Model-based compound hypothesis testing for snATAC-seq data with PACS

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.
Patricia M. Williams Term Professor, Department of Biology
Adjunct 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 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, 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 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. The level of accessibility of CREs usually indicates its activities[10], and in a cell, the activities of CREs are dynamic, dependent on various physiological factors such as cell type[1,3], developmental state[6,7], and spatial location of the tissue[11,12]. Identifying the sets of elements whose accessibility is governed by certain physiological factors is essential in understanding the cis-regulatory codes of biological processes[13,14].

Among all the factors that drive the accessibility of CREs, only some factors are experimentally controlled, for example, tissue type and location of cell collection. 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). Here, we will call the known factors that affect or predict accessibility as independent variables following standard experimental design terminology. We note that sometimes the values of the independent variables 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.

With the emergence of atlas-scale snATAC-seq data collection, available data usually involve multi-factorial predictive variables (e.g., health condition, donor variations, time points). A fundamental question with ATAC-seq data is whether any of the variables significantly affect or predict the accessibility of certain CREs; for example, whether cell type affects accessibility. Existing approaches for hypothesis testing typically involve pairwise testing between two states of a single factor (e.g., tests for Differential Accessible Regions, DARs, between two cell types)[15,16,17]. However, these approaches do not allow testing complex compound hypotheses that involve multiple independent variables. When there are multiple independent variables for a response variable, a standard approach is to model the response by a generalized linear model through an appropriate link function[18]. However, the standard generalized linear model (GLM) framework faces challenges in handling technical biases arising from heterogeneity in sequencing coverage of each cell and overall extreme sparsity of data. To address these limitations, we present a new statistical framework that extends the GLM framework to incorporate sample-specific missing data. Here, we derived a missing-corrected cumulative logistic regression (mcCLR) for the analysis of single cell open chromatin data. Furthermore, we utilized the Firth regularization[19,20] 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 methods for complex compound analysis tasks in snATAC-seq data analysis, 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 design matrix, $F_{C\times J}$ to represent these variables, where C is the number of cells and J is the number of independent 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[21,22]. As snATAC-seq can recover quantitative information on the density and distribution of nucleosomes[17,23], 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[17]. For standard snATAC-seq experiments, PIC counts follow a size-filtered signed Poisson (ssPoisson) distribution for a given Tn5 insertion rate[17]. 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**, lower 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:

$$\begin{aligned}
\text{logit}(\text{P}(Y_{cm} \geq 1)) &= \alpha^{(1)} + \textstyle\sum_{j=1}^{J} \beta_j F_{cj}\,, \text{ where } \text{P}(Z_{cm} \geq 1) = \text{P}(Y_{cm} \geq 1) q_c \\
\text{logit}(\text{P}(Y_{cm} \geq 2)) &= \alpha^{(2)} + \textstyle\sum_{j=1}^{J} \beta_j F_{cj}\,, \text{ where } \text{P}(Z_{cm} \geq 2) = \text{P}(Y_{cm} \geq 2) q_c \\
&\ldots \\
\text{logit}(\text{P}(Y_{cm} \geq T)) &= \alpha^{(\text{T})} + \textstyle\sum_{j=1}^{J} \beta_j F_{cj}\,, \text{ where } \text{P}(Z_{cm} \geq T) = \text{P}(Y_{cm} \geq T) q_c
\end{aligned}$$

(Eq. 1)

where 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 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 independent variables (and their interactions) on accessibility can be tested by the null hypothesis of $\beta_i = 0$ with a likelihood ratio test. 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 $\text{P}(Y_{cm} \geq 1 | f_c)$ 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[24]). Here, we developed a regularized model to resolve the issues

with sparsity in snATAC-seq data by generalizing the Firth logistic regression model[19,25], 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.

Application of PACS to cell type identification

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[26], two mouse kidney datasets[6], and one marmoset brain dataset[27]. In the two human cell line datasets, the cell line labels are annotated by their SNP information[26], 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[27], 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 if there are 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 independent 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 the standard setting of a single factor model (cell types) for type I error and power, for PACS and four existing methods: ArchR[26], Seurat/Signac[16], snapATAC[15], and Fisher's exact test. ArchR conducts the Wilcoxon rank-sum test on the subsampled cells from the initial groups, where the number of sequencing reads between two subsamples is matched. Seurat utilizes the standard logistic regression model[28], but with group labels as the dependent variable and read counts and total reads as independent variables. The sparsity problem that can result in perfect separability is not resolved in this method. SnapATAC conducts a test on the pseudo-bulk data of two groups and utilizes the edgeR[29] regression-based 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 the human cell line dataset[26]. Regions with non-trivial insertion rate differences (i.e., effect size greater than 0.1) were considered to have true cell type effects, while the remaining regions were set to the same insertion rates as their average rates, and thus having no differential effect. We randomly sampled 10,000 non-differential features to assess the type I error and 10,000 differential features to evaluate power, with varying numbers of cells in each group (from 250 to 1000). **Fig. 3a** shows that PACS controlled type I error at the specified level across all conditions. Among the methods that control type I error, 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, and the ad hoc "bvc" choice in snapATAC may result in a miscalibrated test with a low type I error and power. The q-q plots of the five methods are shown in **Supplementary Fig. 2a-e**.

To evaluate the performance under a multi-factor model, we next simulated another snATAC-seq dataset with two spatial locations (S1 and S2) and two cell types (T1 and T2). We introduced sample imbalance by setting S1 to contain 1600 T1 cells and 800 T2 cells, and S2 to contain 400 T1 cells and 1200 T2 cells. The spatial effect term was considered to affect features both with and without cell type effects. Specifically, one-third of the features with (and without) cell type effects were assumed to also have spatial effects, with fold change in accessibility of 0.75 or 0.125. For the methods that cannot directly test effects for multiple factors, two strategies were used. The first is called the "naïve test", where spatial location is ignored, and the test is conducted between two cell types. The second is called the "stratified test", where we stratified the dataset by spatial location and conducted a pairwise test between cell types on each stratum, followed by using the standard Fisher combination test to combine p-values (**Methods**). Across all methods and test strategies, only snapATAC (naïve and stratified), ArchR-stratified, and PACS controlled type I error at the specified level (**Fig. 3c**); PACS remained the most powerful test and detected 7.6, 5.9, and 1.2-fold more true differential peaks compared with those identified by snapATAC-naïve, snapATAC-stratified, and ArchR-stratified, respectively (**Fig. 3d**, **Supplementary Table 3**).

We then simulated a time-series dataset with five time points, to evaluate our model performance for ordinal covariates. We assumed two temporal trends of accessibility, linear and quadratic trends. To put this in a biological setting, the quadratic trend may represent the presence of an acute spike response and the linear trend may represent temporally accumulating chronic responses. The PACS framework could detect both linear and quadratic signals, and its power is dependent on the "effect sizes" defined as the log fold change of accessibility between the highest and lowest accessibility (**Fig. 3e-f**).

We also evaluated the PACS model in real datasets. As the ground truth is unknown, we utilized a sampling-based approach. We used randomly permuted cell type labels to estimate the type I error. To evaluate power, we conducted tests on cell types and treated the consensus DAR set from all methods as "true DARs" (after type I error control, see **Methods**). For the standard two-group DAR test, our method consistently controlled type I error and achieved high power, across different datasets (**Fig. 3g-h**, **Supplementary Fig. 2f-i**).

Taken together, we demonstrated with simulated and real datasets that PACS is a flexible test framework with well-calibrated test statistics.

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[30] 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[31], Seurat[32], and Harmony[30]), 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.[33] 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[34,35] 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[36] (**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[37]. 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[38]. 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[39]. 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[35], 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[40,41,42,43]. 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[44], and *TBX21* is an immune cell TF that also directs T-cell homing to pro-inflammatory sites via regulation of *CXCR3* expression[45]. **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 is characterized by uneven data capturing and data sparsity. For scRNA-seq data, data normalization has been an essential step for adjusting for uneven data capturing; however, in scATAC-seq data, such a notion does not exist, which remains a challenge for data analysis. Here, PACS resolves the issue of sequencing coverage variability in scATAC-seq data by combining a probability model of the underlying group-level accessibility with an independent cell-level capturing probability. We applied PACS to tasks of (supervised) cell type annotation, showed its improved performance compared with the Naïve Bayes model that does not consider cell-specific capturing probability.

With more data being generated for different tissue conditions, atlas-level data integration is essential for understanding tissue dynamics under various conditions. The cell type annotation framework enabled us to transfer the cell type annotation from reference dataset to another dataset, which resolves one challenge in integrative data analysis. Another challenge of data integration is to jointly model various factors (e.g., cell type, spatial locations) that govern cellular CRE activities. Standard GLM framework could not address the uneven data capturing in snATAC-seq data, so we developed a statistical model that extends the standard GLM framework to account for cell-specific missing data. By utilizing this missing-corrected cumulative logistic regression (mcCLR) model with regularization, PACS can conduct multi-covariate hypothesis tests and can be used for 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.

We have previously derived a parametric model of the snATAC-seq read count, called size-filtered signed Poisson distribution (ssPoisson)[17]. Here, we treat the insertion rate as a latent variable and directly model the paired insertion counts (PIC) of the data with an extended cumulative logistic regression model, which enabled fast and efficient computation. Future research will be conducted to explore the potential 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.

412 **References:**

413

414 1. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory
415 variation. *Nature* **523**, 486–490 (2015).
416 2. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution.
417 *Nat. Commun.* **9**, 3647 (2018).
418 3. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility.
419 *Cell* **174**, 1309-1324.e18 (2018).
420 4. Arda, H. E. *et al.* A Chromatin Basis for Cell Lineage and Disease Risk in the Human
421 Pancreas. *Cell Syst.* **7**, 310-322.e4 (2018).
422 5. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**,
423 5985-6001.e19 (2021).
424 6. Miao, Zhen *et al.* Single cell regulatory landscape of the mouse kidney highlights cellular
425 differentiation programs and disease targets. *Nat. Commun.*
426 doi:https://doi.org/10.1038/s41467-021-222266-1.
427 7. Cusanovich, D. A. *et al.* The cis-regulatory dynamics of embryonic development at single-cell
428 resolution. *Nature* **555**, 538–542 (2018).
429 8. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune
430 cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
431 9. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human
432 hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
433 10. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory
434 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. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* **83**, 373–392 (2023).
14. 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).
15. Fang, R. *et al.* Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).
16. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
17. 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.
18. Agresti, A. *Foundations of linear and generalized linear models*. (John Wiley & Sons, 2015).
19. FIRTH, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).
20. Heinze, G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Stat. Med.* **25**, 4216–4226 (2006).
21. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
22. Li, Y. E. *et al.* An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598**, 129–136 (2021).

23. 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).
24. Agresti, A. *Categorical data analysis*. vol. 792 (John Wiley & Sons, 2012).
25. Heinze, G. & Schemper, M. A solution to the problem of separation in logistic regression. *Stat. Med.* **21**, 2409–2419 (2002).
26. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
27. Bakken, T. E. *et al.* Comparative cellular analysis of motor cortex in human, marmoset and mouse. *Nature* **598**, 111–119 (2021).
28. 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).
29. 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).
30. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
31. 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).
32. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

480 33. Chari, T. & Pachter, L. *The Specious Art of Single-Cell Genomics*.
481 http://biorxiv.org/lookup/doi/10.1101/2021.08.25.457696 (2021)
482 doi:10.1101/2021.08.25.457696.
483 34. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions.
484 *Nat. Biotechnol.* **28**, 495–501 (2010).
485 35. Tanigawa, Y., Dyer, E. S. & Bejerano, G. WhichTF is functionally important in your open
486 chromatin data? *PLOS Comput. Biol.* **18**, e1010378 (2022).
487 36. Ransick, A. *et al.* Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the
488 Mouse Kidney. *Dev. Cell* **51**, 399-413.e7 (2019).
489 37. Nowakowski, T. J. *et al.* Spatiotemporal gene expression trajectories reveal developmental
490 hierarchies of the human cortex. *Science* **358**, 1318–1323 (2017).
491 38. Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics.
492 *Cell Genomics* **2**, 100166 (2022).
493 39. Ai, W., Li, H., Song, N., Li, L. & Chen, H. Optimal Method to Stimulate Cytokine
494 Production and Its Use in Immunotoxicity Assessment. *Int. J. Environ. Res. Public. Health*
495 **10**, 3834–3842 (2013).
496 40. Brignall, R. *et al.* Integration of Kinase and Calcium Signaling at the Level of Chromatin
497 Underlies Inducible Gene Activation in T Cells. *J. Immunol.* **199**, 2652–2667 (2017).
498 41. Shin, H.-J., Lee, J.-B., Park, S.-H., Chang, J. & Lee, C.-W. T-bet expression is regulated by
499 EGR1-mediated signaling in activated T cells. *Clin. Immunol.* **131**, 385–394 (2009).
500 42. Dagan-Berger, M. *et al.* Role of CXCR3 carboxyl terminus and third intracellular loop in
501 receptor-mediated migration, adhesion and internalization in response to CXCL11. *Blood*
502 **107**, 3821–3831 (2006).

503 43. Cooke, M. *et al.* Differential Regulation of Gene Expression in Lung Cancer Cells by
504 Diacyglycerol-Lactones and a Phorbol Ester Via Selective Activation of Protein Kinase C
505 Isozymes. *Sci. Rep.* **9**, 6041 (2019).
506 44. Seo, H. *et al.* Dual-specificity phosphatase 5 acts as an anti-inflammatory regulator by
507 inhibiting the ERK and NF-κB signaling pathways. *Sci. Rep.* **7**, 17348 (2017).
508 45. Stolarczyk, E., Lord, G. M. & Howard, J. K. The immune cell transcription factor T-bet.
509 *Adipocyte* **3**, 58–62 (2014).
510 46. Winship, C. & Mare, R. D. Regression Models with Ordinal Variables. *Am. Sociol. Rev.* **49**,
511 512 (1984).
512 47. Christensen, R. H. B. *Sensometrics: Thurstonian and Statistical Models*. (Technical
513 University of Denmark, 2012).
514 48. Venzon, D. J. & Moolgavkar, S. H. A Method for Computing Profile-Likelihood-Based
515 Confidence Intervals. *Appl. Stat.* **37**, 87 (1988).
516 49. Adey, A. C. Tagmentation-based single-cell genomics. *Genome Res.* **31**, 1693–1705 (2021).
517 50. Duttke, S. H., Chang, M. W., Heinz, S. & Benner, C. Identification and dynamic
518 quantification of regulatory elements using total RNA. *Genome Res.* (2019)
519 doi:10.1101/gr.253492.119.
520 51. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29
521 (2021).
522
523

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[26] (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[27] (the Brain Cell Data Center RRID SCR_017266; https://biccn.org/data),
human PBMC time-series stimulation data[38] (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[17]. 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:

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

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[17]. 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[46,47] 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[25,48]. 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[49]), we derived the size-filtered signed Poisson (ssPoisson) distribution to model this data generation process[17]. With the observed counts, we estimated the insertion rate parameters for two cell types, and regions with true open probability difference greater than 0.05 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

692 Building upon the single factor setting, we further assumed the data to contain two cell types (T1
693 and T2) being sampled from two spatial locations (S1 and S2). The goal was to infer cell-type-
694 specific differential features while accounting for the spatial effect. We introduced sample
695 imbalance as frequently seen in real datasets. To be precise, we considered that S1 contained 1600
696 T1 cells and 800 T2 cells, while S2 contained 400 T1 cells and 1200 T2 cells. The spatial effect
697 was considered to affect features both with and without a cell type effect. Specifically, a third of
698 the features with (and without) a cell type effect showed an accessibility difference across batches,
699 with a fold change of 0.75 or 0.125. The peak by cell count data generation procedure is the same
700 as for the single factor setting.
701
702 *Data simulation for time-series differential test*
703 To evaluate model performance in situations where the design matrix contains ordinal covariates,
704 we simulated time-series snATAC-seq data across five time points. We assumed linear and
705 quadratic temporal effects on accessibility and set the effect size (log fold change) to be 0.3 or 0.5
706 between the two groups. The baseline accessibility was generated from the cell line data and the
707 peak by cell count data generation procedure is the same as for the single factor setting.
708
709 *Evaluating type I error and power in real datasets*
710 To estimate type I error in real data where the ground truth is unknown, we used a label
711 permutation approach, where the data in one cell type were divided randomly into two groups, and
712 a differential test was conducted between these groups. As this is randomly assigned, all features
713 were believed to be non-DAR, so the proportion of P values smaller than 0.05 is the empirical type
714 I error using real data. Then, we set the fifth rank percentile as the correct critical value for those
715 methods with type I errors greater than 0.05. We next conducted a test with two different cell types
716 using the calibrated critical values for each method. Since we do not know the true DAR set, we
717 defined the pseudo-true DAR peaks as the union DAR set of all tested methods, using their
718 corresponding new critical values. Power for each method was then calculated by the number of
719 DARs detected divided by the number of pseudo-true DARs. This approach is adopted from Ref.[17].
720
721 *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[16] 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.[33] 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

753 element m_{ij} representing the number of cells in the cell type i and batch j. Then the expected data-
754 level batch mixing score in the setting of no batch effect is given by
755
756 $$\mathrm{E}[\text{batch mixing score}] = \frac{1}{\sum_{i,j} m_{ij}}\{\sum_i[\sum_j m_{ij}(\frac{\sum_{k\neq j} m_{ik}}{\sum_k m_{ik}})]\} \quad \text{(Eq. 15)}$$
757
758 The normalized batch mixing score is the batch mixing score divided by the expected score under
759 random mixing, and thus a higher normalized batch mixing score indicates better mixing across
760 samples.
761
762 *Processing developing human brain data*
763 This dataset contains 18 specimens collected from human donors. For our study, we excluded
764 samples with unknown spatial locations (GW17, GW18, GW21) or samples not from the cortex
765 (MGE_GW20 and MGE_twin34). Here we focused on the excitatory neuron lineage, including
766 radial glia (RG), intermediate progenitor cells (IPCs), early excitatory neurons (earlyEN), deep
767 layer excitatory neurons (dlENs), and upper layer excitatory neurons (ulENs). We further excluded
768 the insular region for having too few cell counts (645 cells across five cell types). The data matrix
769 was saved as a binary matrix, so we implemented the missing-corrected logistic regression model
770 for the analyses of this data.
771
772 *DAR identification in the developing human brain data*
773 We constructed two models to identify the significant region effect of the excitatory neuron lineage.
774 Specifically, to identify the region effect, the systematic component of the PACS model is
775 specified as:
776
777 $$\alpha + \sum_{k=2}^{K}\gamma_k 1(G_c = g_k) + \sum_{l=2}^{L}\zeta\, 1(S_c = S_l) + \sum_{t=2}^{M}\tau_m 1(D_c = D_m) \quad \text{(Eq. 16)}$$
778
779 where G is the index of cell type, S is the index of spatial location, and D is the index of the donor.
780 The null hypothesis for the test is $H_0: \zeta = 0$. To identify the cell type specific region effect, we
781 additionally included the interaction terms between each cell type and spatial location, and the test
782 was conducted for each interaction term.

783

784 *Motif enrichment analysis*

785 The motif enrichment analysis was conducted with Homer[50]. The list of significant DAR peaks is
786 used as input for the analysis, with the size of the search region specified as 300 bp around the
787 peak center. The reported motif enrichment scores are FDR-corrected P values from the known
788 motif results.

789

790 *DAR identification in the human PBMC treatment data*

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

793

794 $$\alpha + \sum_{k=2}^{K} \gamma_k 1(G_c = g_k) + \kappa E + \sum_{t=2}^{M} \omega_t 1(D_c = D_t) \quad \text{(Eq. 17)}$$

795

796 where G is the index of cell type, E is the experimental time index (0, 1, 2 corresponds to control,
797 1h, and 6h after treatment, respectively), and D is the donor index. The null hypothesis for the test
798 is $H_0: \kappa = 0$.

799

800 *Gene and pathway enrichment with GREAT*

801 We used the GREAT method (v. 4.0.4) to conduct gene and enrichment analysis[34], with DARs as
802 input and default parameter settings. The output from GREAT for the human PBMC data can be
803 found in the **Supplementary Tables 17-18**.

804

805 **Figure legends:**

806

807 **Figure 1. PACS modeling framework.**

808 Upper panel: Illustration of the latent accessibility of cells. Multiple factors including cell types,
809 developmental stages, spatial locations etc. determines the chromatin structure and configurations
810 of corresponding cell groups. These different chromatin structures result in the variable Tn5
811 insertion rates in the ATAC-seq experiments. The readout of ATAC assays are paired insertion
812 counts (PIC), which are crude measures of latent insertion rates.

Lower 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.

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

a-d. Parameter estimation accuracy evaluated using simulation data. 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).

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

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).

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

a-b. Type I error and power evaluation using single-factor simulation data.

c-d. Type I error and power evaluation using two-factor simulation data. Methods with "-n" represents the setting of Naïve test, where other factors are ignored when testing the factors of interest. Methods ending with "-s" represent the stratified test where we stratify on other factors and test the factors of interest within the strata.

e. Illustration of linear and quadratic effects of treatment on accessibility across time points. Effect sizes are defined as the fold change between the highest accessibility over the lowest accessibility, across five time points.

f. Evaluation of power in detecting linear and quadratic temporal effects using simulated data with different effect sizes.

g-h. Type I error and power evaluation using empirical adult mouse kidney data.

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

843 **a-b**. UMAP dimension reduction plot constructed with all features (a) or after excluding features
844 with significant batch effect (b), colored by batch labels. Features with batch effect are detected
845 with PACS differential test module, and FDR multiple testing correction is applied.
846 **c-d**. UMAP dimension reduction plot constructed with all features (a) or after removing features
847 with batch effect (b), colored by cell types.
848 **e**. IGV plot of peak summits around cell type-specific genes identified by PACS, for PCT and PST
849 cell types. The list of cell type specific genes is generated with GREAT enrichment analysis using
850 differentially accessible peaks.
851 **f**. Heatmap of normalized gene expression z scores for the scRNA-seq data from male (-m) and
852 female (-f) kidneys. The list of genes match those from the panel **f**.
853
854 **Figure 5. Application of PACS to the developing human brain data.**
855 **a**. Illustration of the developing human brain dataset. The subset of data we analyzed are composed
856 of samples from three donors across six brain anatomical regions, and we focused on the excitatory
857 neuron lineage.
858 **b-c**. UMAP visualization of the data complexity, with points colored by cell type (b) or anatomical
859 regions (c). RG, radial glia; IPC, intermediate (neuro-) progenitor cells; earlyEN, early excitatory
860 neurons; dlEN, deep layer excitatory neurons; ulEN, upper layer excitatory neurons; M1, primary
861 motor cortex; Parietal, dorsolateral parietal cortex; PFC, dorsolateral prefrontal cortex; Somato,
862 primary somatosensory cortex; Temporal, temporal cortex; V1, primary visual cortex.
863 **d**. Motif enrichment results for PFC- and V1-specific peaks identified using PACS. PWM, position
864 weight matrix.
865 **e**. Accessibility z score of PFC and V1 peaks across five cell types.
866
867 **Figure 6. Application of PACS to time-series dataset from human PBMC treatment data.**
868 **a**. Factor landscape of the PBMC treatment dataset. Here, another layer of factor is the four
869 different treatments, which can also be jointly considered in the model, but for demonstration
870 purposes, we only focus on the PMA treatment effect. The control time point is considered as time
871 0, and the times of one and six hours after treatment are considered as time 1 and 2, respectively.
872 **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.

Supplementary Figure 1.

a-b. Parameter estimation accuracy evaluated using simulation data. 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).

c-j. Confusion matrix between true cell type labels and PACS-inferred (or Naïve Bayes-inferred) cell type labels for four datasets.

Supplementary Figure 2.

a-e. Quantile-quantile plots for P values under the null for five testing methods, using simulated data with no insertion rate difference.

f-i. Type I error and power evaluation using empirical cell line mixture data or marmoset brain data.

Supplementary Figure 3.

a-b. UMAP dimension reduction plot constructed after excluding features with significant batch effect (P value < 0.05, no FDR correction), colored by batch labels (a) or cell types (b). Features with batch effect are detected with PACS differential test module.

Supplementary Figure 4.

a. UMAP dimension reduction plot constructed with all features, colored by batch labels. This panel is identical to **Fig. 4a**, and is displayed here for examining feature plots in panels **b-l**.

b-l. Feature plots for top significant batch effect peaks determined by PACS.

Supplementary Figure 5.

a-c. Violin plots that summarize number of fragments in each cell across different donors (a), brain regions (b), or cell types (c), for the human brain data.

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.

Fig. 1

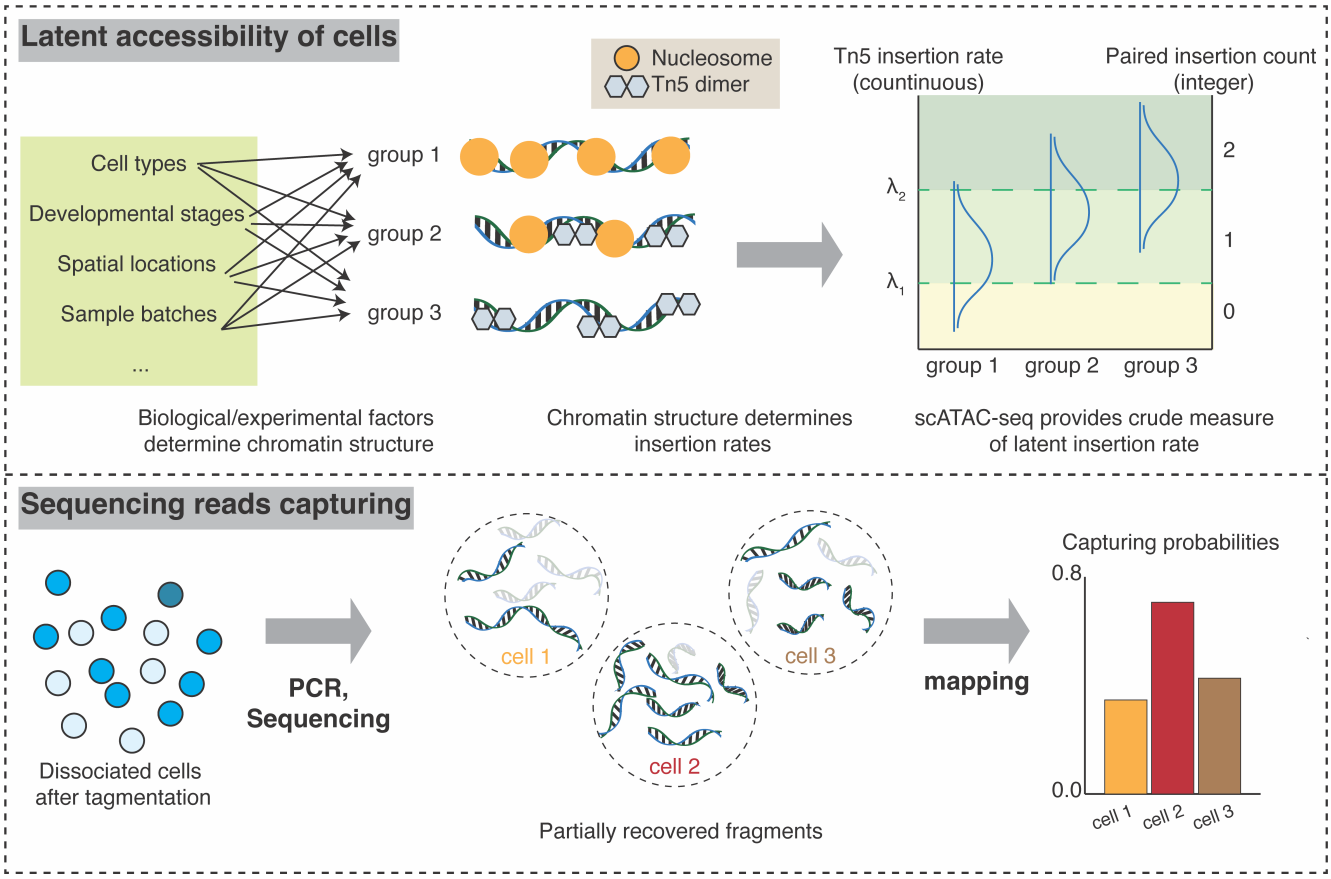

Latent accessibility of cells
Nucleosome
Tn5 dimer
Tn5 insertion rate (countinuous)
Paired insertion count (integer)
Cell types
Developmental stages
Spatial locations
Sample batches
...
group 1
group 2
group 3
λ2
λ1
2
1
0
group 1
group 2
group 3
Biological/experimental factors determine chromatin structure
Chromatin structure determines insertion rates
scATAC-seq provides crude measure of latent insertion rate
Sequencing reads capturing
Dissociated cells after tagmentation
PCR, Sequencing
cell 1
cell 2
cell 3
Partially recovered fragments
mapping
Capturing probabilities
0.8
0.0
cell 1
cell 2
cell 3

Fig. 2

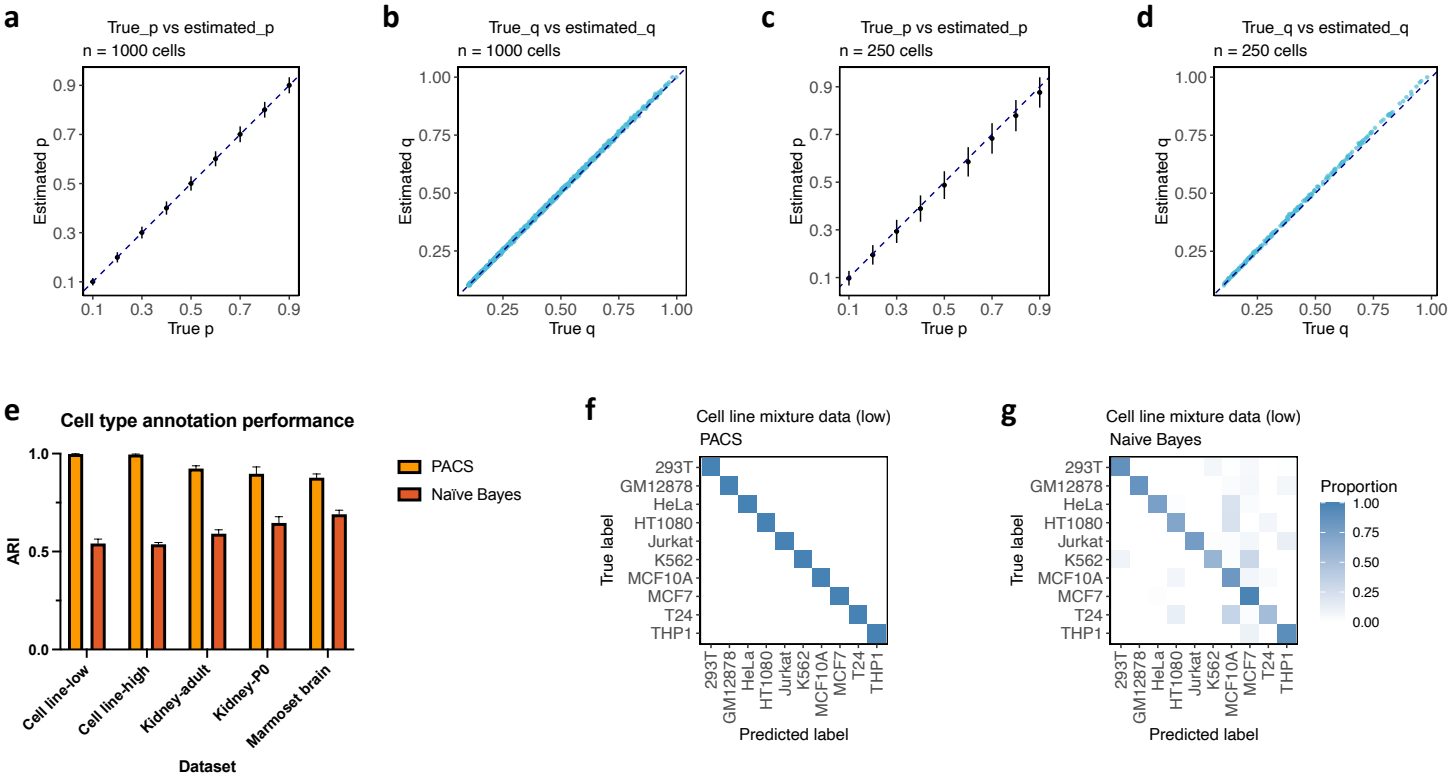

a
True_p vs estimated_p
n = 1000 cells
Estimated p
True p
b
True_q vs estimated_q
n = 1000 cells
Estimated q
True q
c
True_p vs estimated_p
n = 250 cells
Estimated p
True p
d
True_q vs estimated_q
n = 250 cells
Estimated q
True q
e
Cell type annotation performance
ARI
Cell line-low
Cell line-high
Kidney-adult
Kidney-P0
Marmoset brain
Dataset
PACS
Naïve Bayes
f
Cell line mixture data (low)
PACS
True label
Predicted label
293T
GM12878
HeLa
HT1080
Jurkat
K562
MCF10A
MCF7
T24
THP1
g
Cell line mixture data (low)
Naive Bayes
True label
Predicted label
Proportion

Fig. 3

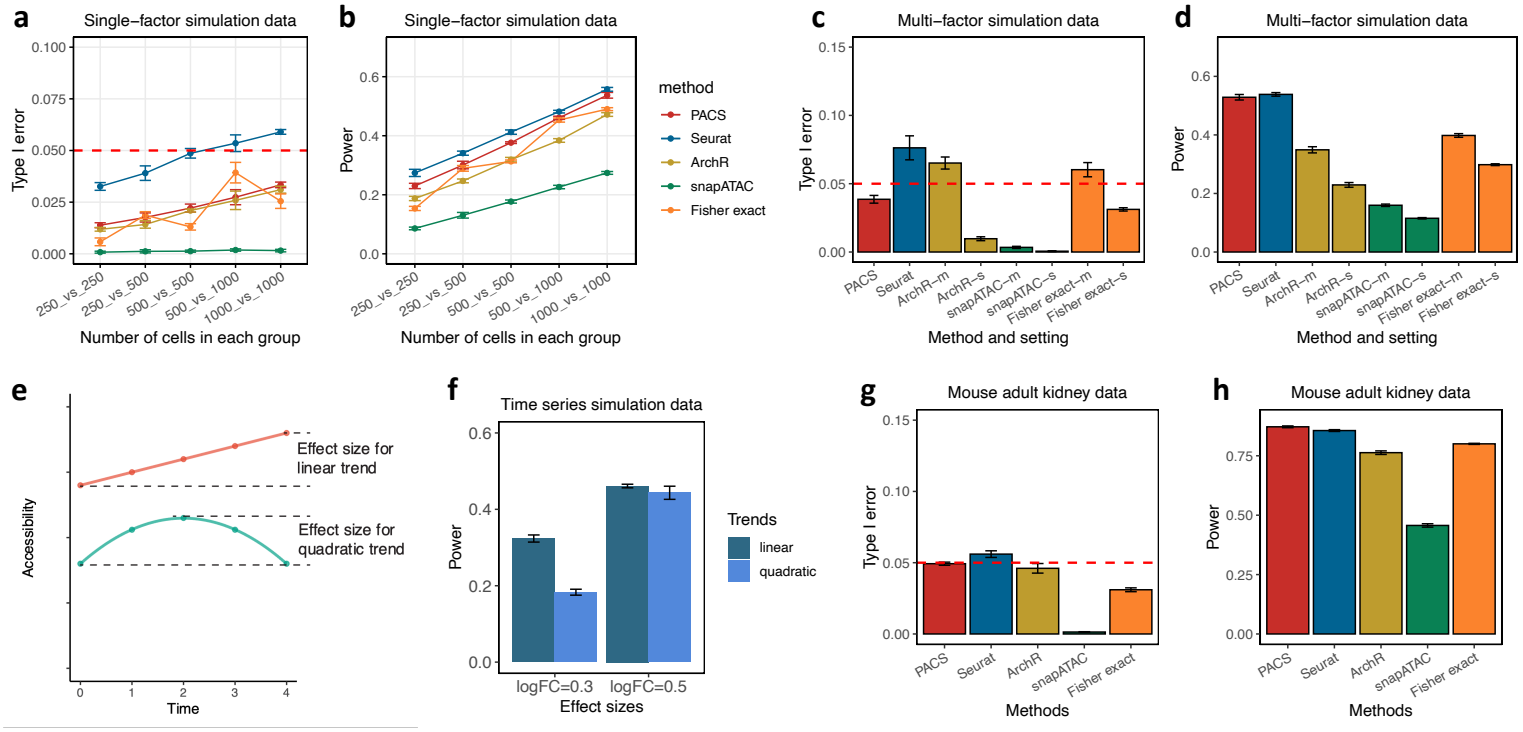

a
Single−factor simulation data
Type I error
0.100
0.075
0.050
0.025
0.000
250_vs_250
250_vs_500
500_vs_500
500_vs_1000
1000_vs_1000
Number of cells in each group
b
Single−factor simulation data
Power
0.6
0.4
0.2
0.0
Number of cells in each group
method
PACS
Seurat
ArchR
snapATAC
Fisher exact
c
Multi−factor simulation data
Type I error
0.15
0.10
0.05
0.00
PACS
Seurat
ArchR−m
ArchR−s
snapATAC−m
snapATAC−s
Fisher exact−m
Fisher exact−s
Method and setting
d
Multi−factor simulation data
Power
0.6
0.4
0.2
0.0
Method and setting
e
Accessibility
Effect size for linear trend
Effect size for quadratic trend
0
1
2
3
4
Time
f
Time series simulation data
Power
0.6
0.4
0.2
0.0
logFC=0.3
logFC=0.5
Effect sizes
Trends
linear
quadratic
g
Mouse adult kidney data
Type I error
0.15
0.10
0.05
0.00
PACS
Seurat
ArchR
snapATAC
Fisher exact
Methods
h
Mouse adult kidney data
Power
0.75
0.50
0.25
0.00
PACS
Seurat
ArchR
snapATAC
Fisher exact
Methods

Fig. 4

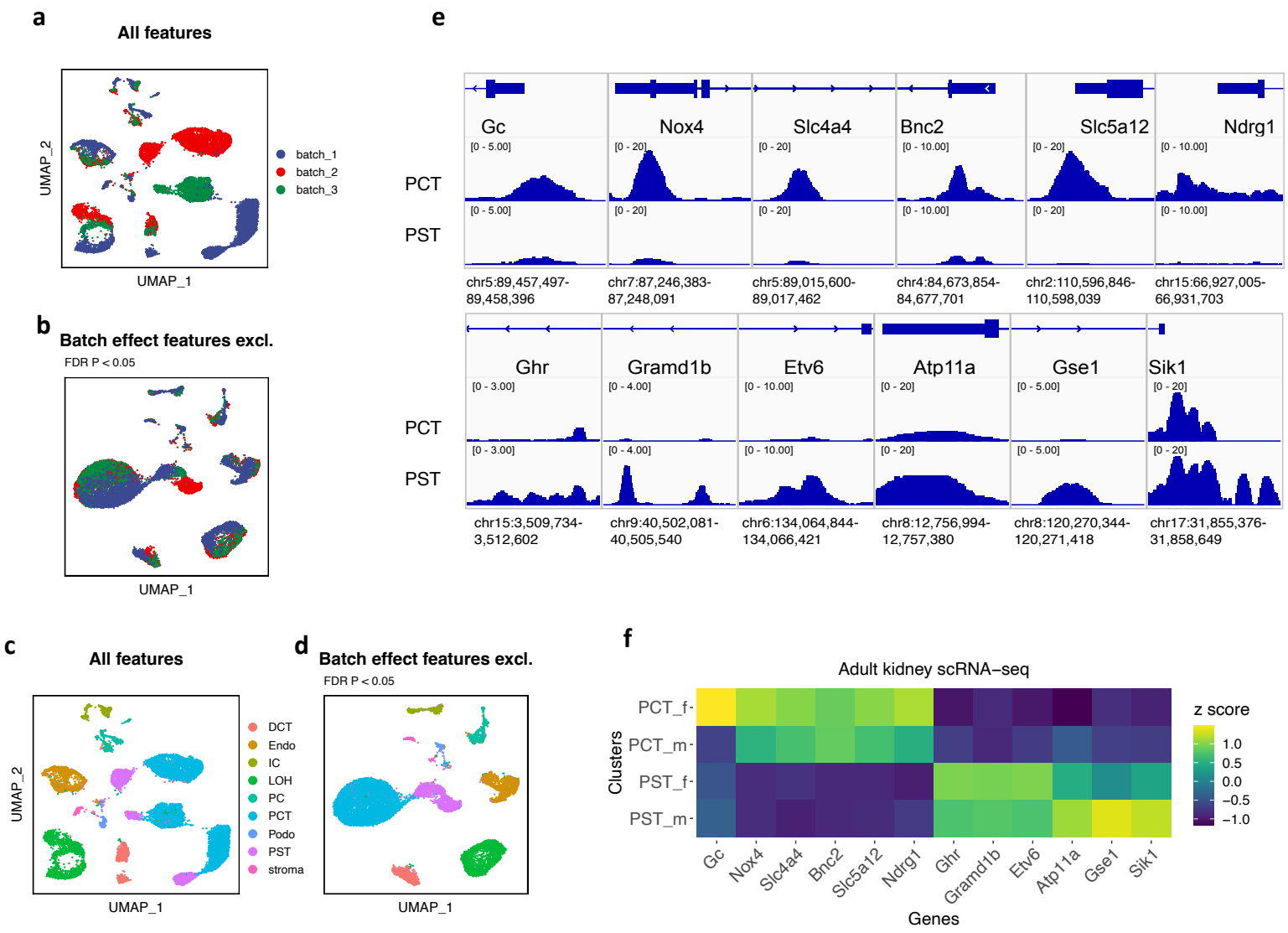
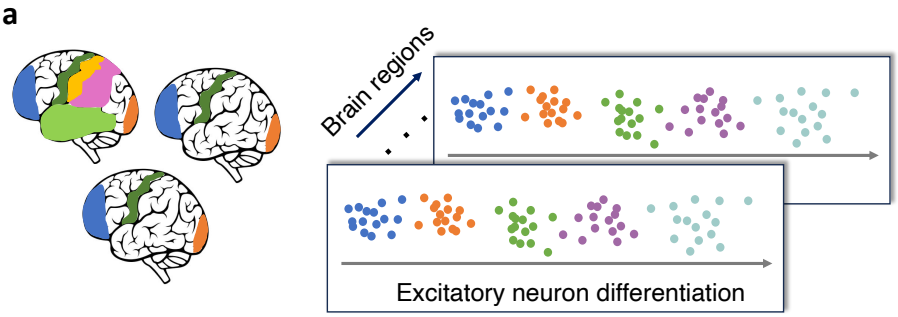
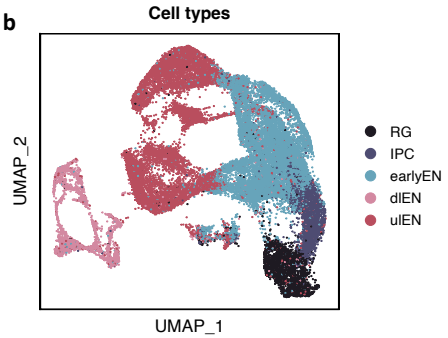
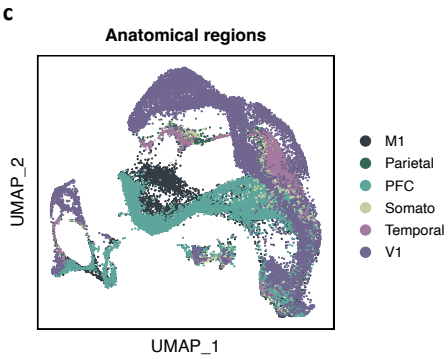

a
All features
UMAP_2
UMAP_1
batch_1
batch_2
batch_3
b
Batch effect features excl.
FDR P < 0.05
UMAP_1
c
All features
UMAP_2
UMAP_1
DCT
Endo
IC
LOH
PC
PCT
Podo
PST
stroma
d
Batch effect features excl.
FDR P < 0.05
UMAP_1
e
PCT
PST
Gc
Nox4
Slc4a4
Bnc2
Slc5a12
Ndrg1
[0 - 5.00]
[0 - 20]
[0 - 20]
[0 - 10.00]
[0 - 20]
[0 - 10.00]
chr5:89,457,497-89,458,396
chr7:87,246,383-87,248,091
chr5:89,015,600-89,017,462
chr4:84,673,854-84,677,701
chr2:110,596,846-110,598,039
chr15:66,927,005-66,931,703
Ghr
Gramd1b
Etv6
Atp11a
Gse1
Sik1
[0 - 3.00]
[0 - 4.00]
[0 - 10.00]
[0 - 20]
[0 - 5.00]
[0 - 20]
chr15:3,509,734-3,512,602
chr9:40,502,081-40,505,540
chr6:134,064,844-134,066,421
chr8:12,756,994-12,757,380
chr8:120,270,344-120,271,418
chr17:31,855,376-31,858,649
f
Adult kidney scRNA–seq
Clusters
PCT_f
PCT_m
PST_f
PST_m
Gc
Nox4
Slc4a4
Bnc2
Slc5a12
Ndrg1
Ghr
Gramd1b
Etv6
Atp11a
Gse1
Sik1
Genes
z score
1.0
0.5
0.0
−0.5
−1.0

Fig. 5







d

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

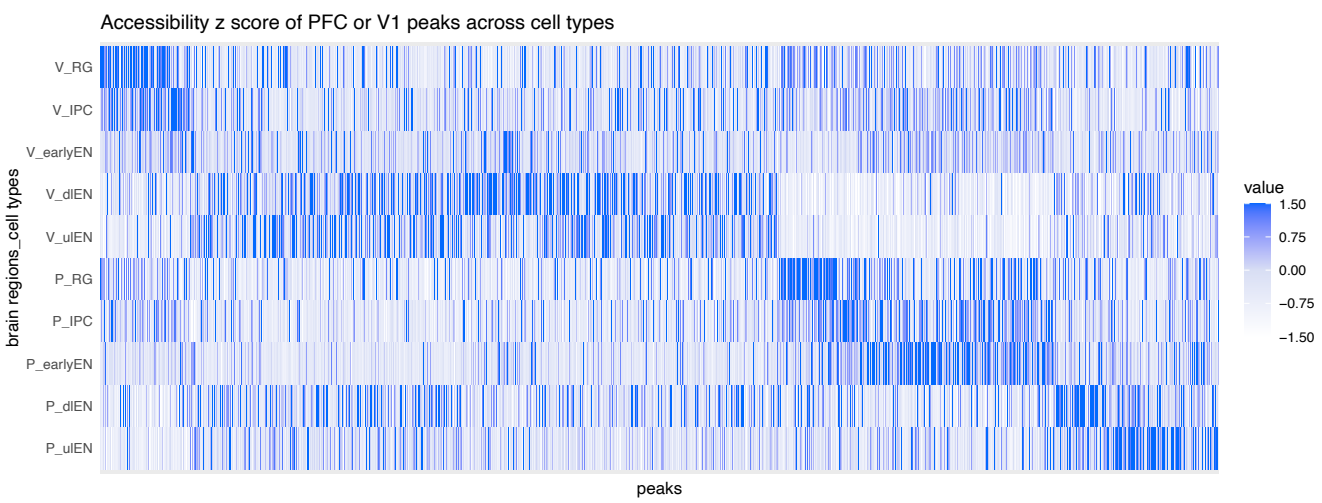
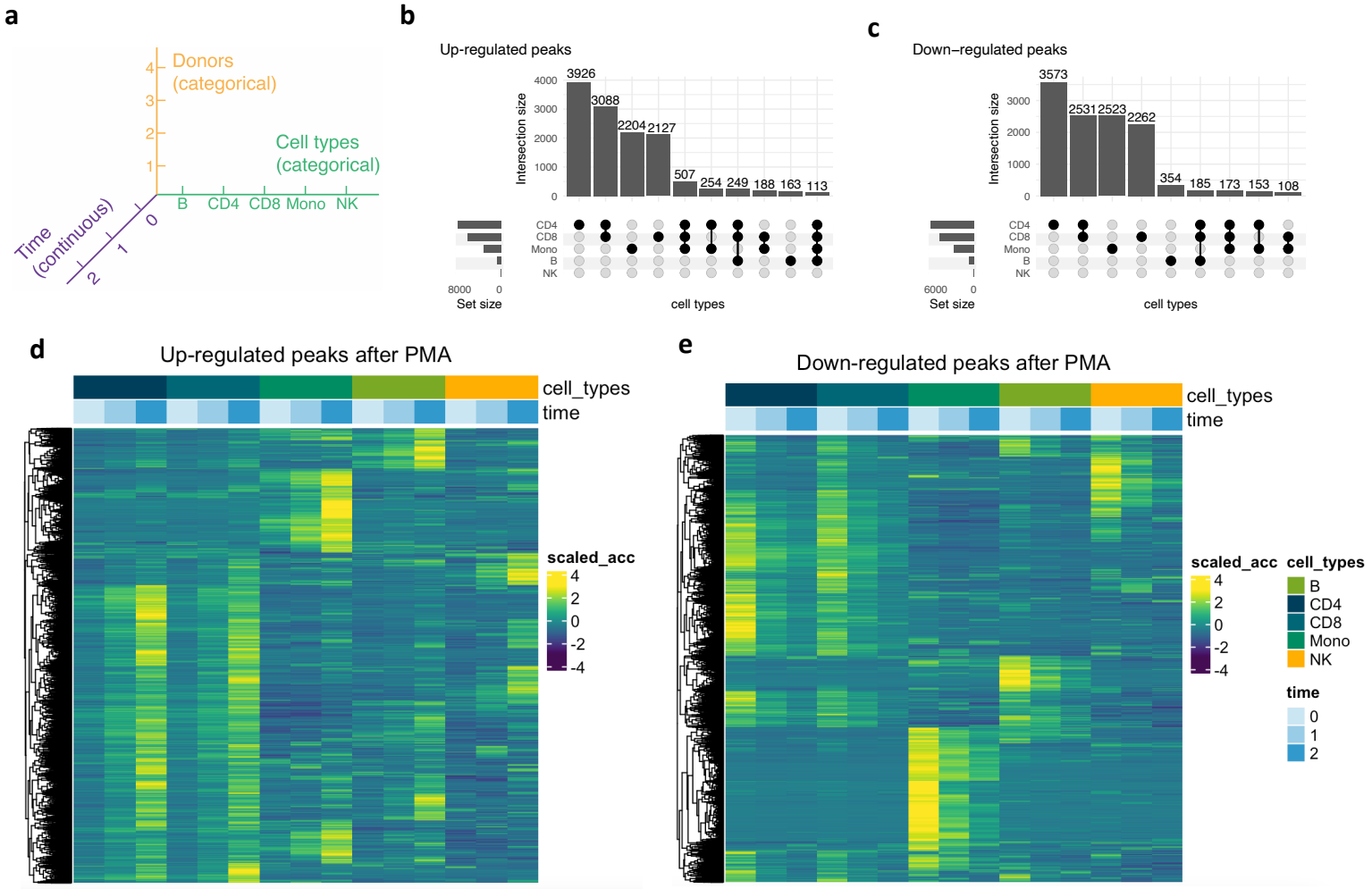


Fig. 6


a
Donors
(categorical)
Cell types
(categorical)
B CD4 CD8 Mono NK
Time
(continuous)
b
Up-regulated peaks
Intersection size
3926 3088 2204 2127 507 254 249 188 163 113
CD4 CD8 Mono B NK
Set size
cell types
c
Down–regulated peaks
3573 2531 2523 2262 354 185 173 153 108
d
Up-regulated peaks after PMA
cell_types
time
scaled_acc
e
Down-regulated peaks after PMA
cell_types
time
scaled_acc
B CD4 CD8 Mono NK
time
0 1 2