PAC allows comprehensive dissection of multiple factors governing chromatin accessibility from snATAC-seq data

Zhen Miao[1,2], Jianqiao Wang[3], Kernyu Park[2], Da Kuang[4], and Junhyong Kim[1,2,4,†]

[1]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[4]Deptartment Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

[†]Corresponding author

Correspondence:
Junhyong Kim, Ph.D.
Christopher Browne Distinguished Professor of Biology
Secondary Professor, Department of Computer and Information Science
University of Pennsylvania
433 S. University Avenue
Philadelphia, PA 19104

Abstract:

Single nucleus ATAC-seq (snATAC-seq) experimental designs have become increasingly complex with multiple factors that might affect chromatin accessibility, including genotype, cell type, tissue of origin, sample location, batch, etc., whose compound effects are difficult to test by existing methods. In addition, current snATAC-seq data present statistical difficulties due to their sparsity and variations in individual sequence capture. To address these problems, we present a zero-adjusted statistical model, Probability model of Accessible Chromatin of Single cells (PACS), that can allow complex hypothesis testing of factors that affect accessibility while accounting for sparse and incomplete data. For differential accessibility analysis, PACS controls the false positive rate and achieves on average a 17% to 122% higher power than existing tools. We demonstrate the effectiveness of PACS through several analysis tasks including supervised cell type annotation, compound hypothesis testing, batch effect correction, and spatiotemporal modeling. We apply PACS to several datasets from a variety of tissues and show its ability to reveal previously undiscovered insights in snATAC-seq data.

Main:

Single nucleus ATAC-seq (snATAC-seq) is a powerful assay for profiling the epigenetic status of open chromatin in individual cells[1,2], and has been applied to study gene regulation across tissues and under various conditions, including homeostasis[3,4,5], development[6,7] or disease[8,9]. The cis-regulatory elements (CREs), modulated by nucleosome turnover and occupancy[10], display variable accessibility across cells. In a cell, the dynamic of CREs activities are dependent on various physiological factors such as cell type[1,3], developmental state[6,7], spatial location of the tissue[11,12], as well as interaction of these factors with genetic variation[13,14]. Identifying the sets of elements whose accessibility is governed by genetic, developmental, and physiological factors is essential in understanding the cis-regulatory codes of biological processes[15,16]. Moreover, prioritizing regulatory element is instrumental for identifying relevant cell types associated with complex traits[5,17–20]. Dissecting genetic, epigenetic, and environmental factors requires analysis methods that allows compound hypothesis testing. However, the discrete and sparse nature of ATAC-seq data presents technical challenges for existing approaches. Here, we establish a model framework

and a new statistical method that allows complex hypothesis testing for single nucleus ATAC-seq data.

Among all the factors that drive the accessibility of CREs, only some factors are experimentally controlled. In a typical single cell experiment, the collection of cells is a random sample of a cell's variable states over the unknown factors (e.g., cell cycle stage, metabolic cycles) while controlling for the known factors (e.g., tissue, location, batch). We note that sometimes the values of these factors are estimated from the data, such as unsupervised inference of cell type labels or time-sequences. Nevertheless, as the data are sampled over unknown microstates and stochastic molecular processes, the latent accessibility of a CRE should be considered as a random variable, even without experimental variability.

A central theme in snATAC-seq data analyses is to understand the dynamic of CRE across cell types and biological conditions. This is typically achieved by conducting a statistical test for Differential Accessible Regions (DARs) between different conditions of hypothetical causal factors that might govern the chromatin structure (e.g., treatment vs control). However, most existing approaches for identifying DARs rely on pairwise testing of one putative causal factor at a time (e.g., ArchR[21], snapATAC[22], or contingency table-based approaches) that does not allow for simultaneously testing multiple factors or testing continuous variables (such as temporal effects). A common practice in literature is to ignore other covariates while testing the factor of interest. However, as schematically diagramed in Fig. 1, testing one factor at a time can create false negatives and false positives in DAR of main factors and cannot test for interactions. Although Seurat/Signac[23,23] uses the logistic regression framework which, theoretically, can conduct multi-factor testing, its current implementation creates spurious dependency problems that result in mis-calibrated tests (Supplemental Notes).

Besides the key problem of inability to hypothesis test multiple causal factors together, there are two additional unresolved technical problems: (1) snATAC-seq data displays heterogeneous sequencing coverage in each cell, so the cells are not directly comparable; (2) snATAC-seq data is sparse but also contains quantitative information at each peak[24]. To address these challenges, we present a new statistical framework, missing-corrected cumulative logistic regression (mcCLR)

for the analysis of snATAC-seq data with modified Firth regularization[25,26] to account for data sparsity.

With this statistical framework, we present our Probability model of Accessible Chromatin of Single cells (PACS), a toolkit for snATAC-seq analysis. PACS allows for complex compound analysis tasks in snATAC-seq data, including cell type classification, feature-level batch effect correction, and spatiotemporal data analysis. With simulated data and real data, we show that PACS effectively controls false positives while maintaining sensitivity for model testing. We apply PACS to a mouse kidney dataset, a developing human brain dataset, and a time-series PBMC treatment dataset, all of which have complex study designs, to demonstrate its capability to model multiple sources of variations for hypothesis-driven biological inference.

Results:

Probabilistic model of accessible peaks and statistical test framework

In the PACS framework, we model the accessibility state of CREs in a single cell as a function of predictive factors such as cell type, physiological/developmental time, spatial region, etc. We use a matrix $F_{C\times J}$ to represent these variables in each cell, where C is the number of cells and J is the number of predictive variables (including dummy variables). Let $Y_{C\times M}$ represent an integer-valued snATAC-seq count matrix across C cells and M genomic regions. For empirical ATAC-seq data, these regions M are determined by data-dependent peak calling, where peaks are regarded as the set of candidate CREs[27,28]. As snATAC-seq can recover quantitative information on the density and distribution of nucleosomes[24,29], we use integer values $Y_{cm} \in \{0,1,2,\dots\}$ to represent the level of accessibility. Existing pipelines diverge in the quantification of snATAC-seq counts, and we propose to use the paired insertion count (PIC) matrix as a uniform input for downstream analyses[24]. For standard snATAC-seq experiments, PIC counts follow a size-filtered signed Poisson (ssPoisson) distribution for a given Tn5 insertion rate[24]. Thus, the integer-valued PIC counts are observed measurements of the latent Tn5 insertion rates and chromatin accessibility (**Fig. 1**, upper panel). Based on this latent variable perspective we developed a proportional odds cumulative logit model to decompose the cumulative distribution of Y_{cm} by its predictive variables F_{c*}.

With cell-specific nucleosome preparation and sequencing depth, the (observed) snATAC-seq output may miss sequence information from certain accessible chromatin (**Fig. 1**, middle panel). Here, we use $R_{C\times M}$, with binary values, to represent the read recovery/capturing status for each cell and region. This matrix encapsulates all the experimental factors (Tn5 activities, sequencing depth, etc.) that result in a disparity of reads recovered across cells. The observed chromatin states, denoted by Z_{CM}, are specified by the element-wise product between the latent accessibility Y_{CM} and the capturing status R_{CM}. Since various experimental factors such as sequencing depth are cell-specific, we further assume the capturing probability $\mathrm{P}(R_{cm} = 1)$ to be unique to each cell but common to all peaks in that cell, and thus we use q_c to denote this conditional read capturing probability in cell c.

Motivated by the latent variable model and to account for cell-specific missing data, we extended the cumulative logit model to simultaneously decompose accessibility as:

$$\mathrm{logit}(\mathrm{P}(Y_{cm} \geq t)) = \alpha^{(t)} + \sum_{j=1}^{J} \beta_j F_{cj}\,, \text{ where } \mathrm{P}(Z_{cm} \geq t) = \mathrm{P}(Y_{cm} \geq t) q_c;\ t \in \{1,2,\dots,T\} \quad \text{(Eq. 1)}$$

where q_c is the capturing probability for a cell c, $\mathrm{P}(Y_{cm} \geq t)$ is the sampling probability of cells with accessibility level greater than or equal to t, $\alpha^{(t)}$ is the intercept term in the tth cumulative logit, and β_j is the coefficient for the jth column of the design matrix. Eq. 1 assumes a proportional odds model, where we have a common set of coefficients β_j for all levels of the cumulative distribution, while allowing for a unique constant term $\alpha^{(t)}$ for each level. Hereafter, we refer to our method as the **mcCLR** model, which stands for the missing-corrected cumulative logit regression model.

With the formulation above, the effect of a complex set of predictive variables (and their interactions) on accessibility can be tested by the null hypothesis of $\beta_i = 0$ with a likelihood ratio test (**Fig. 1**, lower panel). One statistical challenge is to estimate q_c's for each cell. We assumed the same capturing probability within a cell regardless of accessibility across different peaks such

that the problem is tractable and can be computed efficiently. Operationally, we first group the cells by their combination of the treatments and then utilize a coordinate descent algorithm to obtain estimates of $\mathrm{P}(Y_{cm} \geq 1|f_c)$ and q_c (**Methods**).

Another statistical challenge of snATAC-seq is that the data is very sparse, creating a so-called "perfect separation" problem (see[30]). Here, we developed a regularized model to resolve the issues with sparsity in snATAC-seq data by generalizing the Firth logistic regression model[25,31], where we incorporate the cell-specific capturing probability (Eq. 1) into the model (**Methods**). Essentially, a Firth penalty is introduced in the regression model:

$$\log L^*(\boldsymbol{\beta}|\boldsymbol{Z}) = \log L(\boldsymbol{\beta}|\boldsymbol{Z}) + \frac{1}{2}\log|I(\boldsymbol{\beta})| \quad \text{(Eq. 2)}$$

Where L^* represents the penalized likelihood, L is the likelihood of the regression model, and $I(\boldsymbol{\beta})$ is the information matrix. Derivations of the parameter estimation framework are described in the **Methods** section. With the proposed methods, we aim to control type I error more accurately and account for technical zeros (due to uneven data capturing) and sparse data. This regression-based model enables the testing of multiple covariates that jointly determine accessibility, while controlling for other covariates or confounders.

<u>Application of PACS to cell type identification</u>

To demonstrate the effectiveness of our model for separating the latent chromatin accessibility from the capturing probability, we evaluated three model assumptions using the task of (supervised) cell type prediction, where the goal is to predict cell types in a new snATAC-seq dataset given an annotated (labeled) dataset.

We first evaluated the accuracy of the estimation procedure of PACS. We simulated groups of cells with a spectrum of both the underlying probability of accessibility ($\mathrm{P}(Y_{cm} \geq 1)$, or p in short) across peaks, and the capturing probabilities (q) across cells (**Methods**). We then utilized PACS to jointly estimate p and q, with n=1000, 500, or 250 cells. The simulation results show that our estimator can determine both the capturing probabilities and open-chromatin probabilities accurately, with root mean squared errors (RMSE) for the underlying probability of accessibility

from 0.028 (n=1000) to 0.027 (n=250) and RMSE for capturing probability from 0.0067 (n=1000) to 0.012 (n=250, **Fig. 2a-d**, **Supplementary Fig. 1a-b**, and **Supplementary Table 1**).

We next tested PACS by applying it to a cell type label transfer task, comparing it with the Naïve Bayes model. For both models, we started with an estimated $\boldsymbol{p}_g$ for each known cell type group label g, and then applied the Bayes discriminative model to infer the most probable cell type labels for novel unidentified cells. Naïve Bayes does not assume missing data; thus, it ignores the cell-specific capturing probability. The prediction performances were evaluated with ten-fold cross-validation and holdout methods, where the original cell type labels are regarded as ground truth (**Methods**). We tested the methods on five datasets, including two human cell line datasets[21], two mouse kidney datasets[6], and one marmoset brain dataset[32]. In the two human cell line datasets, the cell line labels are annotated by their SNP information[21], so the labels are regarded as gold standards. For the remaining datasets, the original cell type labels are generated by clustering and marker-based annotation, so the labels may have errors.

PACS consistently outperforms the Naïve Bayes model with an average 0.31 increase in Adjusted Rand Index (ARI, **Fig. 2e**), suggesting the importance of considering the cell-to-cell variability in capturing rate. For the gold-standard cell line mixture data, we achieved almost perfect label prediction (ARI > 0.99), while Naïve Bayes had much lower accuracy with an average ARI = 0.54 (**Fig. 2f-g**). For the kidney data[6] and the marmoset brain data[32], PACS still achieved high performance, with average ARI equal to 0.92, 0.90, and 0.88 for the adult kidney, P0 kidney, and marmoset brain data, respectively. The Naïve Bayes model, on the other hand, again produced lower ARI scores, equal to 0.59, 0.65, and 0.69 for the three datasets, respectively (**Supplementary Fig. 1e-h**).

For the holdout experiment, where training and testing is done on different datasets, consistent with the above results, our method shows more accurate cell label prediction than Naïve Bayes (**Supplementary Fig. 1i**). We note that our cell type label prediction approach is very efficient, and the total time for training and prediction takes < 5 min for large datasets (>70,000 cells).

<u>PACS enables parametric multi-factor model testing for accessibility</u>

Identifying the set of CREs regulated by certain physiological cues is essential in understanding functional regulation. For example, differentially accessible region (DAR) analysis tries to determine cell type-specific chromosomal accessibility differences. Most snATAC-seq pipelines adopt RNA-seq differential expression methods to ask whether a peak belongs to a DAR. These approaches generally lack calibration for sparse ATAC data, and the approach of pairwise DAR tests does not allow testing more complex models that might determine peak accessibility (e.g., combination of spatial location, batch effects). With existing methods for DAR detection, commonly adopted approaches are to ignore other factors or stratify by other factors to test the factor of interest, if the predictive variables are nominal (e.g., cell types). However, such tests involve ad hoc partition into levels of the nominal factor and cannot test more complex models including possible metric variables (e.g., developmental time).

To evaluate the performance of the parametric test framework in PACS, we first used simulated data to test a single factor model (cell types as factor). We assessed type I error and statistical power, comparing PACS and four established methods: ArchR[21], Seurat/Signac[23], snapATAC[22], and Fisher's exact test. ArchR conducts Wilcoxon rank-sum test on the subsampled cells from the initial groups, while ensuring parity in the number of sequencing reads across any two samples being tested. Seurat employs a standard logistic regression model[33], but setting the tested factor as the dependent variable and read counts (Tn5 insertions) and total reads as predictive variables. SnapATAC conducts a test on the pseudo-bulk data of two groups and utilizes the edgeR[34] negative binomial test on the pseudo-bulk data with a pre-defined ad hoc variance measure (biological coefficient of variation, bvc = 0.4 for human and 0.1 for mouse data).

To resemble real data, simulated samples were generated by parameterizing the model with the accessibility and capturing probability estimated directly from a human cell line dataset[21] (**Methods**). We randomly sampled varying numbers of cells in each group, ranging from 250 to 1000. **Fig. 3a** shows that Seurat failed to control type I error at the specified significance level. Among the methods demonstrating type I error control, PACS has on average 17%, 19% and 122% greater power than Fisher's exact test, ArchR and snapATAC, respectively (**Fig. 3b**, **Supplementary Table 2**). The reduced power of ArchR is likely due to the subsampling process,

239 and the ad hoc "bvc" choice in snapATAC may result in a miscalibrated test with a low type I error
240 and power. The q-q plots of the five methods are shown in **Supplementary Fig. 2a-e**.
241
242 To evaluate the performance under a multi-factor model, we next simulated another snATAC-seq
243 dataset with a second factor, two spatial locations (S1 and S2) along with the first factor of two
244 cell types (T1 and T2). We evaluated the performance in testing for the main effects or their
245 interactions (**Methods**). For the methods that cannot directly test effects for multiple factors, two
246 strategies were used. The first is called the "naïve test", where the other factor is ignored in testing
247 the factor of interest. The second strategy is called the "stratified test", where we stratified the
248 dataset by the second factor and conducted a pairwise test between the factor of interest on each
249 stratum, followed by *P* value combination test (**Methods**). Across all methods and test strategies,
250 only PACS, snapATAC (naïve and stratified), Fisher-stratified, and ArchR-stratified controlled
251 type I error at the specified level (**Fig. 3c-e**); PACS remained the most powerful test and detected
252 23%, 66%, and 39% more true differential peaks compared with the second most powerful
253 methods (**Supplementary Table 3**). Notably, Seurat could not determine peaks with spatial effects
254 or interaction effects, because the cell type factor is treated as the dependent variable in the
255 regression model. More detailed explanation can be found at the **Supplementary Note 1**.
256
257 We then simulated a time-series dataset with five time points, to evaluate our model performance
258 for ordinal covariates. We assumed two temporal trends of accessibility, linear and quadratic trends.
259 To put this in a biological setting, the quadratic trend may represent the presence of an acute spike
260 response and the linear trend may represent temporally accumulating chronic responses. The
261 PACS framework could detect both linear and quadratic signals, and its power is dependent on the
262 "effect sizes" defined as the log fold change of accessibility between the highest and lowest
263 accessibility (**Fig. 3f-g**).
264
265 We also evaluated the PACS model in real datasets. As the ground truth is unknown, we utilized
266 a sampling-based approach. We used randomly permuted cell type labels to estimate the type I
267 error. To evaluate power, we treated the consensus DAR set from all methods as "true DARs"
268 (after type I error calibration, see **Methods**). For the standard two-group DAR test, our method
269 consistently controlled type I error and achieved high power, across different datasets

(**Supplementary Fig. 2f-k**). We further showed that ignoring confounders such as batch effect could result in substantial false discoveries, up to 47% in the adult mouse kidney data (**Fig. 3h-i**).

PACS identifies kidney cell type-specific regulatory motifs and allows direct batch correction

One important feature of PACS is its ability to handle complex datasets with multiple confounding factors. To test the performance of PACS, we analyzed an adult kidney dataset with strong batch effects[6]. This dataset contains three samples generated independently (in three batches), and the authors identified a strong batch effect. Existing methods for batch correction map the ATAC-seq features to a latent vector space to subtract the batch effects. For example, the original study[6] relies on Harmony[35] to remove the batch effect in latent space for visualization and clustering, but the batch effect is still present in the peak feature sets, which could confound downstream analyses and inferences.

To remove the batch effect at the feature level, we assume that the batch effect will affect (increase or decrease) the accessibility of certain peaks, and these effects are orthogonal to the biological effects. This assumption is necessary for most of the existing batch-effect correction methods (e.g., MNN[36], Seurat[37], and Harmony[35]), as a matter of experimental design. With this assumption, we applied PACS on the adult kidney data, detected significant DAR peaks among batches (P value < 0.05 with or without FDR correction) and removed batch-effect peaks from the feature set. We next implemented Signac to process the original data as well as the batch effect-corrected data, without any other batch correction steps. Dimension reductions with UMAP suggested that the original data contained a strong batch effect, where almost all cell types are separated by batch (**Fig. 4a-b**). After removing the peaks with strong batch effects, the cells are better mixed among batches (**Fig. 4c-d**, **Supplementary Fig. 3a-b**). Note that different cell types are still separated, suggesting the biological differences are (at least partially) maintained. Since UMAP visualization may not fully preserve the actual batch mixing structure, we adopted a batch mixing score from Ref.[38] to quantify the batch effect in the PCA space. The batch mixing score is defined as the average proportion of nearest neighbor cells with different batch identities, where a higher score indicates better mixing between batches, and thus a smaller batch effect (**Methods**). We normalized the mean batch mixing score by dividing it by the expected score under the random

mixing scenario. After batch effect correction with PACS, the normalized mean batch mixing score is 0.358 or 0.417 compared with 0.122 before batch correction.

We next applied our method to identify cell type-specific features while adjusting for batch effect. We focused on the two proximal tubule subtypes, proximal convoluted tubules (PCT) and proximal straight tubules (PST). By fitting our mcCLR model with cell type and batch effect, we identified 19,888 and 62,368 significant peaks for PCT and PST, respectively (FDR-corrected *P* value < 0.05, **Supplementary Tables 4-5**). The original study utilized snapATAC, which reported 23,712 and 36,078 significant peaks for PCT and PST, respectively. With the batch-corrected differential peaks, we then conducted GREAT enrichment analysis[39,40] to identify candidate PCT- and PST-specific genes (**Supplementary Tables 6-7**). We identified *Gc*, *Nox4*, *Slc4a4*, *Bnc2*, *Slc5a12*, and *Ndrg1* genes as top PCT-enriched genes, and *Ghr*, *Gramd1b*, *Etv6*, *Atp11a*, *Gse1*, and *Sik1* as top PST-enriched genes. The associated genomic pile-up figures for the CREs of these genes are shown in **Fig. 4e**, and these findings were supported by a public scRNA-seq dataset[41] (**Fig. 4f**).

<u>PACS dissects complex accessibility-regulating factors in the developing human brain</u>

We applied our method to the human brain dataset[11], which is more challenging due to the complex study design with cells collected from six donors across eight spatial locations. Substantial sequencing depth variations among samples has also been noticed, which further complicated the analysis (**Supplementary Fig. 5a-c**). To study how spatial locations affect chromatin structure, the original reference focused on the prefrontal cortex (PFC) and primary visual cortex (V1) regions, as they were the extremes of the rostral-caudal axis[11]. With the multi-factor analysis capacity of PACS, we conducted analyses to (1) identify the region effect, while adjusting for the donor effect, (2) identify the cell-type specific region effect.

We first examined the marginal effect of brain regions on chromatin accessibility, holding other factors constant (**Methods**). For this, we focused on a subset of three donors where spatial information is retained during data collection (**Fig. 5a-c, Supplementary Table 8**). In total, we identified 146,676 brain region-specific peaks (FDR corrected *P* value < 0.05). Between PFC and V1 regions, we identified 30,455 DAR peaks, ~20% more compared with the original study

(**Supplementary Tables 9-10)**. With the region-specific DARs, we conducted motif enrichment analysis to identify region-specific TFs. For the PFC and V1 regions, we found several signals that were consistent with the original article[11], including PFC-specific motifs *MEIS1*, *TBX21*, and *TBR1*, and V1-specific motifs *MEF2B*, *MEF2C*, *MEF2A*, and *MEF2D*. Moreover, we identified additional V1-specific motifs *ETS* and *ZIC2* (**Fig. 5d**), supported by the scRNA-seq data collected from the same regions[42]. We also noticed that some neuron development-associated TFs, including *OLIG2* and *NEUROG2*, are enriched in both brain regions but with different binding sites, likely due to different co-factors that open different DNA regions. Motif enrichment results for both brain regions are reported in **Supplementary Tables 11-12**.

Next, we used PACS to examine the location effect across different cell types along excitatory neurogenesis. This corresponds to testing the interaction terms between spatial location and cell types, while adjusting for donor effect (**Fig. 5e**). The previous study reported that the chromatin status of the intermediate progenitor cells (IPC) population started to diverge between PFC and V1 regions. Consistent with the article, we identified 2773 significant differential peaks between PFC and V1 at IPC stage, 52% more than snapATAC (**Supplementary Table 13**).

In sum, we show the implementation of PACS for data with three levels of factors: donor, spatial region, and cell type. PACS can be applied to study one factor or the interaction between factors while adjusting for other confounding factors, and test results have higher power.

<u>PACS identifies time-dependent immune responses after stimulation</u>

The existing methods for DAR detection rely on pairwise comparisons, and thus are not applicable to ordinal or continuous factors. One such example is the snATAC-seq data collected at multiple time points. Here, we apply PACS to a peripheral blood mononuclear cell (PBMC) dataset collected at three time points (0h control, 1h, and 6h) after drug treatment[43]. Multiple treatments have been applied separately to cells collected from four human donors. While PACS can simultaneously model all drugs and conditions, we focus on the ionomycin plus phorbol myristate acetate (PMA) treatment to demonstrate the PACS workflow. The factors included in the PACS model are shown in **Fig. 6a**, where cell type and donor effects are categorical, and the time effect is coded as an ordinal variable. Note that time can be alternatively coded as a continuous variable.

We tested the treatment effect by identifying open chromatin regions that show a gradual increase or decrease in accessibility after treatment. In total, we detected 35,356 peaks with a strong treatment effect across five broad cell types (B cell, CD4 T cell, CD8 T cell, Monocyte, and NK cell, **Supplementary Tables 14-16**). Across the cell types, CD4 and CD8 T cells show the most significant changes in chromatin landscape after treatment (**Fig. 6b-c**). This is expected, as PMA can induce T cell activation and proliferation[44]. Among the peaks with significant PMA treatment effect, most become more accessible after treatment, consistent with the activation function of the treatment. We then conducted gene enrichment analysis with GREAT[40], where we identified several GO pathways associated with T cell activation, such as "regulation of T cell differentiation" and "regulation of interleukin-2 production" (**Supplementary Table 17**). We also identified enriched genes including *DUSP5*, *IL1RL1*, *TBX21*, and *CXCR3* (**Supplementary Table 18**), expression of which have been previously reported to be up-regulated in PMA treatment[45,46,47,48]. Notably, *DUSP5* is known to play an essential role in the immune response through regulation of NF-κB as well as ERK1/2 signal transduction[49], and *TBX21* is an immune cell TF that also directs T-cell homing to pro-inflammatory sites via regulation of *CXCR3* expression[50]. **Fig. 6d-e** showed the cell type-specific open chromatin landscape dynamic after the PMA treatment. We noticed that some CREs respond to the treatment effect across all cell types and some CREs become activated in only certain cell types.

Discussion:

Single-cell sequencing data, characterized by uneven data capturing and data sparsity, present significant challenges in data analysis. For scRNA-seq data, data normalization has been an essential step for adjusting for uneven data capturing; however, this approach is not directly applicable to scATAC-seq data, creating a unique challenge. Our method, PACS, addresses this by jointly modeling the group-level underlying accessibility and cell-level sequencing reads capturing. When applied to cell type annotation tasks, PACS showed improved performance compared with the Naïve Bayes model, which does not consider cell-specific capturing probabilities.

The increasing volume of data across various tissue conditions necessitates atlas-level data integration to comprehend tissue dynamics. Our cell type annotation framework facilitates the transfer of cell type annotation from reference dataset to another dataset, thereby overcoming a major hurdle in integrative data analysis. A further challenge of data integration is to jointly model various factors (e.g., genotype, cell type, spatial locations) that govern cellular CRE activities. The standard GLM framework could not address the uneven data capturing in snATAC-seq data, so we developed a statistical model that extends the GLM framework to account for cell-specific missing data. Our model, the missing-corrected cumulative logistic regression (mcCLR) with regularization, enables PACS to perform multi-covariate hypothesis tests including spatial and temporal data analysis. Here we analyzed three empirical datasets from brain, kidney, and blood samples to show the utility and flexibility of our framework in large, complex datasets. Additionally, PACS shows promise in genotype-based chromatin accessibility studies, such as allele imbalance analysis or chromatin accessible quantitative trait locus (caQTL) study. Compared with existing models, PACS offers a more controlled approach for handling multiple covariates in the caQTL analysis.

We have previously derived a parametric model of the snATAC-seq read count, known as the size-filtered signed Poisson distribution (ssPoisson)[24]. In our current work, we treat the insertion rate as a latent variable and directly model the paired insertion counts (PIC) of the data with a regression-based model. This approach greatly enhanced computational efficiency. Future research will delve into the potential applications of parametric distributions. In summary, PACS allows versatile hypothesis testing for the analysis of snATAC-seq data, and its capability of jointly accounting for multiple factors that govern the chromosomal landscape will help investigators dissect multi-factorial chromatin regulation.

418 **References:**

419

420 1. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory
421 variation. *Nature* **523**, 486–490 (2015).
422 2. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution.
423 *Nat. Commun.* **9**, 3647 (2018).
424 3. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.
425 *Cell* **174**, 1309-1324.e18 (2018).
426 4. Arda, H. E. *et al.* A Chromatin Basis for Cell Lineage and Disease Risk in the Human
427 Pancreas. *Cell Syst.* **7**, 310-322.e4 (2018).
428 5. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**,
429 5985-6001.e19 (2021).
430 6. Miao, Zhen *et al.* Single cell regulatory landscape of the mouse kidney highlights cellular
431 differentiation programs and disease targets. *Nat. Commun.*
432 doi:https://doi.org/10.1038/s41467-021-222266-1.
433 7. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell
434 resolution. *Nature* **555**, 538–542 (2018).
435 8. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune
436 cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
437 9. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human
438 hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
439 10. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory
440 epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).

11. Ziffra, R. S. *et al.* Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021).
12. Deng, Y. *et al.* Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**, 375–383 (2022).
13. Turner, A. W. *et al.* Single-nucleus chromatin accessibility profiling highlights regulatory mechanisms of coronary artery disease risk. *Nat. Genet.* **54**, 804–816 (2022).
14. Benaglio, P. *et al.* Mapping genetic effects on cell type-specific chromatin accessibility and annotating complex immune trait variants using single nucleus ATAC-seq in peripheral blood. *PLOS Genet.* **19**, e1010759 (2023).
15. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* **83**, 373–392 (2023).
16. Miao, Z., Humphreys, B. D., McMahon, A. P. & Kim, J. Multi-omics integration in the age of million single-cell data. *Nat. Rev. Nephrol.* **17**, 710–724 (2021).
17. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
18. Sullivan, K. M. & Susztak, K. Unravelling the complex genetics of common kidney diseases: from variants to mechanisms. *Nat. Rev. Nephrol.* **16**, 628–640 (2020).
19. Sheng, X. *et al.* Mapping the genetic architecture of human traits to cell types in the kidney identifies mechanisms of disease and potential treatments. *Nat. Genet.* **53**, 1322–1333 (2021).
20. Yu, F. *et al.* Variant to function mapping at single-cell resolution through network propagation. *Nat. Biotechnol.* **40**, 1644–1653 (2022).

21. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
22. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).
23. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
24. Miao, Z. & Kim, J. Is single nucleus ATAC-seq accessibility a qualitative or quantitative measurement? *bioRxiv* 2022.04.20.488960 (2022) doi:10.1101/2022.04.20.488960.
25. FIRTH, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).
26. Heinze, G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat. Med.* **25**, 4216–4226 (2006).
27. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
28. Li, Y. E. *et al.* An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**, 129–136 (2021).
29. Martens, L. D., Fischer, D. S., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell ATAC-seq analysis. 2022.05.04.490536 Preprint at https://doi.org/10.1101/2022.05.04.490536 (2022).
30. Agresti, A. *Categorical data analysis*. vol. 792 (John Wiley & Sons, 2012).
31. Heinze, G. & Schemper, M. A solution to the problem of separation in logistic regression. *Stat. Med.* **21**, 2409–2419 (2002).
32. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).

33. Ntranos, V., Yi, L., Melsted, P. & Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **16**, 163–166 (2019).

34. Chen, Y., Lun, A. T. L. & Smyth, G. K. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. Preprint at https://doi.org/10.12688/f1000research.8987.2 (2016).

35. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

36. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).

37. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

38. Chari, T. & Pachter, L. *The Specious Art of Single-Cell Genomics*. http://biorxiv.org/lookup/doi/10.1101/2021.08.25.457696 (2021) doi:10.1101/2021.08.25.457696.

39. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

40. Tanigawa, Y., Dyer, E. S. & Bejerano, G. WhichTF is functionally important in your open chromatin data? *PLOS Comput. Biol.* **18**, e1010378 (2022).

41. Ransick, A. *et al.* Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney. *Dev. Cell* **51**, 399-413.e7 (2019).

42. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).

509 43. Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics.
510 *Cell Genomics* **2**, 100166 (2022).
511 44. Ai, W., Li, H., Song, N., Li, L. & Chen, H. Optimal Method to Stimulate Cytokine
512 Production and Its Use in Immunotoxicity Assessment. *Int. J. Environ. Res. Public. Health*
513 **10**, 3834–3842 (2013).
514 45. Brignall, R. *et al.* Integration of Kinase and Calcium Signaling at the Level of Chromatin
515 Underlies Inducible Gene Activation in T Cells. *J. Immunol.* **199**, 2652–2667 (2017).
516 46. Shin, H.-J., Lee, J.-B., Park, S.-H., Chang, J. & Lee, C.-W. T-bet expression is regulated by
517 EGR1-mediated signaling in activated T cells. *Clin. Immunol.* **131**, 385–394 (2009).
518 47. Dagan-Berger, M. *et al.* Role of CXCR3 carboxyl terminus and third intracellular loop in
519 receptor-mediated migration, adhesion and internalization in response to CXCL11. *Blood*
520 **107**, 3821–3831 (2006).
521 48. Cooke, M. *et al.* Differential Regulation of Gene Expression in Lung Cancer Cells by
522 Diacyglycerol-Lactones and a Phorbol Ester Via Selective Activation of Protein Kinase C
523 Isozymes. *Sci. Rep.* **9**, 6041 (2019).
524 49. Seo, H. *et al.* Dual-specificity phosphatase 5 acts as an anti-inflammatory regulator by
525 inhibiting the ERK and NF-κB signaling pathways. *Sci. Rep.* **7**, 17348 (2017).
526 50. Stolarczyk, E., Lord, G. M. & Howard, J. K. The immune cell transcription factor T-bet.
527 *Adipocyte* **3**, 58–62 (2014).
528 51. Winship, C. & Mare, R. D. Regression Models with Ordinal Variables. *Am. Sociol. Rev.* **49**,
529 512 (1984).
530 52. Christensen, R. H. B. *Sensometrics: Thurstonian and Statistical Models*. (Technical
531 University of Denmark, 2012).

53. Venzon, D. J. & Moolgavkar, S. H. A Method for Computing Profile-Likelihood-Based Confidence Intervals. *Appl. Stat.* **37**, 87 (1988).

54. Adey, A. C. Tagmentation-based single-cell genomics. *Genome Res.* **31**, 1693–1705 (2021).

55. Duttke, S. H., Chang, M. W., Heinz, S. & Benner, C. Identification and dynamic quantification of regulatory elements using total RNA. *Genome Res.* (2019) doi:10.1101/gr.253492.119.

Methods

Data availability

We downloaded the following snATAC-seq datasets from public repositories:
mouse kidney data[6] (GEO GSE157079,
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157079),
human cell line data[21] (GEO GSE162690,
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162690),
developing human brain data[11] (GEO GSE163018,
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163018),
marmoset brain data[32] (the Brain Cell Data Center RRID SCR_017266; https://biccn.org/data),
human PBMC time-series stimulation data[43] (GEO GSE178431,
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178431).

Probabilistic model of underlying open chromatin status

Here we model the activity of regulatory elements in each cell type group by the cumulative distribution of the accessibility. The underlying accessibility for a CRE is a function of nucleosome density and turnover rate. As we discuss in the main text, for a particular cell group, the chromatin state should be regarded as a random variable as they are sampled from mixtures of hidden microstates. Here, we expanded the model of accessible chromatin from Ref[24]. Briefly, let $F_{C\times J}$ be a design matrix that summarizes known independent variables (e.g., cell type, developmental time, sample locations, etc.) across C cells, $Y_{C\times M}$ be the underlying (latent) chromatin status across C cells and M regions, where each element represent the accessibility of a genomic region. The goal of PACS is to decompose the (complementary) cumulative distribution of Y_{cm}, i.e., the series of distributions:

$$\mathrm{P}\{Y_{cm} \geq t\} = \sum_{i=t}^{T} \pi_i \text{ for } t = 1,2,\dots,T \qquad \text{(Eq. 3)}$$

by predictive independent variables in F_{c*}. Here the maximum value of accessibility we account for, T, is feature specific. To be precise, for a feature m, T is the largest integer such that $\sum_c 1(Z_{cm} \geq t) \geq n_c$ where n_c is a hyperparameter. In our study, n_c is set to be $0.25C$ but based on our evaluation, our model is not sensitive to the choice of n_c.

Model for capturing probability of cell

Due to various experimental factors like enzyme activity and sequencing depth disparities across cells, we introduce $R_{C\times M}$ as a matrix representing the capturing status of each cell and region. Let $Z_{C\times M}$ be the (observed) scATAC dataset, we have $Z = Y \otimes R$, where $\otimes$ denote element-wise Product. We consider R_{cm} to be sampled from a Bernoulli distribution parameterized by q_c, cell-specific capturing probability:

$$R_{cm} \sim \text{Bernoulli}(q_c) \qquad \text{(Eq. 4)}$$

Joint parameter estimation for single-factor scenario

Given a class of data that correspond to a combination of levels of independent variables, we follow the same parameter estimation framework as described in Ref[24]. Briefly, assume we have a genomic region-by-cell (i.e., peak-by-cell) matrix $Z_{C_f\times M}$ with C_f denoting the subset of cells corresponding to some combination of the independent prediction factors. The observed values in $Z_{C_f\times M}$ are ordinal values, but as most of the non-zero scATAC-seq counts are one (typically >70%), we focus on $\mathrm{P}(Y_{C_f m} \geq 1)$ for purposes of q_c estimation. Hereafter, we use the notation $p^{(1)}_{C_f m}$ to represent the (non-zero) open probability of group C_f and feature m. We have further assumed q_c to be identical across different levels of accessibility for a given cell. Due to the data sparsity and the predominant counts of one, this assumption is moderate, and the estimation process will be greatly accelerated with this assumption. We use moment estimator with a coordinate descent algorithm to iteratively update $p^{(1)}_{C_f m}$ given q_c, and update q_c given $p^{(1)}_{C_f m}$. Briefly, we execute the following iteration until convergence:

1. Start with an initial estimate of $p_m^{[0]}$
2. For $t = 1, 2, \ldots$
    a. Compute $q_c^{[t]}$ by:

    $$q_c^{[t]} = \frac{\sum_{m=1}^{M} I(z_{cm} \geq 1)}{\sum_{m=1}^{M} p_m^{[t-1]}} \text{ for } c \in C_f$$

    b. Update $p_m^{[t]}$ by moment estimator:

$$p_m^{[t]} = \frac{\sum_{c \in C_f} I(z_{cm} \geq 1)}{\sum_{c \in C_f} q_c^{[t]}} \text{ for } m \in \{1,2,\dots,M\}$$

where we use superscript $[t]$ to represent the tth iteration, and we omit the subscript C_f and superscript (1) for $p_{C_f m}^{(1)}$.

Uniqueness of parameter estimation

In order for the above joint parameter estimation framework to converge and for the estimated parameters to be uniquely defined, there should be $q_c = 1$ for some cells and $p_{C_f m}^{(1)} = 1$ for some features. In PACS, we conduct a convergence check by requiring a certain proportion of cells (default 10%) to have an estimated capturing probability greater than 0.9. In the case of a cluster of cells being rare or not sufficiently deeply sequenced, the estimates may be unstable, and we recalibrate the estimates for this rare cluster to its most similar cluster to prevent potential false positives. Specifically, let C_{f1} index the rare group of cells; then, to identify the cell groups with the most similar open chromatin profile, we compute the correlation between $p_{C_{f1}*}^{(1)}$ and $p_{C_{fj}*}^{(1)}$ for all other clusters $j = 1, \dots, J$, across all regions. Assuming C_{fn} has the most similar chromatin profile, we rescale the current estimation of $p_{C_{f1}m}^{(1)}$ by the following formula:

$$S = \sum_m p_{C_{fn}m}^{(1)} / \sum_m p_{C_{f1}m}^{(1)}$$

$$p_{C_{f1}*}^{(1)\prime} = p_{C_{f1}*}^{(1)} \times S \qquad \text{(Eq. 5)}$$

where S is the scale factor, $p_{C_{f1}m}^{(1)\prime}$ is the rescaled open probability estimate for the cluster C_{f1} and feature m, and through rescaling, we are essentially assuming that most peaks are not differentially accessible between these two cell types.

Cell type label prediction framework

Given a reference dataset, we estimate the probability of open chromatin $p_{C_g m}^{(1)}$ for each cell type $g \in \{1, \dots, G\}$, using the formula above. With a new set of observations $Z'_{C' \times M}$, we apply the Bayes discriminative model to predict the corresponding cell type labels, $h(Z'_{c*})$.

$$\mathrm{P}(h(Z'_{c*}) = g | Z'_{c*}) = \mathrm{P}(Z'_{c*} | h(Z'_{c*}) = g)\mathrm{P}(h(Z'_{c*}) = g)$$
$$= \mathrm{P}(h(Z'_{c*}) = g) \prod_{m=1}^{M} \left(p_{C_g m}^{(1)} q_c \right)^{Z'_{cm}} \left(1 - p_{C_g m}^{(1)} q_c \right)^{1 - Z'_{cm}} \quad \text{(Eq. 6)}$$

where $\mathrm{P}(h(Z'_{c*}) = g | Z'_{c*})$ represents the posterior probability of cell c being sampled from cell group g, $\mathrm{P}(Z'_{c*} | h(Z'_{c*}) = g)$ represents the conditional probability of observing Z'_{c*} given that the cell c is sampled from cell type g, $\mathrm{P}(h(Z'_{c*}) = g)$ is the prior probability of a new observation belonging to cell group g, which can either be assumed to be a non-informative Dirichlet prior $\mathrm{Dirich}(\delta)$ or estimated based on the cell type composition in reference data. Note that we have a large feature space so this choice will not make a big difference.

Missing-corrected cumulative logistic regression (mcCLR)

Due to high sparsity of scATAC-seq data, perfect separability is common, hindering the parameter estimation in (Eq. 1). To address this issue, we incorporated Firth regularization (Eq. 2). Here we summarize the (unregularized) log likelihood function and information matrix for the cumulative response model and derive the analytical expression for the binary model. The loss function when considering cumulative response is

$$\log L(\boldsymbol{\Pi}, \boldsymbol{y} | \boldsymbol{q}) = \sum_{c=1}^{C} \sum_{t=0}^{T} \log(\tilde{\pi}_{ct}) \mathrm{I}(z_c = t) \quad \text{(Eq. 7)}$$

where C represent the total number of cells, π_{ct} and $\tilde{\pi}_{ct}$ represent the probability of t PIC counts in cell c before and after accounting for cell-specific capturing probability, respectively. Specifically, $\pi_{ct} = P(y_c \geq t) - P(y_c \geq t + 1)$, $\Pi_c = (\pi_{c0}, \pi_{c1}, \pi_{c2}, \dots, \pi_{cT})^{\mathrm{Trans}}$ and $\widetilde{\Pi}_c = Q_c \Pi_c$, where Q_c is the capturing probability matrix of dimension $(T + 1) \times (T + 1)$ specified as

$$Q_c = \begin{bmatrix} 1 & 1-q_c & 1-q_c & \dots & 1-q_c \\ 0 & q_c & 0 & & \\ 0 & 0 & q_c & & \vdots \\ \vdots & & & \ddots & \\ 0 & & \dots & & q_c \end{bmatrix} \quad \text{(Eq. 8)}$$

In our PACS model, an approximated estimation of parameters in the cumulative logit model were obtained using a method described in a previous set of studies[51,52] that based on stacking the data and optimize with binary logistic regression specified by

$$\log L(\boldsymbol{p}, \boldsymbol{z}|\boldsymbol{q}) = \sum_{c=1}^{C}[z_c \log(p_c q_c) + (1-z_c)\log(1-p_c q_c)] \quad \text{(Eq. 9)}$$

$$I(\boldsymbol{\beta}) = F^T W F \text{ where } W = \text{diag}\{\tfrac{p_c q_c\,(1-p_c)\text{^}2}{1-p_c q_c}\} \quad \text{(Eq. 10)}$$

where $p_c = P(z_c = 1)$.

<u>*Parameter estimation for mcCLR*</u>

We implemented both Newton's method and the Iterative Reweighted Least Squares method (IRLS) for parameter estimation. Briefly, for Newton's method, $\boldsymbol{\beta}$ is estimated through the following iteration

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} + I'^{-1}\left(\boldsymbol{\beta}^{(s)}\right) U^*(\boldsymbol{\beta}^{(s)}) \quad \text{(Eq. 11)}$$

where the superscript s represents the iteration, $I' = I$ for the full model and $I' = I_{-\{d\}}$ for the null model of $\beta_{\{d\}} = 0$. The score function $U^*(\beta)$ is given by:

$$U^*(\beta_r) = U(\beta_r) + \frac{1}{2}\text{trace}\left[I(\boldsymbol{\beta})^{-1}\ \frac{\partial I(\boldsymbol{\beta})}{\partial \beta_r}\right]$$

$$= \sum_{c=1}^{C} \frac{x_{cr}(y_c - p_c q_c)(1-p_c)}{1-p_c q_c} + \frac{1}{2}\sum_{c=1}^{C} f_{cr} k_r h_r, \quad (r = 1, \dots, p) \quad \text{(Eq. 12)}$$

where the h_c's are the cth diagonal elements of the "hat" matrix, $H = W^{1/2}F(F^T W F)^{-1}F^T W^{1/2}$, and $k_c = (2p_c^2 q_c - 3p_c + 1)\ /\ (1 - p_c q_c)$.

For the IRLS method, the information matrix I is replaced with an estimate of the information matrix, $\tilde{I}$,

$$\tilde{I}(\boldsymbol{\beta}) = F^T \widetilde{W} F, \text{ where } \widetilde{W} = \text{diag}\{-\frac{[-p_c^2 q_c^2 + q_c(2p_c + z_i - 1) - z_i] p_c(1-p_c)}{(1-p_c q_c)^2}\} \quad \text{(Eq. 13)}$$

Hypothesis testing framework of mcCLR

We utilized a generalized likelihood ratio test framework for hypothesis testing with the mcCLR model, although a Wald-type test can also be derived. As the model contains Firth regularization, we used the profile penalized likelihood approach to obtain P values[31,53]. Specifically, in the null model, the coefficients of interest are set to zero but still left in the model, so that the regularization accounts for the presence of these parameters during optimization.

Data simulation for single factor differential test

To mimic real data, we estimated insertion rates ($\lambda_{C_f m}$) and q_c from the human cell line data and use these values to construct simulated data. Briefly, because viable snATAC-seq reads come from two adjacent Tn5 insertion events that have the right primer configuration (reviewed in[54]), we derived the size-filtered signed Poisson (ssPoisson) distribution to model this data generation process[24]. With the observed counts, we estimated the insertion rate parameters for two cell types, and regions with true insertion rate difference greater than 0.1 were set to be as true differential (H_a) and the remaining region's open probabilities were set equal (by taking the mean) and therefore non-differential (H_0). Based on parametric model of latent and observed accessibility, we first sampled the latent ATAC reads by $\text{ssPoission}(\hat{\lambda}_{C_f m})$ for $f = 1,2$, and then sampled the observing status by Bernoulli distribution parameterized by q_c. The observed data were generated by the element-wise product of these two matrices. We randomly sampled 10,000 non-differential features to assess the type I error and 10,000 differential features to evaluate power. This simulation was conducted under varying numbers of cells in each group (from 250 to 1000), and each scenario was repeated 5 times.

Data simulation for multi-factor differential test

Building upon the single factor setting, we further assumed the data to contain two cell types (T1 and T2) being sampled from two spatial locations (S1 and S2). We evaluated marginal effects and their interactions through separate simulations. To simulate data with marginal effect, the cell type effect is first introduced using the same method as in single-factor simulation, so the effect sizes varies. The spatial effect was then considered to affect features both with and without a cell type effect. Specifically, a third of the features with (and without) a cell type effect showed an accessibility difference across batches, with a log fold change of ± 0.5. We introduced sample imbalance as frequently seen in real datasets. To be precise, we considered that S1 contained 1600 T1 cells and 800 T2 cells, while S2 contained 400 T1 cells and 1200 T2 cells. The peak by cell count data generation procedure is the same as for the single factor setting. To evaluate the performance of methods that do not support multi-factor testing, two strategies were used, naïve test and stratified test as reported in the main text. Following stratified test, we use Edington's p value combination approach as it assumes consistent effect size across strata.

To evaluate interaction effect, we considered two configurations of interaction. In the first configuration, cells of T1 in S1 have high accessibility while other groups have low accessibility, with effect sizes also estimated from the cell line data. In the second configuration, cells of T1 in S1 and cells of T2 in S2 have high accessibility while other groups have low accessibility. The second configuration may not be common in real biological data, but a method with capability of testing interaction effect should be able to identify it.

Data simulation for time-series differential test

To evaluate model performance in situations where the design matrix contains ordinal covariates, we simulated time-series snATAC-seq data across five time points. We assumed linear and quadratic temporal effects on accessibility and set the effect size (log fold change) to be 0.3 or 0.5 between the two groups. The baseline accessibility was generated from the cell line data and the peak by cell count data generation procedure is the same as for the single factor setting. N=1000 cells were sampled for each time point.

Evaluating type I error and power in real datasets

To estimate type I error in real data where the ground truth is unknown, we used a label permutation approach, where the data in one cell type were divided randomly into two groups, and a differential test was conducted between these groups. As this is randomly assigned, all features were believed to be non-DAR, so the proportion of *P* values smaller than 0.05 is the empirical type I error using real data. Then, we set the fifth rank percentile as the correct critical value for those methods with type I errors greater than 0.05. We next conducted a test with two different cell types using the calibrated critical values for each method. Since we do not know the true DAR set, we defined the pseudo-true DAR peaks as the union DAR set of all tested methods, using their corresponding new critical values. Power for each method was then calculated by the number of DARs detected divided by the number of pseudo-true DARs. This approach is adopted from Ref.[24].

Estimating effect size (fold change and accessibility change)

A common practice to determine differential features in single cell data is by setting a cutoff for both *P* value and fold change. In scRNA-seq data analysis, one way to estimate the effect size of a particular variable (predictor) is by calculating the fold change (FC) for the normalized data, obtained by dividing the normalized mean expression of one group by the other group. However, with snATAC-seq data, there is no direct normalization method available, and computing the fold change on raw read counts may lead to inaccuracies due to disparities in data capture. Here, we propose to use the capturing probability-adjusted count to compute fold change (FC) or the arithmetic difference between accessibility (accessibility change, AC) of two cell types. To be precise:

$$\mathrm{FC} = \frac{\sum_{c \in C_1} Z_{cm}/q_c}{\sum_{c \in C_2} Z_{cm}/q_c}, \quad \mathrm{AC} = \sum_{c \in C_1} Z_{cm}/q_c - \sum_{c \in C_2} Z_{cm}/q_c \qquad \text{(Eq. 14)}$$

where m is the feature of interest and C_1 and C_2 are the lists of cells that contain foreground and background cell types.

Processing kidney adult data with Signac

We used Signac[23] to evaluate the effectiveness of our method in correcting for batch effect at the feature level. We follow the standard workflow as recommended in the Signac vignette (https://stuartlab.org/signac/articles/pbmc_vignette.html). Briefly, we used the TF-IDF approach

without feature selection (*min.cutoff = 'q0'*), followed by SVD to reduce dimensionality. We then conduct clustering and UMAP visualization using the dimensions 2 to 30 (as the first LSI dimension usually reflects sequencing depth, per the Seurat tutorial). The sample and cell type labels are retrieved from the annotations in the initial publication.

Batch mixing score calculation

We calculated the batch mixing scores in the PCA space as a measure of batch effect. At the cell level, the batch mixing score is adopted from Ref.[38] and is defined as the proportion of nearest neighbor cells with different batch identities, where a higher score indicates better mixing between batches, and thus a smaller batch effect. At the whole data level, the batch mixing score is defined as the mean batch mixing score across all cells. To calculate the expected batch mixing score for a given dataset when no batch effect is present, let M denote a cell type-by-batch matrix, with each element m_{ij} representing the number of cells in the cell type i and batch j. Then the expected data-level batch mixing score in the setting of no batch effect is given by

$$\mathrm{E}[\text{batch mixing score}] = \frac{1}{\sum_{i,j} m_{ij}}\{\sum_i[\sum_j m_{ij}(\frac{\sum_{k\neq j} m_{ik}}{\sum_k m_{ik}})]\} \quad \text{(Eq. 15)}$$

The normalized batch mixing score is the batch mixing score divided by the expected score under random mixing, and thus a higher normalized batch mixing score indicates better mixing across samples.

Processing developing human brain data

This dataset contains 18 specimens collected from human donors. For our study, we excluded samples with unknown spatial locations (GW17, GW18, GW21) or samples not from the cortex (MGE_GW20 and MGE_twin34). Here we focused on the excitatory neuron lineage, including radial glia (RG), intermediate progenitor cells (IPCs), early excitatory neurons (earlyEN), deep layer excitatory neurons (dlENs), and upper layer excitatory neurons (ulENs). We further excluded the insular region for having too few cell counts (645 cells across five cell types). The data matrix was saved as a binary matrix, so we implemented the missing-corrected logistic regression model for the analyses of this data.

DAR identification in the developing human brain data

We constructed two models to identify the significant region effect of the excitatory neuron lineage. Specifically, to identify the region effect, the systematic component of the PACS model is specified as:

$$\alpha + \sum_{k=2}^{K} \gamma_k 1(G_c = g_k) + \sum_{l=2}^{L} \zeta\, 1(S_c = S_l) + \sum_{t=2}^{M} \tau_m 1(D_c = D_m) \quad \text{(Eq. 16)}$$

where G is the index of cell type, S is the index of spatial location, and D is the index of the donor. The null hypothesis for the test is $H_0: \zeta = 0$. To identify the cell type specific region effect, we additionally included the interaction terms between each cell type and spatial location, and the test was conducted for each interaction term.

Motif enrichment analysis

The motif enrichment analysis was conducted with Homer[55]. The list of significant DAR peaks is used as input for the analysis, with the size of the search region specified as 300 bp around the peak center. The reported motif enrichment scores are FDR-corrected P values from the known motif results.

DAR identification in the human PBMC treatment data

To identify the cell type-specific temporal effect in the PBMC treatment data, the systematic component of the PCAS model is specified as:

$$\alpha + \sum_{k=2}^{K} \gamma_k 1(G_c = g_k) + \kappa E + \sum_{t=2}^{M} \omega_t 1(D_c = D_t) \quad \text{(Eq. 17)}$$

where G is the index of cell type, E is the experimental time index (0, 1, 2 corresponds to control, 1h, and 6h after treatment, respectively), and D is the donor index. The null hypothesis for the test is $H_0: \kappa = 0$.

Gene and pathway enrichment with GREAT

We used the GREAT method (v. 4.0.4) to conduct gene and enrichment analysis[39], with DARs as input and default parameter settings. The output from GREAT for the human PBMC data can be found in the **Supplementary Tables 17-18**.

Supplementary Materials:

Supplementary Figures 1-5

Supplementary Table 1: Parameter estimation using simulated data

Supplementary Table 2: Type 1 error and power of different methods using simulated data (one-factor setting)

Supplementary Table 3. Type 1 error and power of different methods using simulated data (two-factor setting)

Supplementary Table 4. PCT specific peaks in the adult kidney data

Supplementary Table 5. PST specific peaks in the adult kidney data

Supplementary Table 6. GREAT gene enrichment results of PCT specific peaks

Supplementary Table 7. GREAT gene enrichment results of PST specific peaks

Supplementary Table 8. Number of cells in across spatial regions and donors

Supplementary Table 9. V1 specific peaks in the developing human brain data

Supplementary Table 10. PFC specific peaks in the developing human brain data

Supplementary Table 11. Homer motif enrichment results of the V1 region in the human developing brain data

Supplementary Table 12. Homer motif enrichment results of the PFC region in the human developing brain data

Supplementary Table 13. Number of differential peaks between PFC and V1 across excitatory neuron lineage in the developing human brain data

Supplementary Table 14. Significant up-regulated peaks after treatment across cell types in the PBMC treatment data

Supplementary Table 15. Significant down-regulated peaks after treatment across cell types in the PBMC treatment data

Supplementary Table 16. Number of significant differential peaks after treatment across five cell types, using PACS or ArchR

Supplementary Table 17. GREAT pathway enrichment results of up-regulated treatment effect peaks in T cells

Supplementary Table 18. GREAT gene enrichment results of up-regulated treatment effect peaks in T cells

Code Availability

PACS is an open-access software available at the GitHub repository https://github.com/Zhen-Miao/PACS. Codes for reproducing the analyses are also available at the GitHub page.

Author Contribution

JK and ZM conceived the study. ZM, JW, and JK designed the statistical model. JW formulated the missing data model for sequencing depth and derived the analytical expression for missing-corrected logistic regression estimation procedure. ZM implemented the model and constructed the software package with feedback from JW, DK, and JK. ZM conducted the simulation and real data analysis with help from KP and DK. JK supervised the work. JK and ZM wrote the manuscript with feedback from JW.

Acknowledgements

This work has been supported in part by the UC2DK126024 grant to JK and also by the Health Research Formula Fund of the Commonwealth of Pennsylvania who did not play a direct role in the work. We thank Blavatnik Family Fellowship that supported the work of ZM. We thank Dr. Pablo Camara, Dr. Nancy Zhang, Dr. Kui Wang, Dr. Xiangjie Li, Dr. Yinan Lin, Dr. Mengying You and members of Junhyong Kim's lab, especially Erik Nordgren for their constructive suggestions that improved this work. We thank Dr. Kun Zhang and Dr. Jason Buenrostro for sharing the metadata.

Competing interests

The authors declare no competing interest.

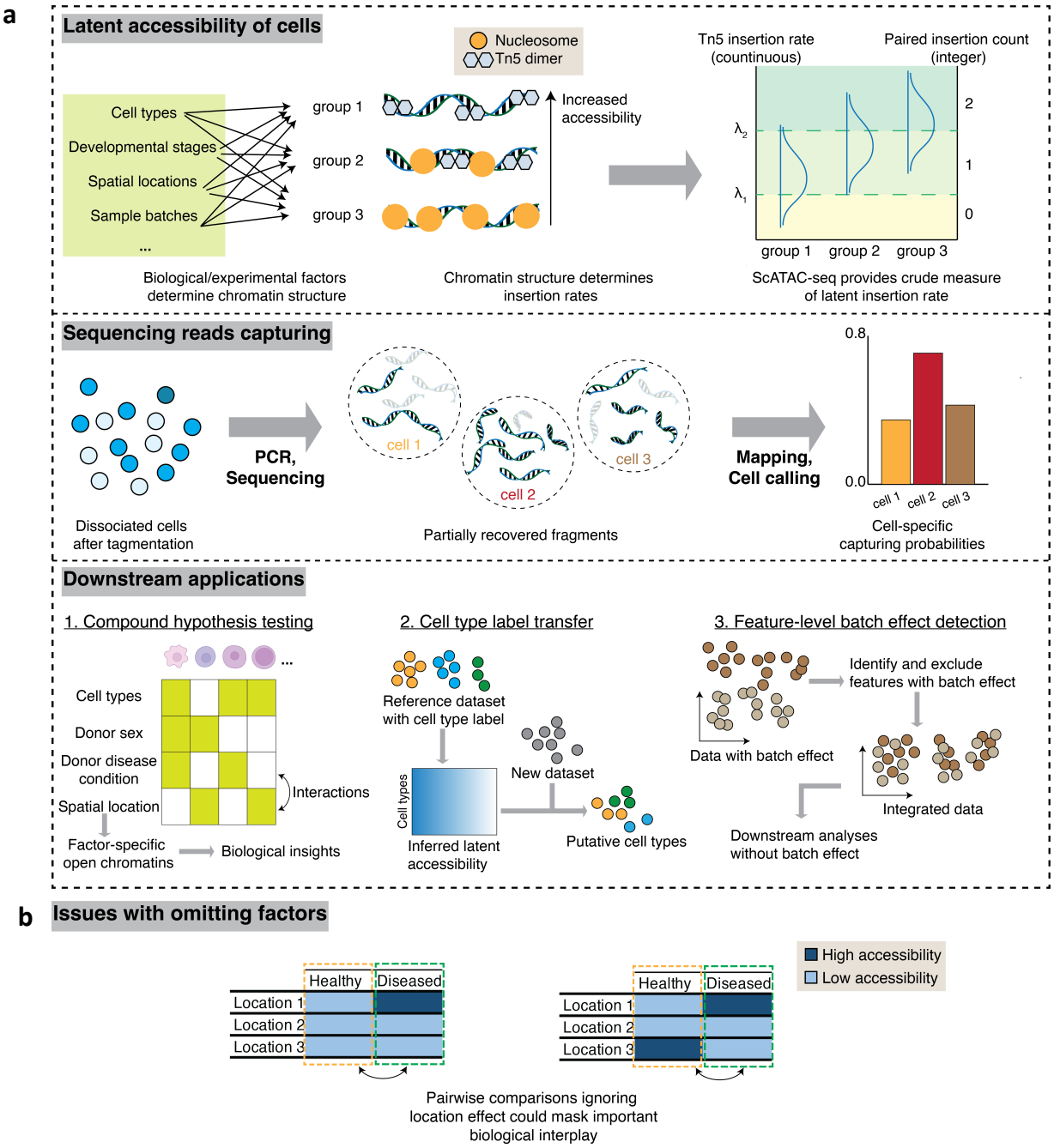


Figure 1. PACS modeling framework.

a. Upper panel: Illustration of the latent accessibility of cells. Multiple factors including cell types, developmental stages, spatial locations etc. determines the chromatin structure and configurations of corresponding cell groups. These different chromatin structures result in the variable Tn5 insertion rates in the ATAC-seq experiments. The readout of ATAC assays are paired insertion counts (PIC), which are crude measures of latent insertion rates. Middle panel: Illustration of the sequencing reads capturing process of snATAC-seq. During PCR and sequencing, fragments in each single cell are partially captured, and after data processing, variable capturing probability should be accounted for in data modeling. Lower panel: Example downstream application of PACS. PACS provide functionalities for compound hypothesis testing, cell type label transfer, and feature-level batch effect correction.

b. Example chromatin configurations where ignoring factors in conducting Differential Accessible Regions (DARs) test may result in false interpretation.

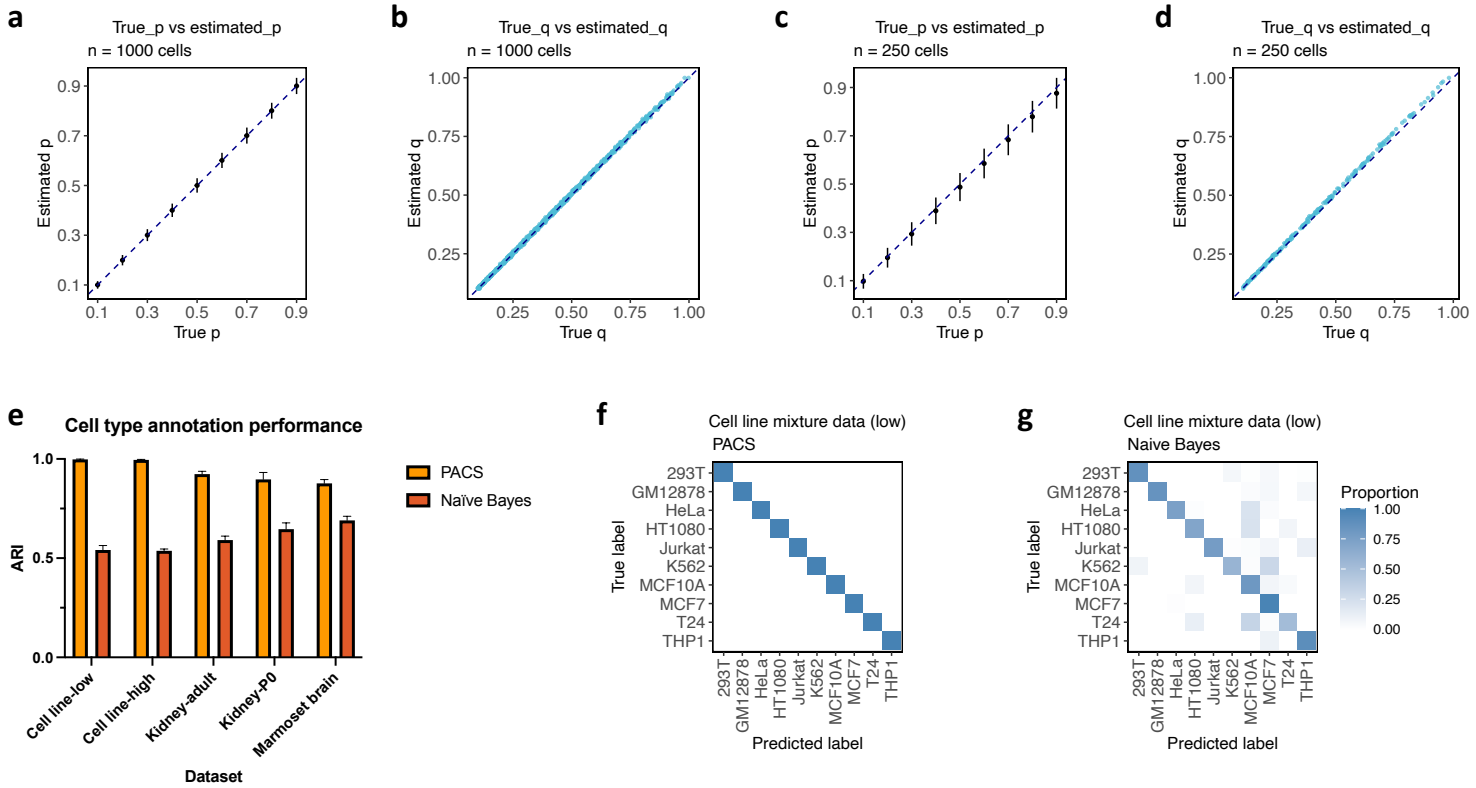


Figure 2. Parameter estimation evaluation and application to cell type annotations.

a-d. Parameter estimation accuracy evaluated using simulation data, for n=1000 cells (**a-b**) or n=250 cells (**c-d**). Here p represents P(y >= 1) and q represents the capturing probability. For this panel and all panels below, the error bars indicate the standard deviation across repeated simulations (n=5). Parameter estimation accuracy with n=500 cells is shown in **Supplementary Fig. 1**.

e. Comparison of cell type annotation adjusted rand index (ARI) between PACS and Naïve Bayes method for five datasets.

f. Confusion matrix between true cell type labels and PACS-inferred cell type labels for the human cell line mixture data (low cell loading setting). The confusion matrices for other datasets are in the Supplementary Figure 1.

g. Confusion matrix between true cell type labels and Naïve Bayes-inferred cell type labels for the human cell line mixture data (low cell loading setting).

a

		PACS	Seurat	ArchR	snapATAC	Contingency table
Method overview	Model type	mcCLR	Logistic regression	Wilcoxon rank sum test	edgeR regression	Fisher's exact test
	Sequencing depth adjustment	Probability model	Total reads as a covariate in regression	Downsampling cells	Normalizing in pseudobulk	✗
	Model-based control of sampling imbalance	✓	✗	✗	✓	✗
	Model-based control of data sparsity	✓	✗	✗	✗	✗
Multi-factor statistical tests	Multi-factor testing (built in)	✓	✓	✗	✗	✗
	Interaction between factors	✓	✗	✗	✗	✗
	Batch effect peak identification	✓	✗	✗	✗	✗
	Time series data analysis	✓	✗	✗	✗	✗

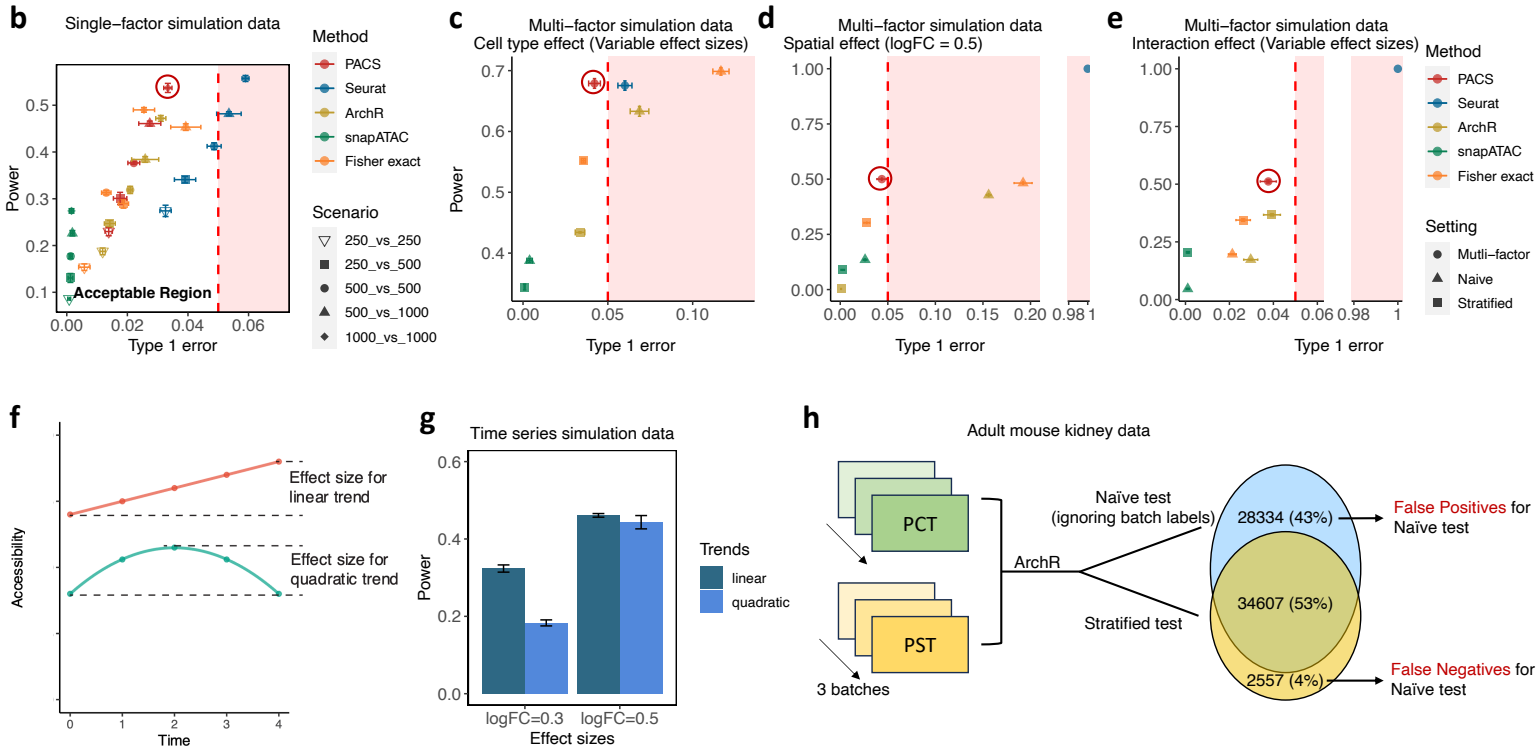


Figure 3. Compound hypothesis testing with PACS is sensitive and specific.

a-b. Type I error and power of different methods evaluated with single-factor simulation data. For this panel and all panels below, all tests are two-sided, and the error bars indicate the standard deviation across repeated simulations (n=5).

c-e. Type I error and power of different methods evaluated with two-factor simulation data. The best-performing method is indicated by a circle with matched color. Cell type and spatial effects are jointly simulated, and **c** and **d** are the test of marginal effects, **e** is the test of interaction effects.

f. Illustration of linear and quadratic effects of treatment on accessibility across five time points. Effect sizes are defined as the log fold change (logFC) between the highest accessibility over the lowest accessibility.

g. Evaluation of power in detecting linear and quadratic temporal effects using simulated data with different effect sizes. N=1000 cells were sampled for each time point.

h. The false positive and false negative rates from ignoring batch covariates in testing cell type effect with ArchR, for the adult mouse kidney data.

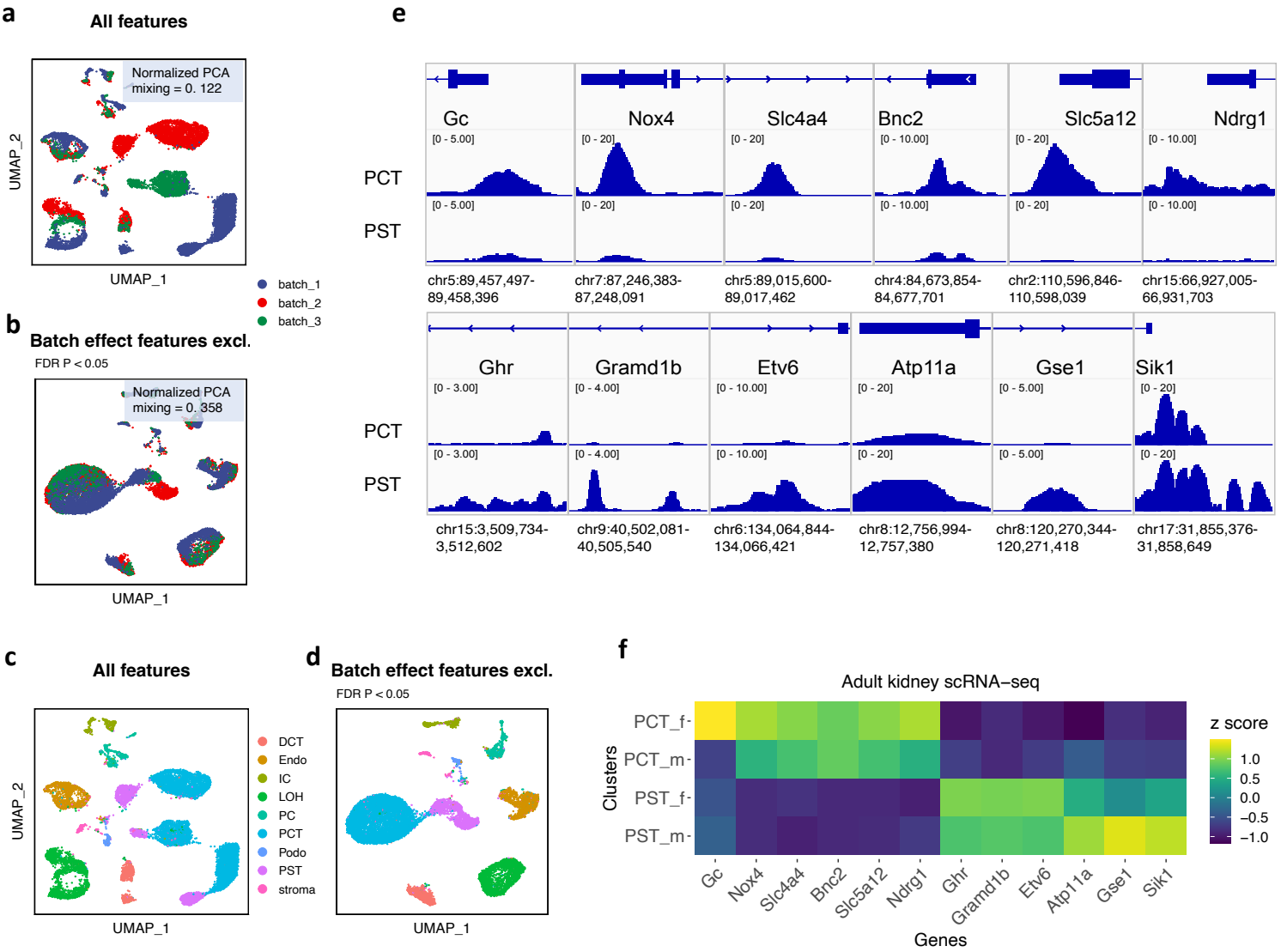


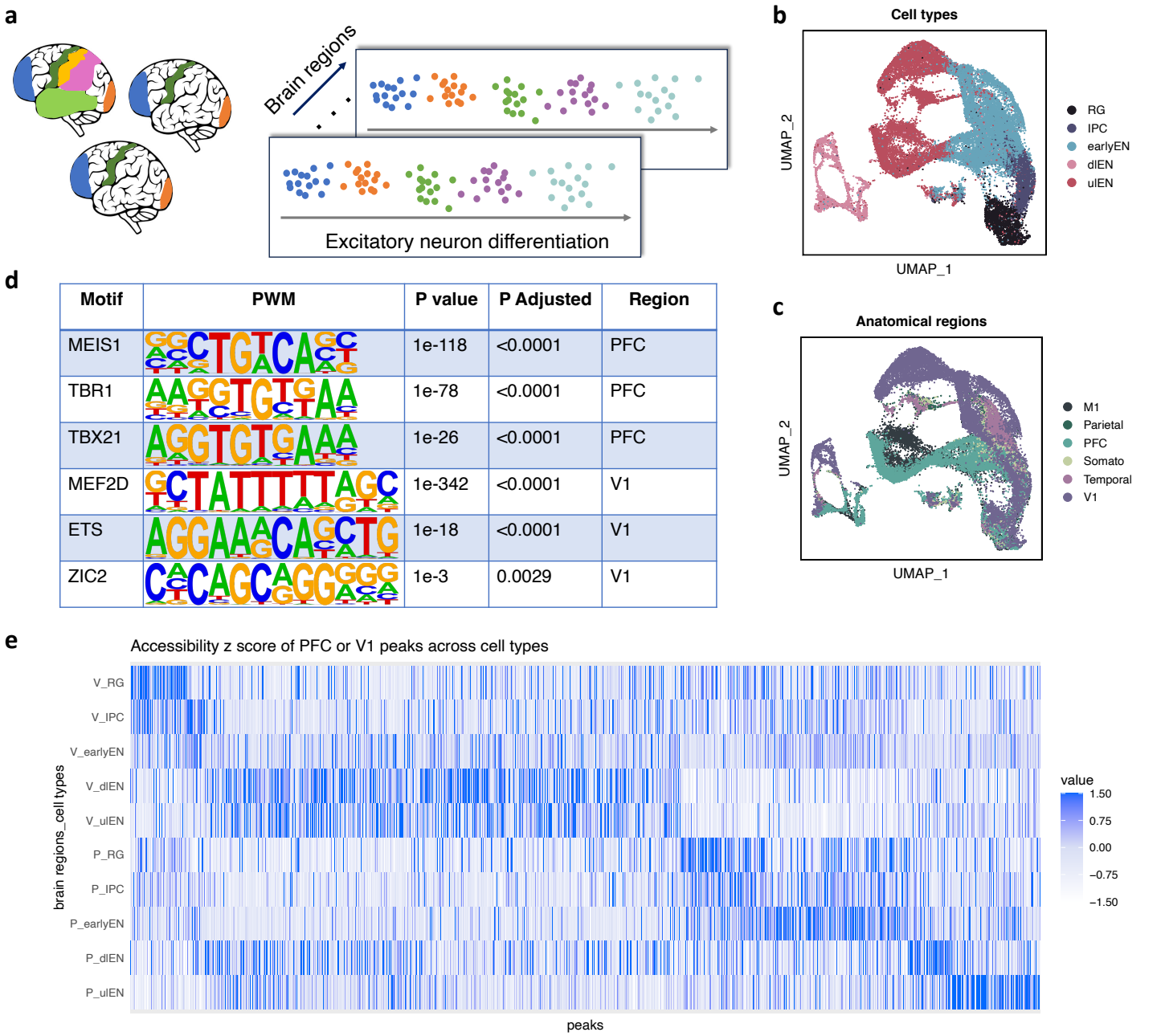
Figure 4. Application of PACS to the mouse kidney dataset.

a-b. UMAP dimension reduction plot constructed with all features (**a**) or after excluding features with significant batch effect (**b**), colored by batch labels. Features with batch effect are detected with PACS differential test module, and FDR multiple testing correction is applied. Normalized PCA mixing represents the normalized mixing score calculated in the PCA space, with 1 being no batch effect and 0 being strongest batch effect.

c-d. UMAP dimension reduction plot constructed with all features (**c**) or after removing features with batch effect (**d**), colored by cell types.

e. IGV plot of peak summits around cell type-specific genes identified by PACS, for PCT and PST cell types. The list of cell type specific genes is generated with GREAT enrichment analysis using differentially accessible peaks.

f. Heatmap of normalized gene expression z scores for the scRNA-seq data from male (-m) and female (-f) kidneys. The list of genes match those from the panel **f**.



Motif	PWM	P value	P Adjusted	Region
MEIS1		1e-118	<0.0001	PFC
TBR1		1e-78	<0.0001	PFC
TBX21		1e-26	<0.0001	PFC
MEF2D		1e-342	<0.0001	V1
ETS		1e-18	<0.0001	V1
ZIC2		1e-3	0.0029	V1

Figure 5. Application of PACS to the developing human brain data.

a. Illustration of the developing human brain dataset. The subset of data we analyzed are composed of samples from three donors across six brain anatomical regions, and we focused on the excitatory neuron lineage.

b-c. UMAP visualization of the data complexity, with points colored by cell type (**b**) or anatomical regions (**c**). RG, radial glia; IPC, intermediate (neuro-) progenitor cells; earlyEN, early excitatory neurons; dlEN, deep layer excitatory neurons; ulEN, upper layer excitatory neurons; M1, primary motor cortex; Parietal, dorsolateral parietal cortex; PFC, dorsolateral prefrontal cortex; Somato, primary somatosensory cortex; Temporal, temporal cortex; V1, primary visual cortex.

d. Motif enrichment results for PFC- and V1-specific peaks identified using PACS. PWM, position weight matrix.

e. Accessibility z score of PFC and V1 peaks across five cell types.

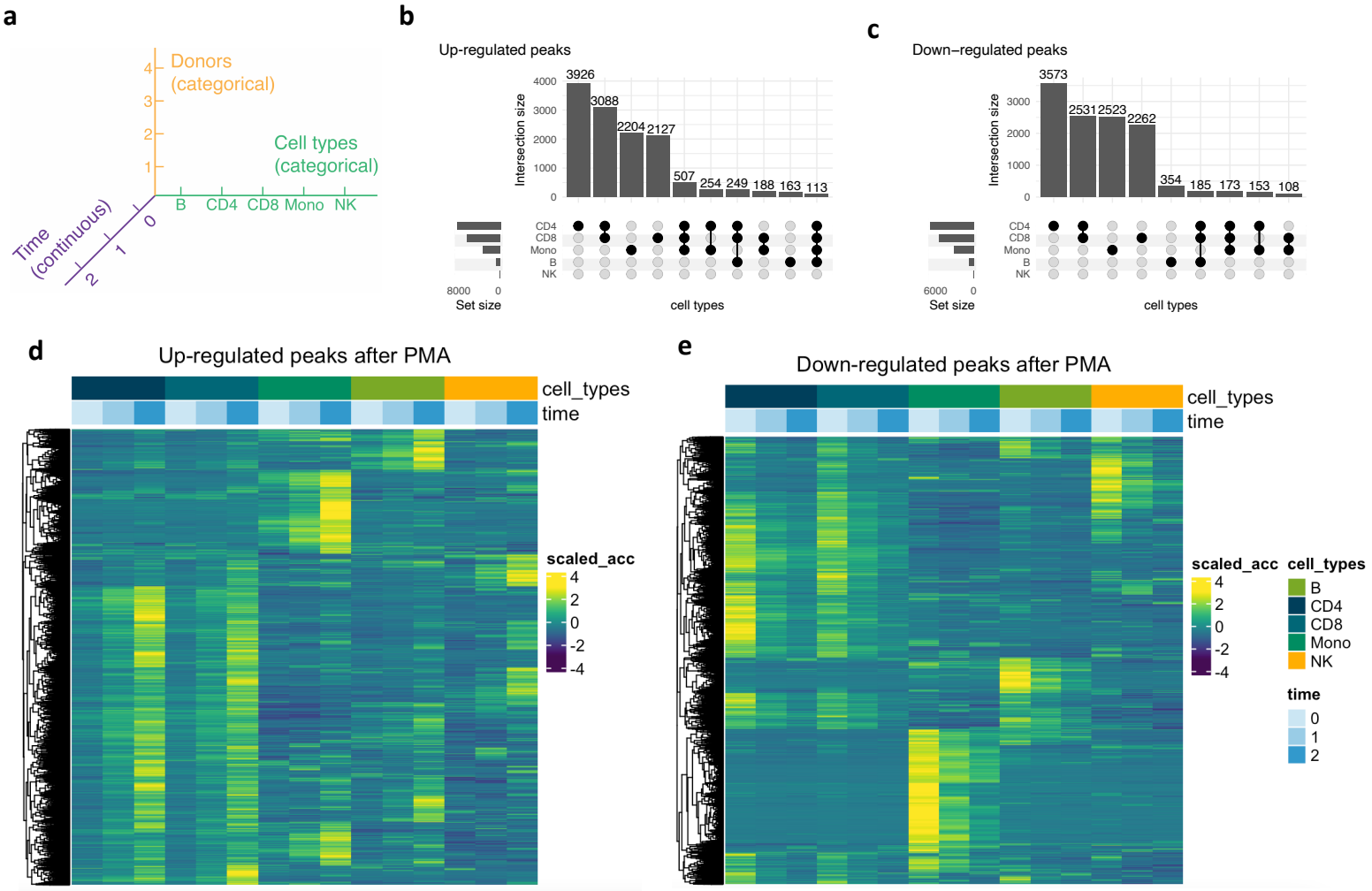


Figure 6. Application of PACS to time-series dataset from human PBMC treatment data.

a. Factor landscape of the PBMC treatment dataset. Here, another layer of factor is the four different treatments, which can also be jointly considered in the model, but for demonstration purposes, we only focus on the PMA treatment effect. The control time point is considered as time 0, and the times of one and six hours after treatment are considered as time 1 and 2, respectively.

b-c. Summary of significant up- or down- regulated peaks after PMA treatment for each cell type.

d-e. Heatmap of significant up- or down- regulated peaks after PMA treatment, grouped by time point and cell type. The color scale (scaled_acc) represents the accessibility z score.