

Best practices for differential accessibility analysis in single-cell epigenomics

Received: 23 May 2023

Accepted: 24 September 2024

Published online: 11 October 2024

 Check for updatesAlan Yue Yang Teo^{1,2}, Jordan W. Squair^{1,2,3,6}✉, Gregoire Courtine^{1,2,3,6}✉ & Michael A. Skinnider^{1,2,4,5,6}✉

Differential accessibility (DA) analysis of single-cell epigenomics data enables the discovery of regulatory programs that establish cell type identity and steer responses to physiological and pathophysiological perturbations. While many statistical methods to identify DA regions have been developed, the principles that determine the performance of these methods remain unclear. As a result, there is no consensus on the most appropriate statistical methods for DA analysis of single-cell epigenomics data. Here, we present a systematic evaluation of statistical methods that have been applied to identify DA regions in single-cell ATAC-seq (scATAC-seq) data. We leverage a compendium of scATAC-seq experiments with matching bulk ATAC-seq or scRNA-seq in order to assess the accuracy, bias, robustness, and scalability of each statistical method. The structure of our experiments also provides the opportunity to define best practices for the analysis of scATAC-seq data beyond DA itself. We leverage this understanding to develop an R package implementing these best practices.

Protocol registration

The Stage 1 protocol for this Registered Report was accepted in principle on 25th October 2023. The protocol, as accepted by the journal, can be found at <https://doi.org/10.6084/m9.figshare.24541816.v1>.

The remarkable diversity of cell types and tissues that compose the human body arise from a single genome. This diversity is orchestrated by cell-type- and context-specific epigenetic programs that regulate the accessibility of specific regions of the genome¹. Epigenetic regulatory programs also choreograph tissue- and cell-type-specific responses to the myriad of endogenous and exogenous perturbations that humans encounter in their lifetime². The fundamental role of these regulatory programs in health and disease spurred the development of technologies to measure the accessibility of the genome^{3,4}. In turn, catalogues of genome accessibility promise to expose regulatory mechanisms that can be targeted therapeutically, and provide genetic access to molecularly-defined cell types^{5–8}.

Initial attempts to dissect these regulatory mechanisms employed bulk assays that aggregated signals over the multitude of different cell types within complex tissues^{9–13}. More recently, the advent of methods to measure genome accessibility at single-cell resolution enabled the discovery of regulatory programs underlying the identity and differentiation of individual cell types^{14–19}. Among these, the assay for transposase-accessible chromatin by sequencing (ATAC-seq) has emerged as the workhorse of single-cell epigenomics. Landmark single-cell ATAC-seq (scATAC-seq) studies established atlases of chromatin accessibility during fetal development^{20–22}, throughout the nervous system^{23–25}, and even within the entire human body²⁶. Exponential increases in the scale and availability of scATAC-seq are now

¹Defitech Center for Interventional Neurotherapies (.NeuroRestore), EPFL/CHUV/UNIL, Lausanne, Switzerland. ²NeuroX Institute and Brain Mind Institute, School of Life Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. ³Department of Clinical Neuroscience, Lausanne University Hospital (CHUV) and University of Lausanne (UNIL), Lausanne, Switzerland. ⁴Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. ⁵Ludwig Institute for Cancer Research, Princeton University, Princeton, NJ, USA. ⁶These authors jointly supervised this work: Jordan W. Squair, Gregoire Courtine, Michael A. Skinnider. ✉e-mail: jordan.squair@epfl.ch; gregoire.courtine@epfl.ch; skinnider@princeton.edu

triggering a transition from atlases of healthy tissues to dissection of cell-type-specific regulatory responses to disease and experimental perturbations^{7,27–30}.

The rapid evolution and widespread application of technologies to interrogate chromatin accessibility at single-cell resolution is exposing a lack of consensus on how to analyse the resulting data. Even fundamental questions, such as whether chromatin accessibility should be considered a qualitative or quantitative measurement, remain debated^{31,32}. Arguably the most important of these questions is how to identify differentially accessible (DA) regions of the genome. DA analysis is the methodological framework that permits the discovery of regulatory programs directing cell identity and perturbation responses. However, existing analysis packages implement markedly different approaches to DA^{33–43}, and analytical workflows implemented in different laboratories bear little resemblance to one another. This discordance not only raises the question of which DA methods are the most accurate, but also whether widely used DA methods are statistically valid or even prone to false discoveries.

Methods for differential analysis of bulk ChIP-seq and ATAC-seq datasets have previously been benchmarked^{44–46}. These benchmarks not only compared individual DA methods, but also evaluated the impact of factors such as sequencing depth, number of replicates, or the characteristics of the underlying signal (e.g., broad histone modifications vs. sharp transcription factor (TF)-binding events^{45,46}). However, these benchmarks generally relied on simulations or a small number of case studies with unclear ground truth, and did not address the new challenges raised by single-cell epigenomics, including the analysis of markedly sparser datasets comprising thousands of cells.

In this Registered Report, we present a systematic comparison of statistical methods that have been applied to identify DA regions in scATAC-seq. To enable a comprehensive analysis, we undertook an

exhaustive review of the literature to chart the landscape of statistical methods that have been used to perform DA analysis. We then applied each of these methods to a compendium of scATAC-seq datasets with matching bulk ATAC-seq or scRNA-seq datasets that provide a basis for comparison. We carried out a series of experiments to quantitatively assess the accuracy, bias, robustness, and scalability of each method for single-cell DA analysis. Moreover, these experiments also afforded us the opportunity to study more fundamental questions in the analysis of scATAC-seq, such as whether measurements within individual cells should be treated as qualitative or quantitative observations. The results of these experiments suggest best practices for the analysis of scATAC-seq datasets more broadly. We have translated this understanding into an R package that implements these best practices in order to empower users to perform accurate DA analysis in scATAC-seq data.

Results

Lack of consensus in single-cell DA analysis

We first sought to map the landscape of statistical methods that have been used to perform DA analysis. For this purpose, we conducted a comprehensive survey of the single-cell epigenomics literature. We identified a total of 118 primary publications that reported single-cell epigenomic datasets (Supplementary Data 1). This survey confirmed the widespread adoption of scATAC-seq (Fig. 1a). Chronological analysis revealed that the number of cells profiled in any given study has increased exponentially over time (Fig. 1b). Among the 13 statistical methods for DA analysis that were detected in this survey, the Wilcoxon rank-sum test was the most widely used (Fig. 1c). However, no method was used in more than 15 studies, and many DA methods were used in just one or two published analyses. Beyond this unsettling observation, we also observed disagreement on fundamental principles of DA analysis, such as whether or not to binarize

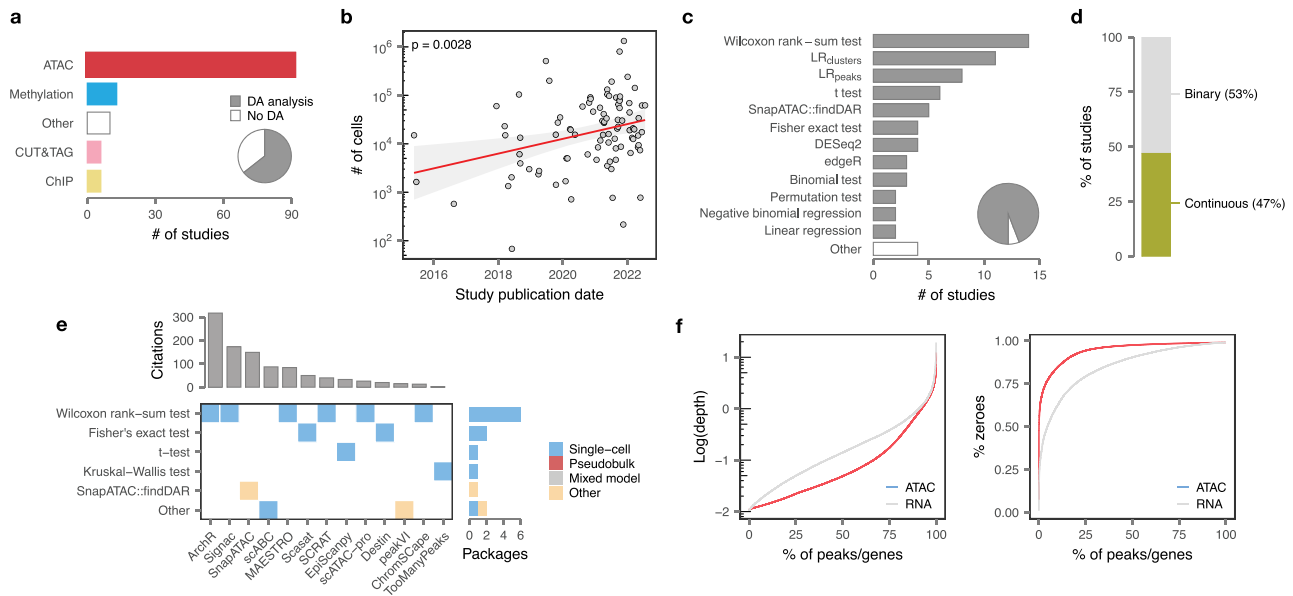


Fig. 1 | Landscape of DA analysis for single-cell epigenomics. **a** Experimental techniques used in 118 primary publications that reported single-cell epigenomic datasets. Inset pie chart shows the proportion of studies (64%) that reported a DA analysis. **b** Number of single cells profiled by scATAC-seq in 91 primary studies, shown as a function of publication date to highlight exponential scaling of scATAC-seq over time. Trend line and inset p-value, linear regression; shaded area, 95% confidence interval. **c** Statistical methods for DA analysis employed in 118 single-cell epigenomics papers. DA methods shown in grey will be considered in our analysis. “Other” includes four additional methods used in just a single study. Inset pie chart shows the total proportion of single-cell epigenomics papers (94%) that employed a DA analysis method considered in this Registered Report. **d** Proportions of single-

cell epigenomics studies that have treated the data as binary or continuous, respectively, during DA analysis. **e** Default statistical methods for DA analysis implemented in 13 single-cell analysis packages. Top, number of citations per package. Right, total number of analysis packages in which each DA analysis method is implemented as the default. **f** Cumulative distribution functions showing statistical properties of RNA-seq and ATAC-seq data from matching single cells. ATAC-seq features (peaks) are characterized by a lower average sequencing depth and a higher proportion of zeroes. Data is from a 10x multiome dataset from the mouse spinal cord (Methods). Source data are provided as a Source Data file (Source Data 1).

measures of genome accessibility (Fig. 1d). This lack of consensus is reflected in the variety of DA methods that are implemented by default within the most widely used analysis packages for scATAC-seq (Fig. 1e).

Our survey of the single-cell epigenomics literature highlighted that the most widely used statistical methods in this field are based on, or identical to, methods that were originally developed for scRNA-seq. The application of similar statistical methodologies contrasts with the different facets of biology measured by each technology. Whereas scRNA-seq measures expressed genes, scATAC-seq measures the accessibility of the entire genome, and accordingly, we identified dramatic differences in the statistical features of scATAC-seq compared to scRNA-seq data (Fig. 1f). In particular, scATAC-seq measures a larger number of features compared to scRNA-seq, and each of these features are quantified by fewer reads and in fewer cells⁴⁷. These biological and technological differences raise the possibility that the statistical methods used for DE analysis of scRNA-seq data may be ill-suited to DA analysis of scATAC-seq data.

Epistemological framework for biologically accurate DA analysis

If the most prevalent statistical methods are not optimised for DA analysis, then they may overlook biological differences, or conversely, could lead to spurious discoveries. These possibilities compelled us to conduct a comprehensive comparison of DA methods for scATAC-seq data. As a prerequisite for such a comparison, we first recognized the necessity of an epistemological framework that would capture the biological accuracy of these methods⁴⁸. Whereas the majority of benchmarks in computational biology rely on simulated datasets, we previously showed that these simulations fail to appreciate essential aspects of biological data generation, and therefore lead to unreliable conclusions⁴⁹. Instead, we showed that comparisons of statistical methods based on real datasets with experimental ground truth do capture biological differences in the performance of these methods. In the context of scRNA-seq data, we showed that a close approximation to the ground truth can be obtained from matched bulk and scRNA-seq performed on the same population of purified cells, exposed to the same perturbation, and sequenced in the same laboratory.

We hypothesized that a similar epistemological framework would allow us to establish the biological accuracy of statistical methods for DA analysis. To enable this framework, we identified a series of published datasets in which matched bulk and single-cell ATAC-seq were used to profile the same populations of purified cells within the same laboratories. These matched bulk datasets provide a mechanism to evaluate the biological accuracy of single-cell DA methods.

We also postulated that the development of single-cell multi-omic assays⁵⁰, in which the epigenome and transcriptome are profiled in the same individual cells, would provide the opportunity to extend this epistemological framework. Concretely, epigenomic measurements from multi-omic assays can be aggregated to the level of genes, and therefore, we reasoned that these assays would allow us to compare DA and DE within the same individual cells. The biological hypothesis underlying this experiment is that DE genes expressed across biological conditions are likely to have promoters that are DA within the same individual cells. Previous work has established that this assumption holds across the genome as a whole when DE and DA are measured systematically^{51–53}.

The third and final component of our epistemological framework is the recognition that the biological accuracy of any statistical method is contingent on its ability to avoid producing false discoveries⁴⁹. We therefore compared DA methods based on their ability to avoid false discoveries in the absence of any true biological differences.

Experiment 1: Evaluating single-cell DA methods with matched bulk data

We first used our assembled compendium of matched bulk and single-cell ATAC datasets to evaluate the biological accuracy of each single-cell DA method, using the bulk data as a reference. Our survey of the literature identified five studies in which matching single-cell and bulk epigenomics data were collected from the same populations of purified cells and sequenced within the same laboratory (Fig. 2a). These studies collected between two to four scATAC-seq libraries per condition (Supplementary Data 2). We performed DA analysis of both the bulk and single-cell ATAC-seq datasets, using each of the 11 DA methods that had been employed by at least two publications at the time of our literature review. We measured the concordance between single-cell and bulk DA analyses using the area under the concordance curve (AUCC)⁵⁴, as employed by previous benchmarks of DE methods for single-cell transcriptomics data^{49,55}.

In our primary analysis, we observed that most DA methods achieved comparable performance, with relatively small differences separating the ten top-performing methods (Fig. 2b). Among these, methods that aggregated cells within biological replicates to form so-called ‘pseudobulks’ consistently ranked near the top. In contrast, negative binomial regression and a previously described permutation test¹⁹ were outliers that achieved substantially lower concordance to the bulk data than other DA methods.

We conducted a series of sensitivity analyses to test the robustness of these observations. First, we found that the performance of single-cell DA methods was largely unchanged when applying different DA approaches to establish the experimental ground truth within the bulk data (Supplementary Fig. 1a, b). Moreover, we obtained broadly consistent results when varying the number of top-ranked DA peaks used to calculate the AUCC, or when filtering peaks that were not accessible in at least 5% of cells (Supplementary Fig. 1c). Because calculating the concordance between DA analyses of single-cell and bulk ATAC-seq requires a matching set of peaks to be defined in both datasets, we also explored the impact of varying this peak set. In all of the above analyses, we had called peaks in the matched bulk ATAC-seq data, but widely used software packages such as Signac³⁵, ArchR³³, and SnapATAC³⁴ instead call peaks in ‘pseudobulk’ samples created by pooling the single-cell data. We found that this procedure improved concordance between single-cell and bulk ATAC overall, but that the relative performance of single-cell DA methods remained similar (Supplementary Fig. 1d). To address the possibility that artefacts in peak calling might confound these results, we also devised a procedure to introduce spurious peaks into the peak sets, and found that the relative performance of the single-cell DA methods was largely robust to the presence of noise (Supplementary Fig. 1e). Finally, we evaluated concordance separately for peaks in promoter versus enhancer regions, and found that differences between DA methods were apparent primarily in the latter (Supplementary Fig. 1f).

Experiment 2: Evaluating single-cell DA methods with single-cell multi-omics

We next identified four studies that employed multi-omic assays to quantify both gene expression and chromatin accessibility across tens of thousands of nuclei, each profiling between two and 13 replicates (Supplementary Data 2), and leveraged these datasets to repeat the comparisons of single-cell DA methods described in Experiment 1, now using gene expression as the reference (Fig. 2c). To enable comparison between the ATAC and RNA modalities, we aggregated chromatin accessibility around promoters into gene-level activity scores, as implemented in a number of widely used software packages^{33,35,56}. We then tested for differences in chromatin accessibility at the gene level, using the DE results from the matching RNA modality to define the reference. Moreover, we performed GO enrichment analyses of the differentially accessible genes, and

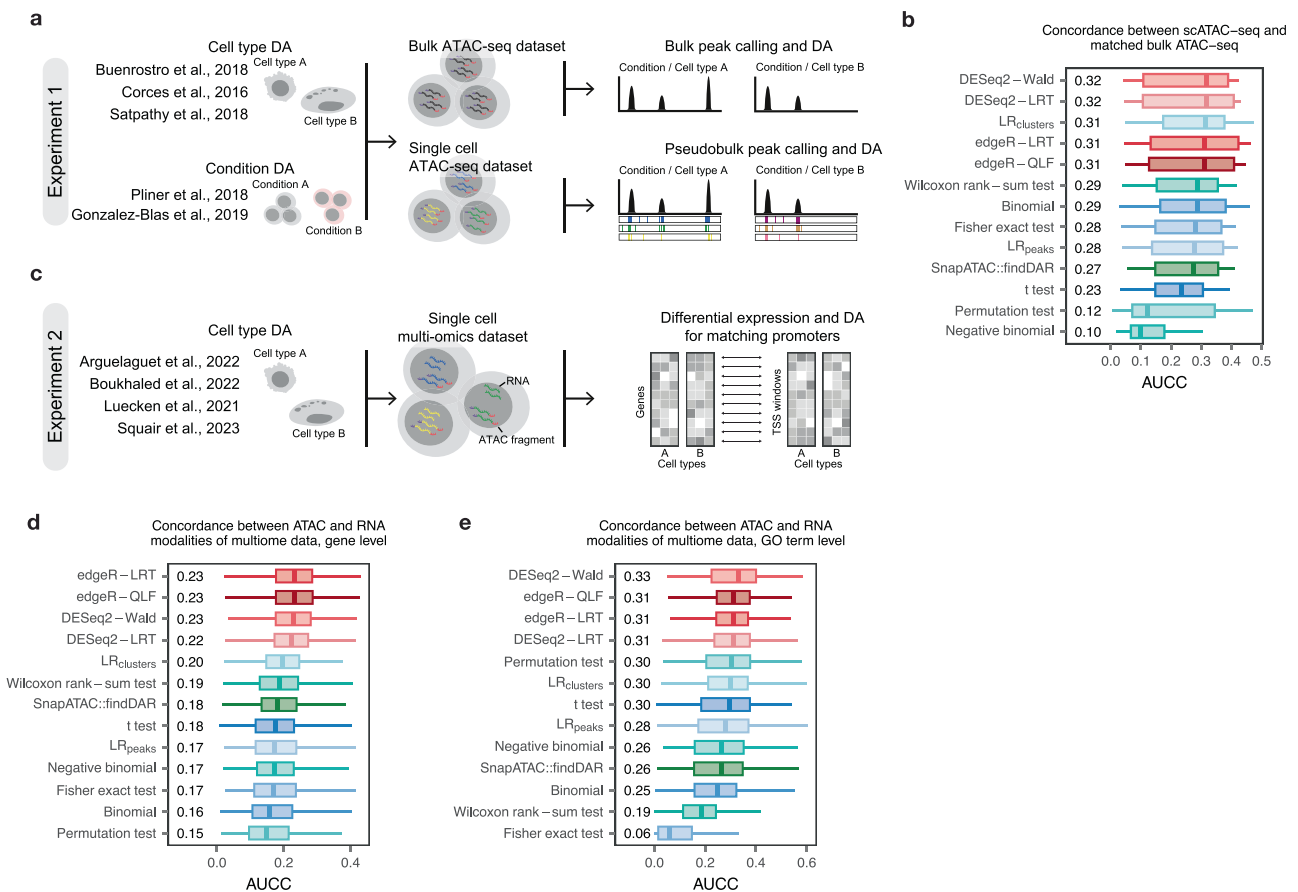


Fig. 2 | Evaluating single-cell DA methods with matched bulk data and single-cell multi-omics. **a** Design of Experiment 1. DA analysis was performed between cell types or conditions for single-cell datasets with matching bulk ATAC-seq as a reference. Peaks were called either in the bulk ATAC-seq data (primary analysis) or in pseudobulk single-cell data (sensitivity) analysis. **b** Area under the concordance curve (AUC) for single-cell DA methods in Experiment 1, using matching bulk ATAC-seq as a reference ($n = 16$ comparisons). Inset text shows the median AUC. Methods that aggregate counts within replicates to form ‘pseudobulks’ are shown in shades of red; one method that aggregates counts across replicates is shown in

green; and methods that do not aggregate information across cells are shown in blue. **c** Design of Experiment 2. DA analysis was performed between cell types using matched scRNA-seq data from the same cell as a reference, comparing DA of genomic intervals around the TSS to gene-level differential expression ($n = 306$ comparisons). **d** Area under the concordance curve (AUC) for single-cell DA methods in Experiment 2, using matching snRNA-seq as a reference. Inset text shows the median AUC. **e** As in **d** but showing concordance at the level of GO terms enriched among DA peaks. Source data are provided as a Source Data file (Source Data 2).

evaluated the concordance between GO enrichment analyses of the ATAC and RNA modalities.

In our primary gene-level analysis, we identified a more pronounced difference between DA methods that aggregated cell-level chromatin accessibility profiles into ‘pseudobulks’ and those that did not, as compared to Experiment 1 (Fig. 2d). The former outperformed the latter, although all pseudobulk methods achieved comparable performance to one another.

We conducted a series of sensitivity analysis to assess the robustness of this result. First, we verified that these results were not confounded by genes with overlapping promoter regions, since removing these genes led to slightly improved concordance but had little impact on the relative performance on single-cell DA methods (Supplementary Fig. 2a). We also confirmed that these results were largely unchanged when (i) applying different statistical approaches to establish the experimental ground truth within the RNA modality; (ii) varying the number of top-ranked DA genes used to calculate the AUC, or (iii) filtering genes whose promoters were not accessible in at least 1% of cells (Supplementary Fig. 2b–d). Finally, previous work has shown that inferences about DE are generally more accurate for highly expressed genes^{57,58}, whereas identifying instances of true DE among lowly-expressed genes can be challenging⁴⁹. These observations raised the possibility that

inaccurate inferences about DE for lowly-expressed genes might confound our analysis. Therefore, we re-evaluated the concordance after excluding the bottom tercile of lowly-expressed genes, and found that the relative performance of DA methods was essentially unchanged (Supplementary Fig. 2e).

In our primary Gene Ontology-level analysis, we again observed that the four top-performing methods all aggregated cells to form pseudobulks (Fig. 2e). We then repeated each of the above sensitivity analyses at the level of Gene Ontology terms. In general, the relative performance of single-cell DA methods was robust to the procedures used to identify the experimental ground truth within the RNA modality, varying the number of top-ranked GO terms used to calculate the AUC, or filtering genes whose promoters were not accessible in at least 1% of cells (Supplementary Fig. 3b–d). In contrast, the relative performance of single-cell DA methods was somewhat more sensitive to the removal of genes with overlapping promoters or by the removal of lowly-expressed genes in the RNA modality, both of which improved the performance of the t -test (Supplementary Fig. 3a, e).

Notably, in all twelve primary and sensitivity analyses at the gene and GO term levels, a pseudobulk DA method achieved the best performance. Moreover, no pseudobulk DA method ranked among the bottom half of lowest-performing DA methods in any of these analyses.

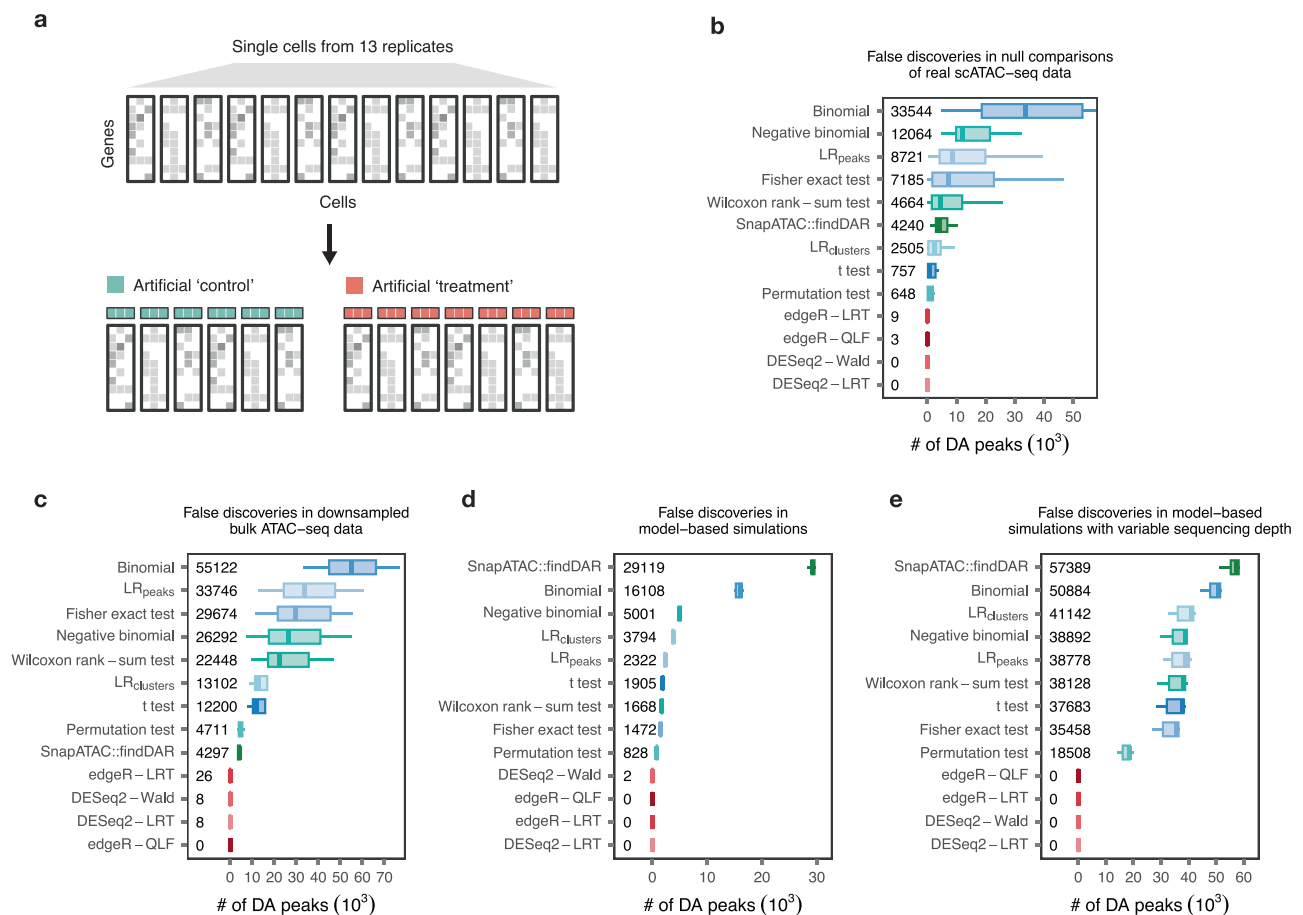


Fig. 3 | False discoveries in single-cell DA analysis. **a** Schematic overview of Experiment 3.1. Bone marrow mononuclear cells from healthy donors were profiled by Luecken et al.⁵⁹ in 13 independent replicates. For each cell type, half of these replicates were assigned to an artificial ‘control’ group, and the other half to an artificial ‘treatment’ group. DA analysis was then performed between cells from randomly assigned replicates. **b** Number of DA peaks detected between randomly assigned replicates at 5% FDR within each cell type in random comparisons of published scATAC-seq data ($n = 21$ comparisons). Inset text shows the median number of DA peaks per method. **c** As in **b** but showing the number of DA peaks

detected at 5% FDR in DA analysis of downsampled bulk ATAC-seq libraries without biological differences between experimental conditions. **d** As in **b** but showing the number of DA peaks detected at 5% FDR in model-based simulations of scATAC-seq data without biological differences between experimental conditions. **e** As in **b** but showing the number of DA peaks detected at 5% FDR in model-based simulations of scATAC-seq data without biological differences between experimental conditions, with variation in sequencing depth between libraries. Source data are provided as a Source Data file (Source Data 3).

Experiment 3: False discoveries in single-cell DA

The possibility of profiling hundreds of thousands of cells with single-cell epigenomics presents both opportunities and challenges. On one hand, the statistical power afforded by scATAC-seq data should enable the detection of subtle changes in chromatin accessibility between conditions or cell types. On the other hand, these subtle changes may represent technical artefacts or even false discoveries rather than true biological differences. We exposed similar opportunities and challenges in our recent analysis of DE methods for scRNA-seq, wherein we found that many of the most widely used methods can produce thousands of false discoveries in routine experiments⁴⁹. We investigated whether similar phenomena arise in scATAC-seq data. For this purpose, we conducted a series of analyses to assess the emergence of false discoveries in DA analyses of both real and simulated datasets.

We first investigated the emergence of false discoveries in random comparisons of cells from the same experimental condition. For this purpose, we repurposed a large scATAC-seq dataset⁵⁹ to create artificial comparisons between cells from identical experimental conditions, with the expectation that any regions that are called as DA reflect statistical false discoveries rather than instances of true DA (Fig. 3a). In our primary analysis, we found that several widely-used DA methods identified thousands of differentially accessible peaks in the absence of

any biological differences (Fig. 3b). Strikingly, the three most frequently used DA methods in our survey of the literature (Wilcoxon rank-sum test, LR_{clusters}, and LR_{peaks}; Fig. 1c) identified a median of 4664, 2505, and 8721 DA peaks within any given cell type. Conversely, methods that aggregated single-cell chromatin accessibility profiles to form pseudobulks never identified more than a median of 9 DA peaks in these random comparisons.

The premise of the above experiment is that any regions that are identified as DA between pairs of randomly assigned replicates are unlikely to represent biological differences. However, the nature of real-world data does not allow us to formally exclude the possibility that heterogeneity between replicates introduces true biological differences. Therefore, we complemented our re-analysis of published data with two simulation studies. In the first of these, we downsampled bulk ATAC-seq libraries to simulate scATAC-seq data with no biological cell-to-cell variation at all³⁴. In our primary analysis, we again observed that several widely-used DA methods produced thousands of false discoveries, but that pseudobulk methods were much less prone to false discoveries (Fig. 3c).

To achieve more precise control over the composition and properties of simulated scATAC-seq datasets, we leveraged a simulation framework originally developed for scRNA-seq data⁶⁰. We fit the

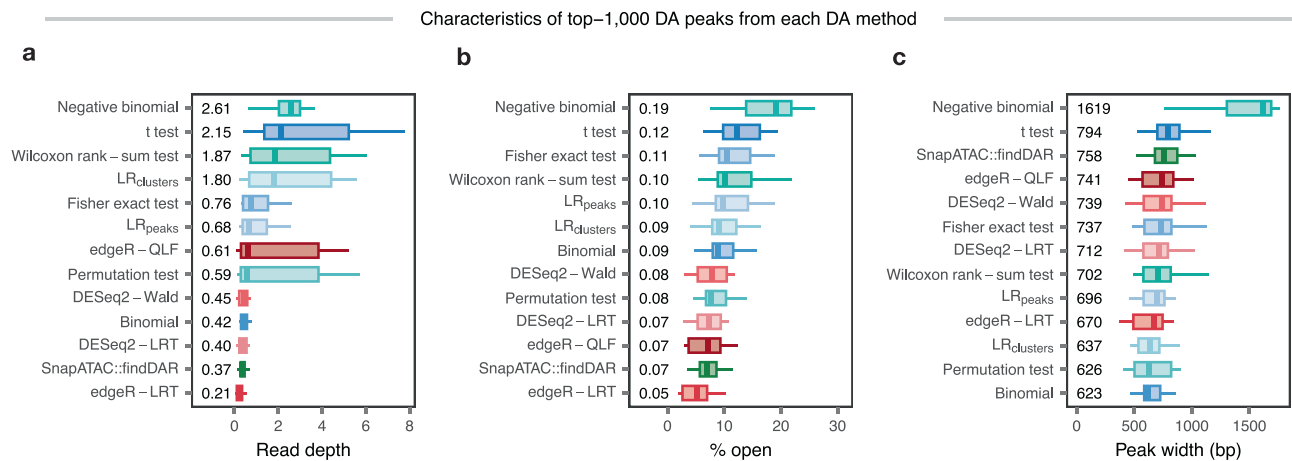


Fig. 4 | Biases in single-cell DA analysis. **a** Mean read depth of the top-1000 DA peaks identified by each single-cell DA method in published scATAC-seq datasets ($n = 16$ comparisons). Inset text shows the median across comparisons. **b** As in **a** but

showing the proportion of cells in which these peaks are open. **c** As in **a** but showing the width of each peak.

parameters of this simulation to scATAC-seq data, and used the resulting model to simulate datasets that varied in the number of cells and libraries sequenced and the degree of technical variation between libraries. In our primary analysis, we again identified thousands of false discoveries from widely-used DA methods (Fig. 3d). However, we observed differences in the specific DA methods that were most prone to producing false discoveries in this simulation setup, as compared to either null comparisons of real scATAC-seq data or the simulation above based on downsampling bulk ATAC-seq data. We cannot exclude that these differences reflect an artefact of this simulation framework.

In secondary analyses, we established that the number of false discoveries was exacerbated by sequencing a greater number of cells or increasing technical variation between libraries, whereas it was reduced by profiling a larger number of replicates (Supplementary Fig. 4a–c). These observations are consistent with the notion that accounting for biological and technical variation between replicates is critical to controlling the false discovery rate⁴⁹.

In our primary analysis, each replicate was simulated with the same sequencing depth, which is in contrast to the variable sequencing depths of replicates in real-world datasets. Therefore, we also carried out a secondary analysis whereby we deliberately simulated replicates with different sequencing depths (Supplementary Fig. 4d). This simulation markedly increased the number of false discoveries returned by most single-cell DA methods, although pseudobulk methods remained robust to false discoveries (Fig. 3e).

Together, these experiments emphasized that many widely used single-cell DA methods can produce thousands of false discoveries. Notably, these false discoveries were exaggerated when simulating more technically or biologically heterogeneous libraries. Conversely, DA methods that aggregated cells within replicates to form pseudobulks showed a markedly improved capacity to avoid false discoveries. We therefore also studied other strategies to account for variation between replicates in DA analysis. First, we explored the impact of including replicate as a fixed effect in generalized linear models. Second, we also considered replicate as a random effect by fitting negative binomial generalized linear mixed models (GLMMs)⁶¹. These experiments revealed that incorporating replicate as a fixed effect did not decrease, and sometimes increased, the number of false discoveries. However, incorporating replicate as a random effect enabled DA analysis without false discoveries (Supplementary Fig. 4e).

Experiment 4: Evaluating biases of single-cell DA methods

It is now well-established that many statistical methods for DE analysis of both bulk and single-cell RNA-seq data exhibit biases in the types of

genes that are preferentially identified as DE^{49,55,57,58,62}. However, it remains unclear whether similar biases may affect DA analyses of scATAC-seq data. We hypothesized that the same biases that affect scRNA-seq data could manifest in scATAC-seq data as a bias of DA methods to preferentially identify peaks that are open in a larger proportion of cells, peaks that are supported by a greater number of sequencing reads, or peaks that are wider. To address these possibilities, we conducted a series of experiments using both real and simulated data to address whether these biases affected DA analysis of scATAC-seq data.

We first characterized the properties of DA peaks in the same published scATAC-seq datasets examined in Experiment 1. To control for differences in the total number of peaks called as DA by each statistical method, we ranked peaks by their p-values and limited our analysis to the top-1000 DA peaks called by each method. In our primary analysis, we identified a number of differences in the characteristics of the peaks preferentially called as DA by each method. Methods that treated chromatin accessibility as a quantitative measurement (*t*-test, Wilcoxon rank-sum test, negative binomial regression, LR_{clusters}) preferentially called peaks supported by a greater number of reads, and open in a greater number of cells, as being DA (Fig. 4a, b). In contrast, methods that treated accessibility as a binary phenotype (Fisher's exact test, LR_{peaks}, binomial test, permutation test), as well as pseudobulk DA methods and SnapATAC's findDAR test, exhibited less bias towards highly accessible peaks. We observed less variability in the widths of peaks preferentially called as DA by each method, with the notable exception of negative binomial regression, which consistently called wider peaks as DA (Fig. 4c).

In secondary analyses, we varied the number of top-ranked DA peaks considered by characterizing the top-500 or top-5000 DA peaks called by each method. The trends observed for the top-1000 DA peaks were broadly conserved in this analysis (Supplementary Fig. 5).

Next, we sought to specifically characterize the peaks that were spuriously called as DA in Experiment 3. We binned peaks into deciles according to each of the properties studied above, and computed both the absolute number as well as the proportion of false discoveries arising from each decile. Across all three false discovery experiments, most DA methods exhibited biases towards peaks supported by a large number of reads, which were open in a greater proportion of cells, and broader peaks (Supplementary Figs. 6–8). In contrast, methods that aggregated cells to form pseudobulks exhibited a lesser degree of bias, when they identified any false discoveries at all. Unexpectedly, we found that SnapATAC's findDAR test exhibited an unusual pattern of bias,

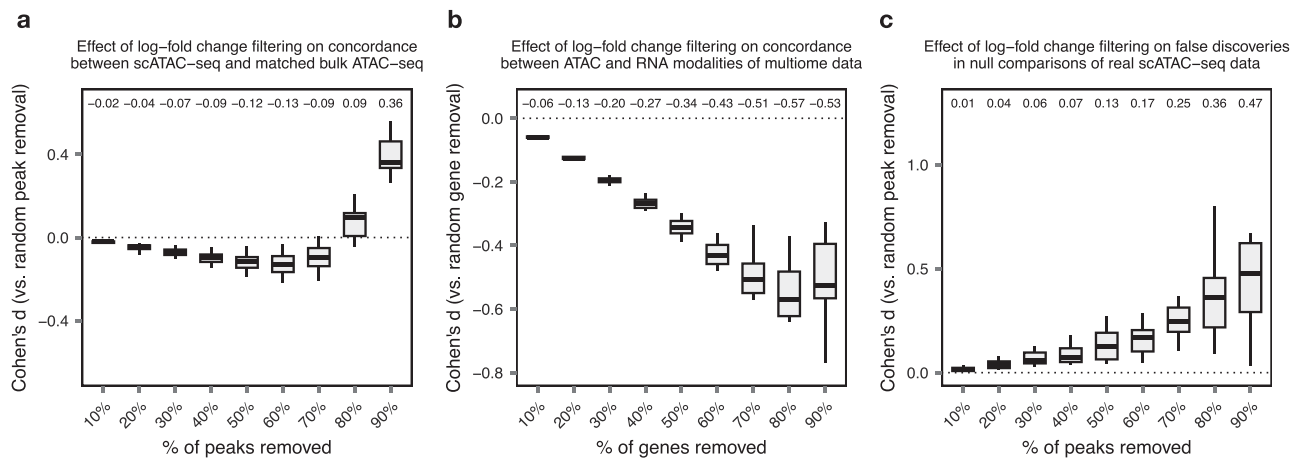


Fig. 5 | Impact of log-fold change filtering on single-cell DA analysis. **a** Effect size (Cohen's *d*) of increasingly stringent log-fold change filtering on the AUCC between single-cell and bulk ATAC-seq DA, relative to the removal of an equivalent number of peaks selected at random ($n = 16$ comparisons). Inset text shows the median Cohen's *d*. **b** As in **a** but for the AUCC between the ATAC and RNA modalities of single-cell multi-omics data ($n = 306$ comparisons). **c** As in **a** but for the number of false discoveries in null comparisons of published scATAC-seq data ($n = 21$ comparisons).

whereby false discoveries preferentially arose from genes with intermediate expression.

Experiment 5: Impact of log-fold change filtering

Differential analyses of ATAC-seq data may discard differentially accessible regions with a log-fold change below an arbitrary threshold, on the grounds that regions with small fold changes are unlikely to be biologically relevant⁴⁴. Similar practices are ubiquitous in the analysis of other sequencing-based technologies, such as RNA-seq, where the choice of log-fold change threshold may alter the biological interpretation of a given experiment⁶³. These observations motivated an empirical assessment of the impact of log-fold change filtering on single-cell DA analysis.

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

We first examined the concordance of DA between scATAC-seq and matching bulk ATAC-seq. As expected, increasingly stringent log-fold change filters increased concordance for all DA methods. However, the same effect was also seen when removing equivalent numbers of random peaks (Supplementary Fig. 9). When controlling for random peak filtering, we observed that the concordance initially decreased when filtering up to 70% of peaks based on log-fold change, and then increased when filtering 80% or more of peaks. To summarize these trends, we visualized the effect size of log-fold change filtering across all DA methods, compared to random peak filtering (Fig. 5a).

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

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

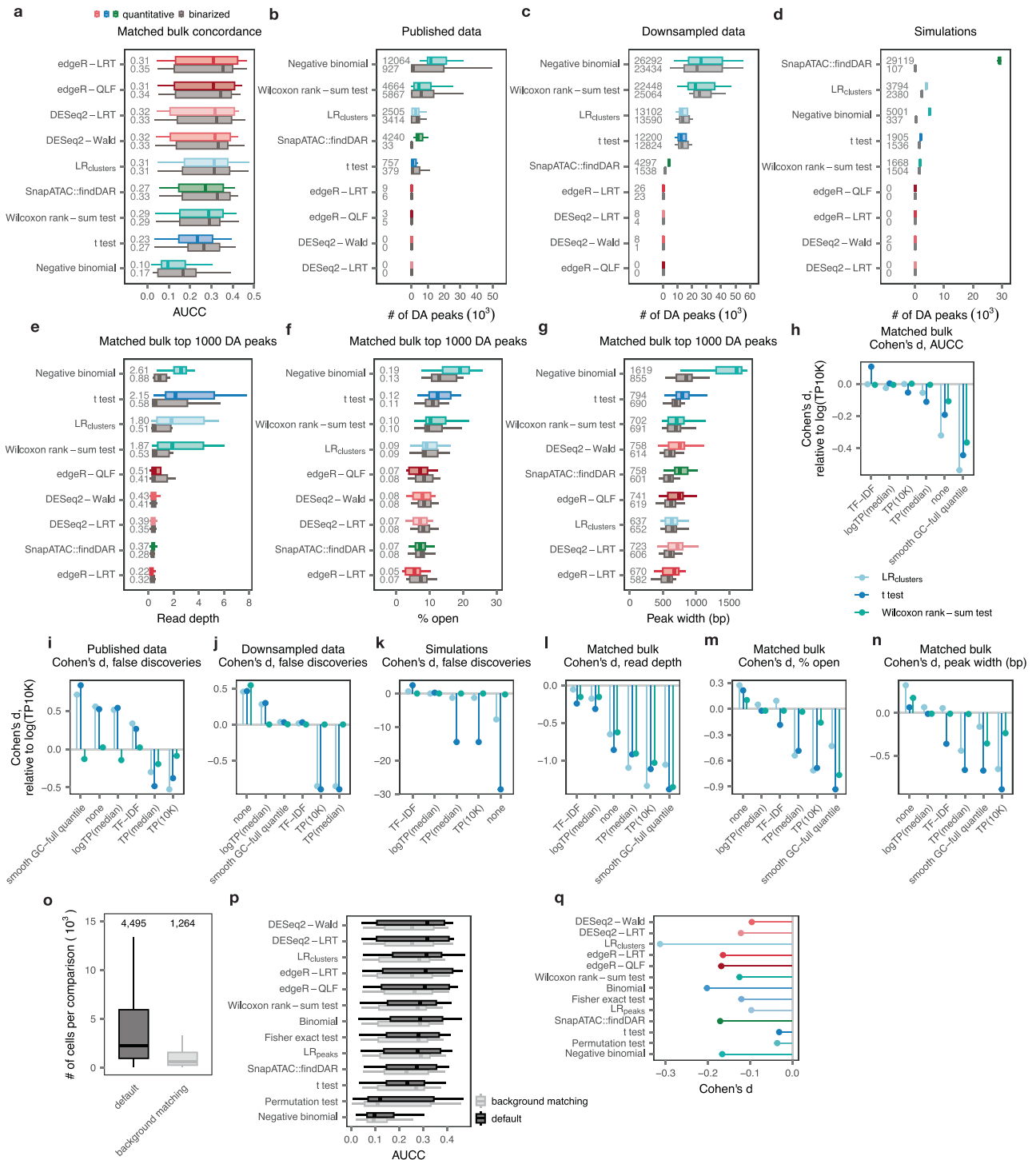
Experiment 6: Best practices for scATAC-seq analysis

Our survey of the literature identified substantial discordance not just in the methods used for DA analysis, but even in the representations of scATAC-seq data that are provided as input to these methods (Fig. 1d). Indeed, fundamental concepts in the representation of scATAC-seq data remain subjects of ongoing debate. Perhaps the most notable among such concepts is whether to treat scATAC-seq data as a qualitative measurement by binarizing genome accessibility^{31,32}. We asked whether the setting of DA could provide an opportunity to address this question. Specifically, we hypothesized that the accuracy of DA analysis could be used as a litmus test, in that approaches to scATAC-seq data preprocessing that enable more accurate and robust DA analysis are likely to also provide more accurate results at other steps in the analytical workflow. Accordingly, we repeated a subset of the analyses described above using binarized representations of the scATAC-seq datasets, with the aim of characterizing the effect of binarization on the biological accuracy, false discovery control, and biases of DA methods for scATAC-seq data.

We first tested whether binarizing scATAC-seq data would improve the biological accuracy of DA analysis, as quantified by the concordance to matching bulk ATAC-seq. Surprisingly, we found that binarizing scATAC-seq data generally increased the concordance between scATAC-seq and bulk ATAC-seq data, and the magnitude of this improvement was largest for the worst-performing DA methods (Fig. 6a and Supplementary Fig. 12a). The relative performance of DA methods was largely unchanged by binarization, with the exception of SnapATAC's findDAR test, which improved to rank among the top-performing methods when applied to binarized data.

We next asked whether binarizing scATAC-seq data could mitigate the appearance of false discoveries. This analysis yielded ambiguous results (Fig. 6b–d and Supplementary Fig. 12b–d). A subset of DA methods, including SnapATAC's findDAR test and negative binomial regression, consistently produced fewer false discoveries when applied to binarized data, although the magnitude of this decrease varied markedly across experiments. For the remaining DA methods, binarization did not have a consistent effect on the number of false discoveries.

We then investigated whether binarization modulated the biases of DA methods. Interestingly, in published scATAC-seq data, binarization attenuated the biases of some DA methods towards calling highly accessible peaks as DA (Fig. 6e–f and Supplementary Fig. 13a, b), although this attenuation was less apparent for methods that



aggregated cells to form pseudobulks, since these biases were not apparent to begin with. Moreover, we observed that, for virtually every DA method, binarizing scATAC-seq data led narrower peaks to be called as DA (Fig. 6g and Supplementary Fig. 13c).

We additionally sought to specifically characterize the peaks that were spuriously called as DA before and after binarization, but found that binarization did not alter the more general tendency for false discoveries to preferentially arise from more accessible and wider peaks (Supplementary Figs. 14–16).

Collectively, these results suggest that binarization can improve the biological accuracy of DA analysis, and this effect is observed for virtually every DA method included in our analysis. This observation is

at odds with the argument that binarization is an unnecessary step that discards quantitative data collected in scATAC-seq experiments^{31,32}. We suggest that this apparent paradox can be reconciled by the observation that binarization generally reduces the biases of DA methods towards more accessible and wider peaks, even for DA methods that are less affected by these biases in the first place, and that mitigating these biases can outweigh the potential negative effects of binarization.

Another important open question in the analysis of scATAC-seq data is whether and how this data should be normalized prior to DA analysis. The majority of the DA methods considered in this study operate directly on count matrices or binarized data. However, three

Fig. 6 | Best practices for scATAC-seq analysis. **a** Area under the concordance curve (AUCC) for single-cell DA methods using matching bulk ATAC-seq as a reference, before and after binarization ($n = 16$ comparisons). Inset text shows the median AUCC. **b** Number of DA peaks detected between randomly assigned replicates at 5% FDR in random comparisons of published scATAC-seq data, before and after binarization ($n = 21$ comparisons). Inset text shows the median number of DA peaks. **c** As in **b** but in downsampled bulk ATAC-seq libraries. **d** As in **b** but in model-based simulations of scATAC-seq data. **e** Mean read depth of the top-1000 DA peaks identified by each single-cell DA method in published scATAC-seq datasets, before and after binarization. Inset text shows the median. **f** As in **e** but showing the proportion of cells in which these peaks are open. **g** As in **e** but showing the width of each peak. **h** Effect size (Cohen's d) of alternative approaches to normalization of scATAC-seq data on the AUCC between single-cell and bulk ATAC-seq DA, relative to log-TP10K normalization ($n = 16$ comparisons). **i** As in **h** but

showing the number of DA peaks detected between randomly assigned replicates at 5% FDR in randomized comparisons of published scATAC-seq data ($n = 21$ comparisons). **j** As in **i** but in downsampled bulk ATAC-seq libraries. **k** As in **i** but in model-based simulations of scATAC-seq data. **l** As in **h** but showing the mean read depth of the top-1000 DA peaks identified by each single-cell DA method in published scATAC-seq datasets. **m** As in **l** but showing the proportion of cells in which these peaks are open. **n** As in **l** but showing the width of each peak. **o** Number of cells considered per comparison, before and after controlling for technical covariates using the ArchR background-matching procedure ($n = 322$ comparisons). Inset text shows the median. **p** Area under the concordance curve (AUCC) for single-cell DA methods using matching bulk ATAC-seq as a reference, before and after controlling for technical covariates using the ArchR background-matching procedure ($n = 16$ comparisons). **q** Effect size (Cohen's d) of the ArchR background-matching procedure on the AUCC between single-cell and bulk ATAC-seq.

methods (t -test, Wilcoxon rank-sum test, and LR_{clusters}) are generally applied to normalized data, with the expectation that normalization is required to produce biologically accurate results. We therefore surveyed the literature to identify the approaches to normalization that are most commonly applied to scATAC-seq data, and then evaluated the impact of normalization on the performance of these three DA methods.

We first tested the impact of normalization on the biological accuracy of DA analysis, as quantified by the concordance to matching bulk ATAC-seq. In this analysis, no approach consistently improved over log-TP10K normalization (Fig. 6h and Supplementary Fig. 17a). The lone exception was the combination of TF-IDF normalization and the t -test, which demonstrated improved concordance relative to log-TP10K normalization.

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

We further studied the relationship between normalization and the biases of DA methods. Interestingly, this analysis showed that numerous approaches to normalization can reduce the biases of DA methods towards highly accessible or broad peaks, relative to log-TP10K normalization (Fig. 6l–n and Supplementary Fig. 18). We additionally sought to specifically characterize the peaks that were spuriously called as DA under different normalization strategies, but this analysis failed to identify trends that were consistent across both DA methods and normalization strategies, and no method altered the global pattern whereby false discoveries preferentially arose from more accessible and wider peaks (Supplementary Figs. 19–26).

Collectively, these results do not provide strong support for any alternative approach to normalization as compared to log-TP10K, which is the default approach implemented in several of the most widely used analysis packages^{33,35} for DA analysis. This conclusion is in line with a more systematic benchmark of normalization approaches to scRNA-seq data, which also highlighted the strong performance of log-TP10K normalization⁶⁴.

Last, we evaluated a procedure proposed by the authors of ArchR to control for technical artefacts in DA analysis. For any given set of 'foreground' cells (e.g., cells of a particular type), ArchR constructs a set of 'background' cells of equal size that are matched according to a series of technical properties. By default, TSS enrichment and $\log_{10}(\#$ of fragments) are used to select a matching set of 'background' cells.

This approach inherently entails a trade-off whereby correcting for potentially confounding artefacts comes at the cost of using only a subset of the available data for analysis. Indeed, across all comparisons, we found that the ArchR background matching procedure substantially reduced the number of cells considered in any given DA analysis, from a mean of 4495 to 1264 cells per comparison (Fig. 6o). This decrease in the number of cells considered appeared to outweigh the positive impacts of controlling for technical variation, since we identified a uniform decrease in the AUCC across all DA methods considered (Fig. 6p–q).

Experiment 7: Data requirements for single-cell DA analysis

The scale of scATAC-seq datasets is increasing exponentially (Fig. 1b). On one hand, the availability of hundreds of thousands of cells per dataset could increase the statistical power of DA analysis. On the other hand, past work in single-cell transcriptomics⁶⁵ reported that the number of cell types identified per study was closely linked to the total number of individual cells sequenced in that study. This observation suggests that the increased statistical power afforded by a greater number of cells could be offset by an increased granularity of DA comparisons, each leveraging proportionally fewer cells. Moreover, many of the largest datasets are very sparse: for instance, the single-cell map of chromatin accessibility across 30 tissues reported by Zhang et al.²⁶ demonstrated a median of just ~2800 fragments per cell. These trends raise the question of what minimum sequencing depth or number of cells are required for accurate DA analysis.

To address these questions, we performed downsampling analyses of the sequencing depth or number of cells per dataset. We first studied the effect of sequencing depth on the biological accuracy of DA analysis by downsampling the plate-based datasets used in Experiment 1 to a mean of 500, 1000, 2000, 5000, or 10,000 counts per cell. We quantified the effect of downsampling on the AUCC by calculating Cohen's d relative to the DA analysis on 10,000 counts per cell. We identified a decrease in the concordance between single-cell and bulk DA in each case, as expected, with no evidence of saturation, suggesting that DA analysis continues to benefit from deeper sequencing up to at least 10,000 fragments per cell (Fig. 7a and Supplementary Fig. 27a).

We next investigated the minimum number of cells required for accurate DA analysis. To this end, we downsampled the number of cells within the droplet-based datasets used in Experiment 2. We limited our analysis to comparisons in which at least 1000 cells were sequenced in each experimental condition, and quantified the effect of downsampling on the AUCC by calculating Cohen's d relative to the DA analysis on 1000 cells per condition. In this analysis, the AUCC began to saturate when exceeding 300 cells per condition, although the degree and rate of saturation varied across individual DA methods (Fig. 7b and Supplementary Fig. 27b). These findings suggest that accurate single-cell DA analysis can be achieved with as few as 50–100

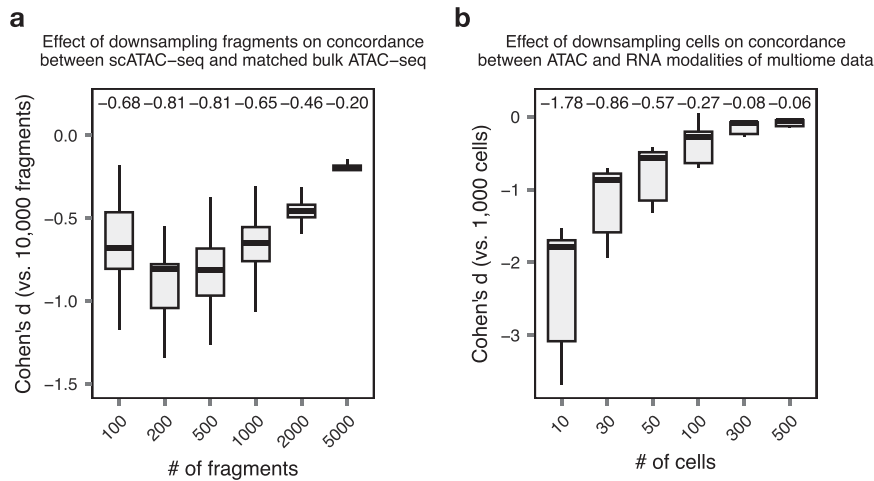


Fig. 7 | Data requirements for single-cell DA analysis. **a** Effect size (Cohen's *d*) of downsampling plate-based scATAC-seq data to a mean of 500, 1000, 2000, 5000 counts per cell on the AUCC for single-cell DA methods using matching bulk ATAC-seq as a reference, relative to DA analysis of the same datasets with a mean of 10,000 counts per cell ($n = 16$ comparisons). Inset text shows the median Cohen's *d*.

b Effect size (Cohen's *d*) of downsampling single-cell multi-omics data to 20, 50, 100, 200, or 500 cells per condition on the AUCC between the ATAC and RNA modalities, relative to DA analysis of the same datasets with 1000 cells per condition ($n = 306$ comparisons). Inset text shows the median Cohen's *d*.

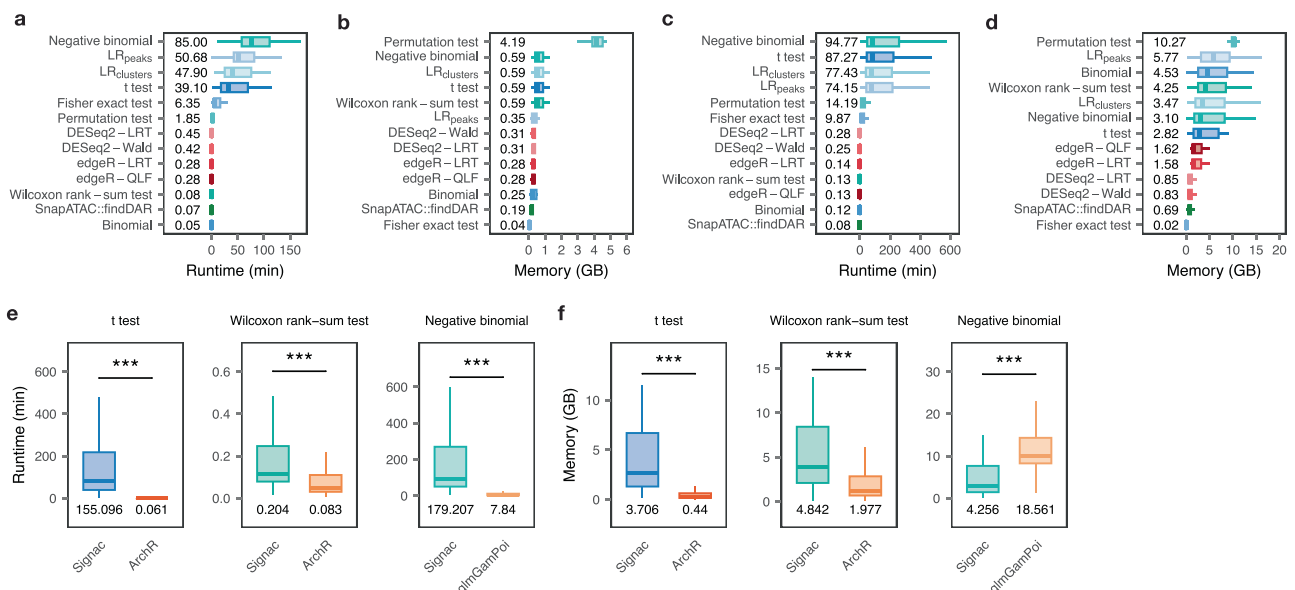


Fig. 8 | Scalability of single-cell DA methods. **a** Wall clock time required by each DA method to execute each comparison in Experiment 1 ($n = 16$ comparisons). Inset text shows the median runtime in minutes. **b** Peak memory usage of each DA method while executing each comparison in Experiment 1 ($n = 16$ comparisons). Inset text shows the median peak memory usage in GB. **c** As in **a** but for each comparison in Experiment 2 ($n = 306$ comparisons). **d** As in **b** but for each

comparison in Experiment 2 ($n = 306$ comparisons). **e** Wall clock time required by alternative implementations of three DA methods (t-test, Wilcoxon rank-sum test, and negative binomial regression). Inset text shows the median runtime in minutes. $***p < 10^{-15}$, two-sided paired *t*-test. **f** As in **e** but showing peak memory usage by each alternative implementation. Inset text shows the median peak memory usage in GB. $***p < 10^{-15}$, two-sided paired *t*-test.

cells, and that sequencing more than 300 cells per condition yields diminishing returns.

Experiment 8: Scalability of single-cell DA methods

The scale of single-cell epigenomics is continuing to increase. This observation underscores the need to evaluate the computational scalability of methods for DA analysis. Accordingly, we measured both the wall time and peak memory usage of each single-cell DA method in Experiments 1 and 2. We observed that the single-cell DA methods compared here varied by several orders of magnitude in both runtime

and memory usage (Fig. 8a-d). Negative binomial regression, both variants of logistic regression, and the t-test typically required more than an hour to perform a single comparison within multiome datasets. Conversely, pseudobulk DA methods and SnapATAC's findDAR test were consistently among the most time- and memory-efficient methods, suggesting these methods could most readily scale to datasets comprising millions of cells. Moreover, whereas the computational requirements of all DA methods scaled with the number of cells in the dataset, this relationship was attenuated for these same five methods (Supplementary Fig. 28).



Fig. 9 | Summary of DA method performance across major evaluation criteria. Methods were grouped into tertiles of high, low, or intermediate performance on the basis of their quantitative performance on each task, as described in the Methods, and ranked by their average performance across all criteria.

For some of the DA methods considered in our analysis, different implementations are provided in widely-used software packages, some of which have been reported to offer speed-ups of several orders of magnitude⁶⁶. These methods include the Wilcoxon rank-sum test (Signac vs. ArchR), the *t*-test (Signac vs. ArchR), and negative binomial regression (Signac vs. glmGamPoi⁶⁷). For these three DA methods, we tested both available implementations. We observed that optimized implementations of these methods can markedly reduce the computational resources required to execute any given DA analysis (Fig. 8e, f).

Performance of DA methods across experiments

To summarize our findings, we divided DA methods into top-, middle-, and bottom-performing tertiles on each task that involved a quantitative comparison of DA methods (Fig. 9). This visualisation reinforced the excellent performance of pseudobulk DA methods across all of the experiments in this study. We suggest that these methods should be considered a first choice approach for the DA analysis of scATAC-seq data. We did not include mixed models in our primary analyses because they had not been used in a published study at the time of our literature review; however, we hypothesize that these may also offer improved performance relative to the methods that are most widely used at present (Supplementary Fig. 4e), albeit at the cost of substantially increased runtime and memory requirements⁴⁹.

Exploratory analyses

During the peer review of this manuscript, the reviewers suggested additional analyses that had not been preregistered. In this section, we present the results of these *post hoc*, unregistered analyses.

First, we found that the concordance between single-cell and bulk ATAC-seq data was substantially higher for peaks located in promoter regions than for those located in enhancers (Supplementary Fig. 1f). Whereas our data does not allow us to establish a definitive causal explanation for this phenomenon, we found that peaks in promoter regions tended to be open in a greater number of cells, and supported by a greater number of reads, as compared to promoter regions (Supplementary Fig. 29). These observations are congruent with observations made in both bulk and single-cell ATAC-seq data^{2,32} and likely explain, at least in part, the greater biological accuracy of DA analysis for peaks in promoter regions. When considered in combination with the observation that the accuracy of DA analysis continues to benefit with increased sequencing depth up to at least 10,000 fragments per cell (Fig. 7a), these observations raise the possibility that increased sequencing depth would specifically benefit DA analysis of enhancers, although this possibility was not directly tested in the present study.

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

Third, the design of Experiments 1, 2, and 3 was preregistered before the results of Experiment 6 became available. In view of the observation that the performance of some DA methods benefited from binarization and/or alternative approaches to normalization, the question arose as to how these preprocessing strategies would affect the summary of DA methods shown in Fig. 9. A similar question arose with respect to the observation that the concordance between single-cell and matched bulk ATAC-seq was generally higher when calling peaks in performance in ‘pseudobulks’ created from single-cell data. To explore these points, we produced additional summary figures that compared DA methods across all primary and secondary analyses in Experiments 1 and 2 (Supplementary Fig. 33) and across all binarization and normalization scenarios in Experiment 6 (Supplementary Fig. 34). These analyses corroborated the observation that some DA methods consistently achieved very good or very poor performance: for instance, in Experiments 1 and 2, DESeq2-LRT ranked among the top tertile of DA methods in all but two analyses, whereas negative binomial regression, the binomial test, and Fisher’s exact test never ranked among the top tertile in any analysis. Considering all binarization and normalization approaches yielded rankings of DA methods that were broadly consistent with those shown in Fig. 9, but with some important differences: notably, SnapATAC::findDAR emerged as a top-performing DA method with respect to concordance with matched bulk ATAC-seq data, albeit not control of the false discovery rate, when applied to binarized data.

Discussion

The increasingly broad adoption of technologies to interrogate the epigenome at single-cell resolution has exposed a lack of consensus on

how to analyse the resulting data. In this study, we formulated and implemented a series of preregistered analyses that aimed to systematically compare statistical methods for DA analysis of epigenomics data. We carried out an extensive review of the literature to identify all of the DA methods that had been applied in published single-cell studies, and established an epistemological framework that would enable a comparison of these methods with respect to their biological accuracy and their propensity to produce false discoveries. These analyses established that statistical methods that aggregated individual cells to form ‘pseudobulks’ generally yielded DA results that were better-aligned with orthogonal measurements of the same biological systems (i.e., matching bulk ATAC-seq or matching RNA-seq from the same cells) than methods that treated individual cells as the biologically relevant unit of observation. Our data further suggested that these differences in performance could be attributed, at least in part, to the appearance of false discoveries in DA analysis of individual cells. In null comparisons of published scATAC-seq data, and in a series of simulation studies, we found that the most widely used DA methods identified thousands of DA peaks in the absence of any underlying biological differences. These false discoveries preferentially arose from wider peaks and peaks accessible in a greater number of cells, mirroring the tendency for false discoveries to arise from highly-expressed genes in single-cell data⁴⁹, and were abrogated both by pseudobulk DA methods and by mixed-effects models that incorporated replicate as a random effect. Remarkably, models that made identical assumptions about the underlying distribution of the count data varied markedly in their ability to avoid false discoveries, depending on whether and how they controlled for biological and technical variation across replicates. Together, these findings suggest that single-cell DA methods must account for technical or biological variation between replicates in order to enable accurate DA analysis and avoid a proliferation of false discoveries.

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

Beyond comparisons of individual methods for single-cell DA analysis, we sought to use DA as a litmus test to identify best practices for the analysis of single-cell epigenomics data more generally. In some cases, these analyses supported clear recommendations. For instance, we found that filtering peaks by their log-fold change between conditions did not consistently improve DA analysis, relative to removing an equal number of peaks chosen at random. Moreover, in comparisons of approaches to the normalization of scATAC-seq data, we did not identify any method that consistently improved performance relative to log-TPI0K normalization. We additionally observed a higher degree of concordance between DA analyses of matched bulk and single-cell ATAC-seq datasets when analyzing peaks called in ‘pseudobulk’ samples created by pooling the single-cell data, which supports the widespread application of this approach to the analysis of scATAC-seq data. In other cases, our results were more ambiguous. Notably, we found that binarization reduced the biases of single-cell DA methods towards peaks accessible in a greater number of cells and to broader peaks, and generally increased the concordance between DA analyses of single-cell and matched bulk ATAC-seq data. Conversely, we found that the binarization did not always decrease (and

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

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

Our study also has a number of limitations. First, our comparisons to parallel bulk ATAC-seq or multiome data leveraged high-throughput assays carried out on matching samples (or even cells) under identical experimental conditions, but a limitation of these analyses is the requirement of the use of statistical methods for DA and DE analysis in these parallel data sets. We carried out sensitivity analyses to test the impact of leaving out any given bulk DA or DE method, which supported the robustness of our conclusions. However, we cannot formally exclude the possibility that these comparisons favor single-cell DA methods that yield output more similar to that of other methods in general, for technical rather than biological reasons. Second, our comparison to multiome data was based on the premise that genes that are DE across biological conditions will also tend to have DA promoters, but exceptions to this assumption are known, notably during differentiation⁶⁸. The majority of the comparisons that we propose entail the comparison of two fully differentiated cell types, mitigating this concern to some degree. Third, our approach to quantifying gene-level accessibility involved aggregating reads around the transcription start site, which required the specification of an arbitrary window within which reads were considered to be associated with promoter accessibility. Fourth, the relative paucity of datasets with paired bulk ATAC-seq data from identical biological conditions means that the datasets used in Experiment 1 comparisons involved comparisons both between and within cell types. However, these are comparisons with different underlying biological motivations, and in particular, the effect sizes of comparisons between cell types are likely to be larger than those within cell types and across conditions.

We provide a common interface to all single-cell DA methods described here by updating our existing R package, *Libra* (available from GitHub at <http://github.com/neurorestore/Libra>). We provide a vignette walking users through the analysis of an example dataset, including easy-to-adapt code that we hope will empower users with limited computational experience and resources to perform accurate DA analysis.

Methods

Single-cell DA methods

We selected eleven single-cell DA methods to compare in this Registered Report, on the basis that each of these methods had been employed by at least two independent publications at the time of our literature review (Fig. 1c). The complete list of DA methods that we studied is as follows:

1. *t*-test
2. Wilcoxon rank-sum test
3. Logistic regression, using cell type or condition as the independent variable (LR_{clusters})
4. Logistic regression, using binary peak accessibility as the independent variable (LR_{peaks})
5. Negative binomial regression
6. Fisher's exact test
7. Binomial test
8. Permutation test
9. SnapATAC::findDAR
10. DESeq2 (Wald test and likelihood ratio test)
11. edgeR (quasi-likelihood *F*-test and likelihood ratio test)

For the *t*-test, Wilcoxon rank-sum test, negative binomial regression, and LR_{clusters} , implementations from Seurat (version 3.1.5)³⁵ were used. The findDAR method was drawn from the SnapATAC package (version 2.0). For LR_{peaks} , Fisher's exact test, binomial test, and permutation test, we developed optimized implementations for sparse single-cell matrices. Finally, for DESeq2 and edgeR, we used the implementations from the Libra package⁴⁹.

We excluded computational tools that do not perform statistical comparisons of individual genomic regions. For instance, we excluded SCENIC, which performs topic modeling of peak matrices to simultaneously cluster peaks and cells^{69,70}.

Evaluating single-cell DA methods with matched bulk data

Our survey of the literature identified five publications in which matching single-cell and bulk epigenomics data was collected from the same population of purified cells and sequenced within the same laboratory. The list of datasets incorporated in Experiment 1 was as follows:

1. Corces et al.⁷¹ collected single-cell and bulk ATAC-seq data from immunophenotypically purified leukemia stem cells (LSCs) and leukemic blast cells (blasts) from donors with acute myeloid leukemia. DA analysis was performed between LSCs and blasts.
2. Buenrostro et al.⁷² collected single-cell and bulk ATAC-seq data from immunophenotypically defined human hematopoietic cell types. For five of these cell types (common myeloid progenitor, CMP; granulocyte-macrophage progenitor, GMP; hematopoietic stem cell, HSC; megakaryocytic-erythroid progenitor, MEP; multipotent progenitor, MPP), at least two independent replicates were collected for both bulk and single-cell data. DA analysis was performed between each pair of cell types.
3. Pliner et al.⁵⁶ collected single-cell and bulk ATAC-seq data from human skeletal muscle myoblasts, before (0 h) and after (72 h) differentiation to myotubes. DA analysis was performed between myoblasts and myotubes.
4. Satpathy et al.⁷³ collected single-cell and bulk ATAC-seq data from immunophenotypically defined CD4⁺ naive and memory T cell subtypes. Among these cell types, naive CD4⁺ T cells and T_H17 CD4⁺ T cells were profiled by both single-cell and bulk ATAC-seq in at least two independent replicates. DA analysis was performed between naive and T_H17 CD4⁺ T cells.
5. Gonzalez-Blas et al.⁶⁹ collected single-cell and bulk ATAC-seq data from melanoma cell lines before (0 h) and 24, 48, or 72 h after siRNA knockdown (KD) of SOX10. DA analysis was performed between SOX10 KD cells at each timepoint and control cells.

FASTQ files were downloaded from GEO (see Data availability statement for accessions) and demultiplexed using custom Python scripts to confirm their availability for the proposed analyses. Demultiplexed FASTQ files for both the single-cell and bulk datasets were trimmed using Trim Galore (version 0.6.6) and aligned to the genome using bwa (version 0.7.17)⁷⁴, as implemented within SnapATAC³⁴. Peaks were called in the bulk data with MACS2 (version 2.2.6)⁷⁵, and the resulting peak set was used to obtain peak count matrices for both the bulk and single-cell data using SnapATAC.

DA analysis of the bulk ATAC-seq data was performed using three widely-used statistical approaches to the analysis of bulk sequencing data, including DESeq2 (version 1.38.1)⁷⁶, edgeR (version 3.40.0)⁷⁷, and limma (version 3.54.0)⁷⁸. Each of these methods was applied using two distinct statistical approaches implemented within the relevant package (i.e. Wald test vs. likelihood ratio test in DESeq2; quasi-likelihood *F*-test vs. likelihood ratio test in edgeR; and trend vs. voom in limma), and the concordance was averaged over all six bulk DA methods. Our expectation was that by considering several different approaches to bulk DA analysis, we would be able to mitigate the degree to which our conclusions are dependent on the results of any individual bulk DA method. However, to confirm the robustness of our conclusions to the inclusion of any individual bulk DA method, we also performed secondary analyses in which we removed individual DA methods from the bulk analysis and re-computed the concordance between single-cell and bulk DA.

Concordance between single-cell and bulk DA analyses was quantified using the area under the concordance curve (AUCC)⁵⁴, as employed by previous studies of DE methods for single-cell transcriptomics data^{49,55}. This metric requires the specification of a parameter *k* that determines the number of top-ranked peaks to be compared in the calculation of the AUCC, on the basis of the *p*-values assigned by each DA method. By default, we set the value of *k* to 5000; we selected this threshold based on our previous benchmark of DE methods for single-cell transcriptomics, which used *k* = 500 to calculate the AUCC, on the basis that there are typically about an order of magnitude more peaks in scATAC-seq count matrices than there are genes in snRNA-seq count matrices. However, we also evaluated the sensitivity of our conclusions to the value of this parameter by repeating the concordance analysis with *k* = 1000. Moreover, because previous benchmarks of DE analysis in single-cell transcriptomics data have established that the performance of some statistical methods is contingent on filtering lowly-expressed genes prior to analysis, we performed another secondary analysis in which we filtered peaks open in less than 5% of cells (i.e., zero counts in >95% of cells) in the single-cell data; no such filtering was applied in the primary analysis.

Measuring concordance between DA analyses of single-cell and bulk ATAC-seq requires a matching set of peaks to be defined in both datasets. In our primary analysis, we obtained such a peak set by calling peaks in the matched bulk ATAC-seq data, under the prespecified hypothesis that analyzing bulk data would yield a higher-quality set of peak definitions. However, this procedure is at odds with the design of most scATAC-seq studies, which instead call peaks in 'pseudobulks' created from single-cell data; this is the strategy implemented in widely used software packages such as Signac, ArchR, and SnapATAC. This divergence raises the possibility that the relative performance of single-cell DA methods might depend on the definitions of the peaks provided as input. To evaluate this possibility, we performed a secondary analysis in which we repeated the calculation of concordance using peaks called from 'pseudobulks' of single-cell data, adapting code from Signac (version 1.13.0, function CallPeaks).

A related issue in the analysis of both single-cell and bulk ATAC-seq data is that a certain proportion of peaks called by any computational method are generally thought to represent artifacts of data

processing rather than true biological signal. To test whether these artifactual peaks could influence the relative rankings of single-cell DA methods, we performed an additional sensitivity analysis in which we deliberately introduced a certain amount of artifactual peaks into the dataset. We achieved this by performing a second round of peak calling with MACS2 in the matched bulk ATAC-seq data at a deliberately increased q-value threshold of 0.1. To avoid analyzing broader peaks throughout the genome, we removed peaks that overlap with the original peak set and then combined the remaining peaks with the original peak set. We then conducted DA analysis using the expanded peak set, in which many of the additional peaks are expected to be artifactual. We repeated the analysis of concordance between single-cell and bulk DA as described above, and used a linear model to formally test for an interaction between DA method performance and the peak set used as input.

A final issue that is relevant to single-cell DA analysis is whether single-cell DA methods exhibit differential performance on peaks located in enhancer versus promoter regions. Enhancer elements are generally supported by fewer read counts than promoters in both bulk and single-cell ATAC-seq data^{2,32}, which raises the possibility that DA inference will likely be less accurate in general for enhancer elements. However, the question of whether certain DA methods are better-suited for analysis of enhancer elements in single-cell data has not, to our knowledge, previously been addressed. To evaluate this possibility, we repeated our analysis of DA concordance between single-cell and matched bulk ATAC-seq data, but calculating concordance separately for peaks located in enhancer versus promoter elements, as derived from ENCODE Registry of Candidate cis-Regulatory Elements⁷⁹.

Evaluating single-cell DA methods with single-cell multi-omics

Our survey of the literature (Fig. 1a) identified three publications in which matched RNA-seq and ATAC-seq data was collected from the same nuclei, using multi-omic assays to profile at least two independent replicates, and which involve a comparison between discrete cell types or conditions. We supplemented these three datasets with our own, newly-collected multi-omic dataset from the mouse spinal cord⁸⁰. The list of datasets incorporated in Experiment 2 was therefore as follows:

1. Arguelaguet et al.⁸¹ collected multi-omic data from 44,455 cells across a time course of mouse embryonic development, then defined cell types using the RNA modality. DA analysis was performed between cell types in this dataset. To avoid conflating peaks that are DA between cell types versus across differentiation trajectories, we limited our analysis to a single timepoint (E8.5, the timepoint with the greatest number of cells).
2. Boukhaled et al.⁸² collected multi-omic data from 16,544 peripheral blood mononuclear cells obtained from patients with melanoma and healthy controls, then defined cell types using the RNA modality. DA analysis was performed between cell types in this dataset. To avoid conflating peaks that are DA between cell types versus between healthy and diseased patients, we limited our analysis to the healthy controls in this dataset.
3. Luecken et al.⁵⁹ collected multi-omic data from 69,249 bone marrow mononuclear cells obtained from healthy donors, then defined cell types using both the RNA and ATAC modalities. DA analysis was performed between cell types in this dataset.
4. Last, we collected a new multi-omic dataset comprising 40,526 cells from the mouse spinal cord before and after spinal cord injury (SCI), and defined cell types using the RNA modality. DA analysis was performed between cell types in this dataset.

Notably, for several of these datasets, the authors of the original studies annotated dozens of cell types within the data, implying hundreds of potential pairwise cell type comparisons. To mitigate the

influence of any single dataset on our results, we randomly selected a maximum of 100 pairwise cell type comparisons per dataset.

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

DE analysis of the scRNA-seq data was performed according to the best practices established in our previous work⁴⁹: specifically, we aggregated cells of a given type within a biological replicate to form pseudobulks, then applied the same six statistical methods described above for the analysis of bulk ATAC-seq data in Experiment 1. We again confirmed the robustness of our conclusions by removing individual DE methods from the scRNA-seq analysis and then re-computing the concordance between the ATAC and RNA modalities.

Concordance between the ATAC and RNA modalities was again quantified using the AUCC, but here reverting the value of the parameter k to the default of 500 used in our previous work. We again evaluated the sensitivity of our analysis to the value of this parameter by re-calculating the concordance with $k=100$ and $k=1000$. Moreover, we repeated our sensitivity analysis of low-abundance features by filtering genes accessible in less than 1% of cells; no such filtering was applied in the primary analysis.

Separately, we leveraged the gene-level activity scores to perform parallel GO enrichment analyses of the ATAC and RNA modalities. Gene set enrichment analysis were performed using the 'fgsea' (version 1.25.1) R package⁸³ with 10^6 permutations, using GO term annotations from the Gene Ontology Consortium website and ranking genes in descending order by the absolute value of the test statistic. GO terms annotated to less than 10 or more than 1000 genes were removed to mitigate the influence of very broad or very specific terms on the analysis. Concordance between ATAC and RNA was evaluated using the AUCC, using $k=100$ by default with the rationale that fewer top-ranked GO terms are generally of interest than are top-ranked genes. We then repeated the sensitivity analyses described above, including (i) removing individual DE methods from the scRNA-seq analysis, (ii) varying the value of k ($k=50$ or 500), and (iii) filtering genes accessible in $<1\%$ of cells.

Previous work has shown that inferences about DE are generally more accurate for highly expressed genes^{57,58}. Conversely, identifying instances of true DE among lowly-expressed genes can be highly challenging⁴⁹. These observations raised the possibility that inaccurate inferences about DE for lowly-expressed genes could confound our analysis. Therefore, we performed a final sensitivity analysis for both the gene- and GO-level concordance, whereby we re-evaluated concordance after excluding the bottom tercile of lowly-expressed genes in the RNA modality.

False discoveries in single-cell DA analysis

To study the appearance of false discoveries in single-cell DA analysis, we leveraged both published and simulated scATAC-seq data.

First, we repurposed the Luecken et al. dataset⁵⁹ introduced in Experiment 2 to create artificial comparisons between cells from identical experimental conditions. Specifically, this dataset consists of bone marrow mononuclear cells obtained from healthy donors,

profiled in 13 independent replicates. We first re-processed the dataset in order to analyze DA at the level of peaks, rather than gene activity scores, by calling peaks with MACS2 and using SnapATAC to obtain a peak count matrix. Then, for each cell type within the dataset, we randomly divided half of these replicates to an artificial 'control' group, and the other half to an artificial 'treatment' group. We then performed DA analysis between cells from randomly assigned replicates, and computed the total number of DA peaks identified at a 5% false discovery rate (FDR) using the 'p.adjust' function with method = 'BH'.

We complemented this experiment with two simulation studies that allowed us to more precisely control the presence and degree of biological variation between libraries. First, we repurposed bulk ATAC-seq libraries to simulate scATAC-seq data with no biological cell-to-cell variation at all by randomly downsampling reads from bulk ATAC-seq libraries to create simulated single-cell ATAC profiles. For this, bulk ATAC-seq data of lymphoblastic cell lines (LCLs) from the 1000 Genomes Project⁸⁴ was downloaded from the ENA in CRAM format, and decompressed to BAM format. Because the original dataset comprises data from 100 individuals, we randomly selected a subset of six libraries. For each of these bulk ATAC-seq libraries, we simulated a scATAC-seq library with no cell-to-cell biological variation by randomly sampling fragments with replacement into individual 'cells', adapting a strategy proposed by the authors of SnapATAC³⁴. We found that sampling 0.1% of the fragments from each library into a dataset comprising 1000 cells yields simulated scATAC-seq data with a mean number of fragments per cell that is comparable to our own scATAC-seq data. Peak calling was performed on the aggregated single-cell profiles from all 1000 cells, and a peak count matrix was obtained using SnapATAC. We then divided half of the simulated scATAC-seq replicates to an artificial 'control' group, and half of these replicates to an artificial 'treatment' group, and performed DA analysis between groups using the same single-cell DA methods as above. Performance was measured by the total number of DA peaks identified by each method at a 5% FDR.

To model biological and technical variation between cells more precisely, we carried out a second simulation study using the Splatter (version 3.10) package⁶⁰. This approach has the advantage of allowing exact control over the degree of technical variation between simulated scATAC-seq libraries, while maintaining a realistic degree of cell-to-cell biological variation. To this end, we fit the parameters of the simulation based on our own newly-collected dataset from the mouse spinal cord. We first re-processed this dataset in order to generate a peak count matrix using MACS2 and SnapATAC (rather than a gene count matrix, as in Experiment 2), and then estimated the parameters of the simulation from the resulting peak count matrix using the 'splatEstimate' function. The 'splatSimulate' function was then used to simulate a range of scATAC-seq datasets, differing in the total number of cells (100, 250, 500, 1000, or 2500 per condition); the total number of libraries (2, 3, 4, 5, or 10 per condition); and the intensity of technical variation between libraries (as controlled by the 'de.loc' parameter, which was varied from 0.5 to 2 in increments of 0.5). The proportion of peaks affected by technical variation (the 'de.prob' parameter) was fixed at 50%. Each of the resulting datasets then had the same number of peaks as our own spinal cord dataset. We randomly assigned half of the simulated scATAC-seq libraries to an artificial 'control' group, and the other half to an artificial 'treatment' group. We then performed DA analysis between randomly assigned groups, and calculated the total number of DA peaks identified by each method at a 5% FDR.

Our past experience using Splatter to simulate scRNA-seq data^{49,85–87} drew our attention to the fact that this package produces simulated datasets in which every simulated library is sequenced to identical depth. This is an unrealistic assumption, since real-world libraries inevitably display some variation in sequencing depth. To evaluate the impact of variation in sequencing depth on our conclusions, we performed an additional sensitivity analysis. We first created

five downsampled versions of our own scATAC-seq dataset, whereby each dataset was downsampled to between 50% and 90% of the full depth, in increments of 10%. We then re-estimated simulation parameters separately from each downsampled dataset, as well as the full dataset, and combined the six sets of simulation parameters to simulate a total of 1000 cells across six libraries, such that each library was simulated from an independent set of simulation parameters. This procedure afforded a simulated dataset identical to those described above, except in that each of the libraries demonstrated a different sequencing depth. We then randomly divided libraries into artificial conditions, performed single-cell DA analysis, and calculated the number of DA peaks at a 5% FDR.

We previously showed that, in DE analysis of single-cell RNA-seq data, accounting for variation between biological replicates is required to achieve control of the false discovery rate. Because some of the single-cell DA methods analyzed here are capable of accounting for this effect, we performed secondary analysis to test the impact of incorporating biological replicate into the underlying model. Specifically, we explored the impact of adding replicate as a covariate to LR_{clusters} and negative binomial regression, using the 'latent.vars' argument implemented in the Seurat function FindMarkers. Moreover, because it may be more appropriate to consider replicate as a random effect rather than a fixed effect, we also explored the impact of fitting negative binomial generalized linear mixed models (GLMMs), using the fast approximations in the NEBULA (version v1.4.1) package⁶¹.

Biases of single-cell DA methods

We conducted a series of experiments using both real and simulated data to evaluate the biases of methods for single-cell DA analysis.

We began by re-analyzing the same published scATAC-seq datasets examined in Experiment 1. We computed three properties for the peaks called as DA by each method, including (i) the mean number of counts supporting each peak across all cells (that is, the mean read depth); (ii) the percentage of cells in which that peak was open; and (iii) the width of that peak, in base pairs. To control for differences in the total number of peaks called as DA by each statistical method, we ranked peaks by their p-values and limit our analysis to the top-1000 DA peaks called by each method. We then performed a sensitivity analysis by increasing or decreasing the number of top-ranked DA peaks that are used to calculate the three summary statistics described above (i.e., the top-500 or the top-5000 DA peaks).

We then re-analyzed the published and simulated scATAC-seq dataset examined in Experiment 3. Here, we leveraged the ground truth afforded by simulation studies to specifically characterize these biases for peaks that are known to represent false discoveries. Specifically, we binned peaks into deciles according to each of the three peak properties, and computed both the absolute number as well as the proportion of false discoveries arising from each decile. The rationale of this analysis was to establish whether there was a relationship between the properties of a given peak (e.g., its overall propensity to be open across all cells in a study) and its likelihood of being spuriously identified as DA.

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

Impact of log-fold change filtering

To study the impact of log-fold change filtering, we again made use of the matching bulk ATAC-seq and single-cell multi-omic datasets employed in Experiments 1 and 2, respectively. We binned peaks (genes) according to their deciles of absolute log-fold change in the scATAC-seq data, and then removed the bottom 10% to 90% of peaks (genes) with the lowest log-fold change in each dataset. This procedure was repeated after removing an equal number of peaks from the

dataset at random. We then computed the difference in the AUCC when filtering by log-fold change versus at random, and summarized the difference across DA methods using Cohen's d . Because log-fold change estimates differ across software implementations, we standardized log-fold change calculation for all DA methods using the code implemented in the Seurat function `FoldChange`³⁵ (Eq. 1):

$$\begin{aligned} \widetilde{X}_1 &= \{ \widetilde{X}_{1,g_1}, \dots, \widetilde{X}_{1,g_m} \}, \widetilde{X}_{1,g_i} = \frac{1}{n_i} \sum_j^{n_i} (X_{g_j} + 1) \\ \log(FC) &= \log(e^{\widetilde{X}_1} - 1) - \log(e^{\widetilde{X}_2} - 1) \end{aligned} \quad (1)$$

Separately, we evaluated the effect of log-fold change filtering on the number of false discoveries in single-cell DA analysis. For this purpose, we repeated the random comparisons of published scATAC-seq data in Experiment 3, and compared the number of false discoveries returned by each DA method after log-fold change filtering versus random peak filtering.

Best practices for scATAC-seq analysis

To evaluate the impact of binarizing scATAC-seq data, we first repeated the DA analysis of the scATAC-seq data used in Experiment 1 after binarizing the single-cell count matrices, and then re-calculated the concordance with the bulk DA results as described above. We then directly compared the AUCCs obtained from analysis of binarized and non-binarized data, and summarized these differences with Cohen's d . We carried out similar analyses of binarized count matrices for the real and simulated datasets used in Experiment 3, and re-calculated the number of false discoveries obtained from DA analysis of binarized datasets. Last, we repeated our analyses of the properties of the top-1000 DA peaks (i.e., mean read depth, proportion of cells in which the peak is open, and peak width) in binarized versions of the same published scATAC-seq datasets) in order to establish the impact of binarization on the biases of each DA method.

We carried out a similar suite of analyses to study the impact of various approaches to the normalization of scATAC-seq data. Whereas most of the DA methods considered in this study operate directly on count matrices or binarized data, three methods (t -test, Wilcoxon rank-sum test, and LR_{clusters}) are generally applied to normalized data, with the expectation that normalization is required to produce biologically accurate results. To identify the most widely used methods for normalizing scATAC-seq data for DA analysis, we reviewed the approaches implemented by published scATAC-seq analysis packages^{33–43}. Based on this review, we complemented our analyses of the log-counts per 10,000 normalization implemented by default in Signac, which calls the `NormalizeData` function from Seurat (version 3.1.15), with additional DA analyses using the following additional normalization strategies: (i) raw counts; (ii) scaling counts to a total of 10,000 per cell without log-transformation, using the “relative counts” implementation in Signac; (iii) scaling counts to the median number of counts per cell^{33,36} without log-transformation; (iv) scaling counts to the median number of counts per cell with log-transformation; and (v) TF-IDF normalization, as implemented in the Signac function `RunTFIDF`. We also evaluated the impact of a normalization method recently developed for bulk ATAC-seq data (“smooth GC-full-quantile”)⁸⁸ that explicitly seeks to regress out technical properties of the underlying sequencing data, notably GC bias. Software implementations from the Signac package were used for all normalization strategies except smooth GC-full-quantile, for which the implementation in the `qsmooth` package (version 1.15.1) was used.

In the final experiment of this section, we evaluated the impact of a proposed approach to controlling for technical covariates implemented in the ArchR (version 1.0.1, function `matchBiasCellGroups`) package³³. Briefly, ArchR does not use all available cells in DA analysis but rather performs a statistical test for DARs on a subset of all cells.

Specifically, for a given set of ‘foreground’ cells (e.g., cells of a user-specified type), ArchR constructs a set of ‘background’ cells of equal size that are matched according to some set of QC features. By default, TSS enrichment and \log_{10} (# of fragments) are used to select a matching set of ‘background’ cells. We compared the concordance between DA analysis of scATAC-seq and matching bulk ATAC-seq before and after applying this background matching procedure to select a subset of cells for DA analysis.

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

Data requirements for single-cell DA analysis

To study the impact of sequencing depth on DA analysis, we downsampled the datasets used in Experiment 1 to a mean of 500, 1000, 2000, 5000, or 10,000 fragment counts per cell. This was achieved by using the ‘`downsampleMatrix`’ function from the ‘`scuttle`’ R package (version 1.8.1) to perform downsampling on the entire matrix rather than on a per-cell basis. We then repeated the AUCC analysis for each downsampled dataset and calculated the difference in AUCC relative to DA analysis with 10,000 counts per cell.

To study the impact of the number of cells profiled on DA analysis, we downsampled the datasets used in Experiment 2 to consider only 20, 50, 100, 200, 500, or 1000 cells per condition. Only the subset comparisons from Experiment 2 that involved at least 1,000 cells per group were used for this analysis. We then repeated the AUCC analysis for each downsampled dataset and calculated the difference in AUCC relative to DA analysis with 1,000 cells per condition.

Scalability of single-cell DA methods

Peak memory usage and wall time were monitored using the ‘`peakRAM`’ R package (version 1.0.3), for each of the DA analyses described in Experiment 1 and 2. Because some of the DA methods studied in this Registered Report have different implementations in widely-used analysis packages, including the Wilcoxon rank-sum test (Signac FindMarkers vs. ArchR `sparseMatWilcoxon`), the t -test (Signac FindMarkers vs. ArchR `sparseMatTTest`), and negative binomial regression (Signac FindMarkers vs. `glmGamPoi` `glm_gp`⁶⁷), we evaluated both implementations of these DA methods.

Visualization

Throughout the paper, the box plots show the median (horizontal line), interquartile range (hinges) and smallest and largest values no more than 1.5 times the interquartile range (whiskers).

Statistics and reproducibility

Sample sizes were chosen based on a review of publicly available data and were determined by the authors of the original studies. No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All of the datasets analyzed in this study are publicly available. A complete list of accessions or URLs is provided below. Simulated datasets generated in this study and intermediate data files necessary to reproduce our results are available via Zenodo (<https://doi.org/10.5281/zenodo.10687784>).

The complete list of accessions for published datasets is as follows:

- Experiment 1: Evaluating single-cell DA methods with matched bulk data
 - Corces et al.⁷¹: scATAC-seq, [GSE74310](#); bulk ATAC-seq, [GSE74912](#)
 - Buenrostro et al.⁷²: scATAC-seq, [GSE96772](#); bulk ATAC-seq, [GSE74912](#)
 - Pliner et al.⁵⁶: bulk and scATAC-seq, [GSE109828](#)
 - Satpathy et al.⁷³: scATAC-seq, [GSE107816](#); bulk ATAC-seq, [GSE107223](#)
 - Gonzalez-Blas et al.⁶⁹: bulk and scATAC-seq, [GSE114557](#)
 - Experiment 2: Evaluating single-cell DA methods with single-cell multi-omics
 - Arguelaguet et al.⁸¹: [GSE205117](#)
 - Boukhaled et al.⁸²: [GSE199994](#)
 - Luecken et al.⁵⁹: [GSE194122](#)
 - New multi-omic dataset of SCI: [GSE230765](#)
 - Experiment 3: False discoveries in single-cell DA
 - Luecken et al.⁵⁹: [GSE194122](#)
 - 1000 Genomes Project LCLs⁸⁴: ERP110508 [<https://www.ebi.ac.uk/ena/browser/view/PRJEB28318>].
- Source data are provided with this paper.

Code availability

Source code necessary to reproduce our analyses is available via GitHub (<http://github.com/neurorestore/DA-analysis>)⁸⁹. Libra is available via GitHub (<http://github.com/neurorestore/Libra>).

References

1. Carter, B. & Zhao, K. The epigenetic basis of cellular heterogeneity. *Nat. Rev. Genet.* **22**, 235–250 (2021).
2. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).
3. Hesselberth, J. R. et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
4. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
5. Graybuck, L. T. et al. Enhancer viruses for combinatorial cell-subclass-specific labeling. *Neuron* **109**, 1449–1464.e13 (2021).
6. Mich, J. K. et al. Functional enhancer elements drive subclass-selective expression from mouse to primate neocortex. *Cell Rep.* **34**, 108754 (2021).
7. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
8. Egervari, G. et al. Chromatin accessibility mapping of the striatum identifies tyrosine kinase FYN as a therapeutic target for heroin use disorder. *Nat. Commun.* **11**, 4634 (2020).
9. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
10. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
11. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
12. Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
13. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
14. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
15. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
16. Jin, W. et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature* **528**, 142–146 (2015).
17. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
18. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
19. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
20. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
21. Trevino, A. E. et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053–5069.e23 (2021).
22. Ziffra, R. S. et al. Single-cell epigenomics reveals mechanisms of human cortical development. *Nature* **598**, 205–213 (2021).
23. Herring, C. A. et al. Human prefrontal cortex gene regulatory dynamics from gestation to adulthood at single-cell resolution. *Cell* **185**, 4428–4447.e28 (2022).
24. Corces, M. R. et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer’s and Parkinson’s diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
25. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598**, 86–102 (2021).
26. Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001.e19 (2021).
27. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).
28. Meijer, M. et al. Epigenomic priming of immune genes implicates oligodendroglia in multiple sclerosis susceptibility. *Neuron* **110**, 1193–1210.e13 (2022).
29. Pritykin, Y. et al. A unified atlas of CD8 T cell dysfunctional states in cancer and infection. *Mol. Cell* **81**, 2477–2493.e10 (2021).
30. Morabito, S. et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer’s disease. *Nat. Genet.* **53**, 1143–1155 (2021).
31. Miao, Z. & Kim, J. Uniform quantification of single-nucleus ATAC-seq data with Paired-Insertion Counting (PIC) and a model-based insertion rate estimator. *Nat. Methods* **21**, 32–36 (2024).
32. Martens, L. D., Fischer, D. S., Theis, F. J. & Gagneur, J. Modeling fragment counts improves single-cell ATAC-seq analysis. *Nat. Methods* **21**, 28–31 (2024).
33. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
34. Fang, R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12**, 1337 (2021).
35. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
36. Danese, A. et al. EpiScanpy: integrated single-cell epigenomic analysis. *Nat. Commun.* **12**, 5228 (2021).
37. Ashuach, T., Reidenbach, D. A., Gayoso, A. & Yosef, N. PeakVI: a deep generative model for single-cell chromatin accessibility analysis. *Cell Rep. Methods* **2**, 100182 (2022).

38. Baker, S. M., Rogerson, C., Hayes, A., Sharrocks, A. D. & Rattray, M. Classifying cells with Scasat, a single-cell ATAC-seq analysis tool. *Nucleic Acids Res.* **47**, e10 (2019).
39. Schwartz, G. W., Zhou, Y., Petrovic, J., Pear, W. S. & Faryabi, R. B. TooManyPeaks identifies drug-resistant-specific regulatory elements from single-cell leukemic epigenomes. *Cell Rep.* **36**, 109575 (2021).
40. Prompsy, P. et al. Interactive analysis of single-cell epigenomic landscapes with ChromSCape. *Nat. Commun.* **11**, 5702 (2020).
41. Zamanighomi, M. et al. Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.* **9**, 2410 (2018).
42. Yu, W., Uzun, Y., Zhu, Q., Chen, C. & Tan, K. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome Biol.* **21**, 94 (2020).
43. Ji, Z., Zhou, W. & Ji, H. Single-cell regulome data analysis by SCRAT. *Bioinformatics* **33**, 2930–2932 (2017).
44. Gontarz, P. et al. Comparison of differential accessibility analysis strategies for ATAC-seq data. *Sci. Rep.* **10**, 10150 (2020).
45. Eder, T. & Grebien, F. Comprehensive assessment of differential ChIP-seq tools guides optimal algorithm selection. *Genome Biol.* **23**, 119 (2022).
46. Steinhauser, S., Kurzawa, N., Eils, R. & Herrmann, C. A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief. Bioinforma.* **17**, 953–966 (2016).
47. Chen, H. et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019).
48. Mehta, T., Tanik, M. & Allison, D. B. Towards sound epistemological foundations of statistical methods for high-dimensional biology. *Nat. Genet.* **36**, 943–947 (2004).
49. Squair, J. W. et al. Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
50. Method of the Year 2019. Single-cell multimodal omics. *Nat. Methods* **17**, 1 (2020).
51. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
52. Zhu, C. et al. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).
53. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
54. Irizarry, R. A. et al. Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2**, 345–350 (2005).
55. Sonesson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
56. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018).
57. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
58. Rapaport, F. et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14**, R95 (2013).
59. Luecken, M. D. et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)* 13 (2021).
60. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
61. He, L. et al. NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun. Biol.* **4**, 629 (2021).
62. Sonesson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinforma.* **14**, 91 (2013).
63. Dalman, M. R., Deeter, A., Nimishakavi, G. & Duan, Z.-H. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinforma.* **13**, S11 (2012).
64. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat. Methods* **20**, 665–672 (2023).
65. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
66. Korsunsky, I., Nathan, A., Millard, N. & Raychaudhuri, S. Presto scales Wilcoxon and auROC analyses to millions of observations. *BioRxiv* <https://doi.org/10.1101/653253> (2019).
67. Ahlmann-Eltze, C. & Huber, W. glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **36**, 5701–5702 (2021).
68. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e20 (2020).
69. Bravo González-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
70. Bravo González-Blas, C. et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *BioRxiv* <https://doi.org/10.1101/2022.08.19.504505> (2022).
71. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
72. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548.e16 (2018).
73. Satpathy, A. T. et al. Transcript-indexed ATAC-seq for precision immune profiling. *Nat. Med.* **24**, 580–590 (2018).
74. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
75. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
76. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
77. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
78. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
79. ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
80. Skinnider, M. A. et al. Single-cell and spatial atlases of spinal cord injury in the *Tabulae Paralytica*. *Nature* **631**, 150–163 (2024).
81. Argelaguet, R. et al. Decoding gene regulation in the mouse embryo using single-cell multi-omics. *BioRxiv* <https://doi.org/10.1101/2022.06.15.496239> (2022).
82. Boukhaled, G. M. et al. Pre-encoded responsiveness to type I interferon in the peripheral immune system defines outcome of PD1 blockade therapy. *Nat. Immunol.* **23**, 1273–1283 (2022).
83. Korotkevich, G. et al. Fast gene set enrichment analysis. *BioRxiv* <https://doi.org/10.1101/060012> (2016).
84. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
85. Skinnider, M. A. et al. Cell type prioritization in single-cell data. *Nat. Biotechnol.* **39**, 30–34 (2021).
86. Kathe, C. et al. The neurons that restore walking after paralysis. *Nature* **611**, 540–547 (2022).

87. Squair, J. W., Skinnider, M. A., Gautier, M., Foster, L. J. & Courtine, G. Prioritization of cell types responsive to biological perturbations in single-cell data with Augur. *Nat. Protoc.* **16**, 3836–3873 (2021).
88. Van den Berge, K. et al. Normalization benchmark of ATAC-seq datasets shows the importance of accounting for GC-content effects. *Cell Rep. Methods* **2**, 100321 (2022).
89. Teo, A. Y. Y. et al. Best practices for differential accessibility analysis in single-cell epigenomics. DA-analysis. <https://doi.org/10.5281/zenodo.13772270> (2024).

Acknowledgements

This work was supported by the Swiss National Science Foundation (grant nos. 310030_192558 and 320030-227881 to G.C. and PZ00P3_208988 to J.W.S.); Wings for Life (to M.A.S.); Princeton Precision Health (to M.A.S.); Friedrich Flick Förderungsstiftung through Wings for Life (to G.C.); Wyss Center for Bio and Neuroengineering (to G.C.); and the ALARME Foundation (to G.C.).

Author contributions

A.Y.Y.T., J.W.S., G.C., and M.A.S. conceived and designed experiments. A.Y.Y.T. conducted experiments. A.Y.Y.T. and M.A.S. analysed the data. A.Y.Y.T., J.W.S., G.C., and M.A.S. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-53089-5>.

Correspondence and requests for materials should be addressed to Jordan W. Squair, Gregoire Courtine or Michael A. Skinnider.

Peer review information *Nature Communications* thanks Ivan Costa, Jeff Vierstra and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024