitative data,** since the binary case is a subset of the ordinal data.

To further examine the question of added value in quantitative treatment as noted by the reviewer, we selected the 10X Genomics multiome PBMC-10k dataset with ~27% non-binary counts for additional study. We compared our full model (cumulative logistic + logistic) against the simpler logistic regression version of PACS. In this dataset, due to the higher overall quantitative counts, 12.3% peaks have substantial proportion of non-binary counts to be considered for the cumulative model (i.e., T>=2 using the notation in the manuscript). We then quantitatively compared the type 1 error and power of the full model with the reduced model (that is, with the binarized data). We observe an increase in power as expected (average increase in power over five repetitions from 0.86 to 0.92, see **Supplementary Fig. 6 below**). We note that the slight increased type 1 error of PACS from binary to non-binary could be associated with our approximation in parameter estimation (see **Methods**). Finally, for completeness, we also tested Seurat and snapATAC on the same dataset; as in our other evaluations, both methods had higher Type I error than the nominal critical value. These new results have now been added to the manuscript.



Supplementary Fig. 6. Type 1 error and Power comparison on the quantitative peaks between PACS-full, PACS-binary, Seurat and snapATAC2 using the 10X Genomics PBMC-10k dataset.

(2) Probability model of accessibility vs using total read counts as a covariate in regression

In PACS, through our probability model, we separated the effect of incomplete data capturing from the latent accessibility. The latent accessibility is an interpretable quantity that represents the probability of a peak being accessible in a given cell type (group), independent of experimental data capturing effects.

This interpretable probability model enables PACS not only to conduct statistical test of differential accessible regions (DARs) but also to transfer cell type label annotations. In our main text, we described the corresponding approach and demonstrated improved performance over existing naïve Bayes-based methods. However, using total read counts as a covariate cannot achieve this task.

We have updated the manuscript accordingly to include the discuss of quantitative treatment of scATAC-seq data, and we hope the reviewer will agree with the technical advantages and the novel applications allowed by our proposed method.

Reviewer #2 (Remarks on code availability):

I had a very quick review of the code repository. The package seems well organised but the analysis

notebooks are relatively limited. I suggest releasing all analysis notebooks/scripts for reproducibility.

We thank the reviewer for this comment, in our updated package, we have released the analyses codes for the real data analyses as well as the simulations. In addition, we included new functionalities to run our code using Seurat object as input so people can seamlessly integrate PACS in their workflow.