SnapATAC: A Comprehensive Analysis Package for Single Cell ATAC-seq

Rongxin Fang[1,2], Sebastian Preissl[3], Yang Li[2], Xiaomeng Hou[3], Jacinta Lucero[4], Xinxin Wang[3], Amir Motamedi[5], Andrew K. Shiau[5], Xinzhu Zhou[6], Fangming Xie[7], Eran A. Mukamel[7], Kai Zhang[2], Yanxiao Zhang[2], M. Margarita Behrens[4], Joseph R. Ecker[4,8], and Bing Ren[2,3,9*]

1. Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA 92093, USA
2. Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA
3. Center for Epigenomics, Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093, USA
4. The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
5. Small Molecule Discovery Program, Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA
6. Biomedical Science Graduate Program, University of California San Diego, La Jolla, CA 92093, USA
7. Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92037, USA
8. Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
9. Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, UCSD Moores Cancer Center, La Jolla, CA 92093, USA

*Correspondence to: biren@ucsd.edu

Abstract

Identification of the *cis*-regulatory elements controlling cell-type specific gene expression patterns is essential for understanding the origin of cellular diversity. Conventional assays to map regulatory elements via open chromatin analysis of primary tissues is hindered by heterogeneity of the samples. Single cell analysis of transposase-accessible chromatin (scATAC-seq) can overcome this limitation. However, the high-level noise of each single cell profile and the large volumes of data could pose unique computational challenges. Here, we introduce SnapATAC, a software package for analyzing scATAC-seq datasets. SnapATAC can efficiently dissect cellular heterogeneity in an unbiased manner and map the trajectories of cellular states. Using the Nyström method, a sampling technique that generates the low rank embedding for large-scale dataset, SnapATAC can process data from up to a million cells. Furthermore, SnapATAC incorporates existing tools into a comprehensive package for analyzing single cell ATAC-seq dataset. As demonstration of its utility, SnapATAC was applied to 55,592 single-nucleus ATAC-seq profiles from the mouse secondary motor cortex. The analysis revealed ~370,000 candidate regulatory elements in 31 distinct cell populations in this brain region and inferred candidate transcriptional regulators in each of the cell types.

Introduction

A multicellular organism comprises diverse cell types, each highly specialized to carry out unique functions. Each cell lineage is established during development as a result of tightly regulated spatiotemporal gene expression programs[1], which are driven in part by sequence-specific transcription factors that interact with *cis*-regulatory sequences in a cell-type specific manner[2]. Thus, identifying the *cis*-elements and their cellular specificity is an essential step towards understanding the developmental programs encoded in the linear genome sequence.

Since the *cis*-regulatory elements are often marked by hypersensitivity to nucleases or transposases when they are active or poised to act, approaches to detect chromatin accessibility, such as ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing)[3] and DNase-seq (DNase I hypersensitive sites sequencing)[4] have been widely used to map candidate *cis*-regulatory sequences. However, conventional assays that use bulk tissue samples as input cannot resolve cell-type specific usage of *cis* elements and lacks the resolution to study their temporal dynamics. To overcome these limitations, a number of methods have been developed for measuring chromatin accessibility in single cells. One approach involves combinatorial indexing to simultaneously analyze tens of thousands of cells[5]. This strategy has been successfully applied to embryonic tissues in *D. melanogaster*[6], developing mouse forebrains[7] and adult mouse tissues[8]. A related method, scTHS-seq (single-cell transposome hypersensitive site sequencing), has also been used to study chromatin landscapes at single cell resolution in the adult human brains[9]. A third approach relies on isolation of cell using microfluidic devices (Fluidigm, C1)[10] or within individually indexable wells of a nano-well array (Takara Bio, ICELL8)[11]. More recently, single cell ATAC-seq analysis has been demonstrated on droplet-based platforms[12,13], enabling profiling of chromatin accessibility from hundreds of thousands cells in a single experiment[13]. Hereafter, these methods are referred to collectively as single cell ATAC-seq (scATAC-seq).

The growing volume of scATAC-seq datasets coupled with the sparsity of signals in each individual profile due to low detection efficiency (5-15% of peaks detected per cell)[7] present a unique computational challenge. To address this challenge, a number of

77 unsupervised algorithms have been developed. One approach, chromVAR[14], groups
78 similar cells together by dissecting the variability of transcription factor (TF) motif
79 occurrence in the open chromatin regions in each cell. Another approach employs the
80 natural language processing techniques such as Latent Semantic Analysis (LSA)[8] and
81 Latent Dirichlet Allocation (LDA)[15] to group cells together based on the similarity of
82 chromatin accessibility. A third approach analyzes the variability of chromatin
83 accessibility in cells based on the k-mer composition of the sequencing reads from each
84 cell[13,16]. A fourth approach, Cicero[17], infers cell-to-cell similarities based on the gene
85 activity scores predicted from their putative regulatory elements in each cell.
86
87 Because the current methods often require performing linear dimensionality reduction
88 such as singular value decomposition (SVD) on a cell matrix of hundreds of thousands of
89 dimensions, scaling the analysis to millions of cells remains very challenging or nearly
90 impossible. In addition, the unsupervised identification of cell types or states in complex
91 tissues using scATAC-seq dataset does not have the same degree of sensitivity as that from
92 scRNA-seq[18]. One possibility is that the current methods rely on the use of pre-defined
93 accessibility peaks based on the aggregate signals. There are several limitations to this
94 choice. First, the cell type identification could be biased toward the most abundant cell
95 types in the tissues, and consequently lack the ability to reveal regulatory elements in the
96 rare cell populations that could be underrepresented in the aggregate dataset. Second, a
97 sufficient number of single cell profiles would be required to create robust aggregate
98 signal for creating the peak reference.
99
100 To overcome these limitations, we introduce a software package, Single Nucleus Analysis
101 Pipeline for ATAC-seq – SnapATAC (https://github.com/r3fang/SnapATAC) - that does
102 not require population-level peak annotation prior to clustering. Instead, it resolves
103 cellular heterogeneity by directly comparing the similarity in genome-wide accessibility
104 profiles between cells. We also adopt a new sampling technique, ensemble Nyström
105 method[19,20], that significantly improves the computational efficiency and enables the
106 analysis of scATAC-seq from a million cells on typical hardware. SnapATAC also
107 incorporates many existing tools, including integration of scATAC-seq and scRNA-seq
108 dataset[18], prediction of enhancer-promoter interaction, discovery of key transcription

factors[21], identification of differentially accessible elements[22], construction of trajectories during cellular differentiation, correction of batch effect[23] and classification of new dataset based on existing cell atlas[18], into one single package to maximize its utility and functionalities. Thus, SnapATAC represents a comprehensive solution for scATAC-seq analysis.

Through extensive benchmarking using both simulated and empirical datasets from diverse tissues and species, we show that SnapATAC outperforms current methods in accuracy, sensitivity, scalability and reproducibility for cell type identification from complex tissues. Furthermore, we demonstrate the utility of SnapATAC by building a high-resolution single cell atlas of the mouse secondary motor cortex. This atlas comprises of ~370,000 candidate *cis*-regulatory elements across 31 distinct cell types, including rare neuronal cell types that account for less than 0.1% of the total population analyzed. Through motif enrichment analysis, we further infer potential key transcriptional regulators that control cell type specific gene expression programs in the mouse brain.

Results

Overview of SnapATAC workflow

A schematic overview of SnapATAC workflow is displayed in **Fig. 1**. SnapATAC first performs pre-processing of sequencing reads including demultiplexing, reads alignments and filtering, duplicate removal and barcode selection using SnapTools (https://github.com/r3fang/SnapTools) (**Supplementary Methods**). The output of this pre-processing step is a "snap" (Single-Nucleus Accessibility Profiles) file (**Supplementary Note 1**) specially formatted for storing single cell ATAC-seq datasets (**Supplementary Fig. 1a**). Users could select high quality single cell ATAC-seq profiles for subsequent analysis based on numbers of unique fragments detected from the cell and percentage of promoter-overlapping fragments[24].

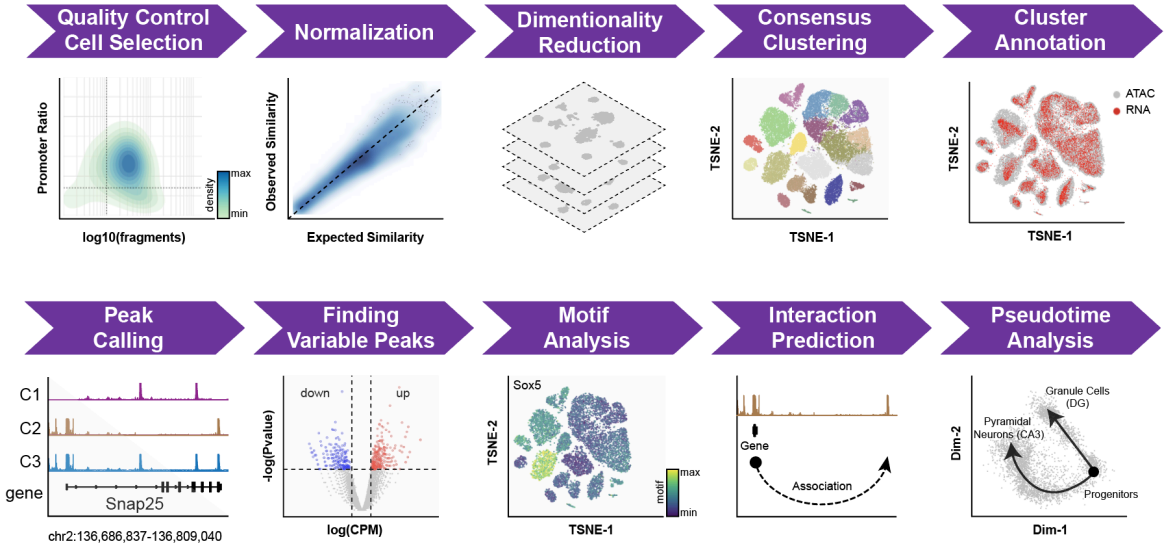


Figure 1. Schematic overview of SnapATAC analysis workflow. See main text for description of each step.

Next, SnapATAC resolves the heterogeneity of cell population by assessing the similarity of chromatin accessibility between cells. To achieve this goal, each single cell chromatin accessibility profile is represented as a binary vector, the length of which corresponds to the number of uniform-sized bins that segment the genome. Through systematic benchmarking, a bin size of 5kb is chosen in this study (**Supplementary Methods** and **Supplementary Fig. 2b**). A bin with value "1" indicates that one or more reads fall within that bin, and the value "0" indicates otherwise. The set of binary vectors from all

148 the cells are converted into a Jaccard similarity matrix, with the value of each element
149 calculated from the fraction of overlapping bins between every pair of cells. Because the
150 value of Jaccard Index could be influenced by sequencing depth of a cell
151 (**Supplementary Methods**), a regression-based normalization method is developed to
152 remove this confounding factor (**Supplementary Methods** and **Supplementary Fig.**
153 **3-4**). Using the normalized similarity matrix, eigenvector decomposition is performed for
154 dimensionality reduction. Finally, in the reduced dimension, SnapATAC uses Harmony[23]
155 to remove potential batch effect between samples introduced by technical variability
156 (**Supplementary Methods**).
157
158 The computational cost of the algorithm scales quadratically with the number of cells. To
159 improve the scalability of SnapATAC, a sampling technique - the Nyström method[19] – is
160 used to efficiently generate the low-rank embedding for large-scale datasets
161 (**Supplementary Methods**). Nyström method contains two major steps: 1) it computes
162 the low dimension embedding for a subset of selected cells (also known as landmarks); 2)
163 it projects the remaining cells to the embedding structure learned from the landmarks.
164 This achieves significant speedup considering that the number of landmarks could be
165 substantially smaller than the total number of cells. Through benchmarking, we further
166 demonstrate that this approach will not sacrifice the performance once the landmarks are
167 chosen appropriately (**Supplementary Methods** and **Supplementary Fig. 5a-c**) as
168 reported before[20].
169
170 Nyström method is stochastic and could yield different clustering results in each sampling.
171 To overcome this limitation, a consensus approach is used that combines a mixture of
172 low-dimensional manifolds learned from different sets of sampling (**Supplementary**
173 **Methods**). Through benchmarking, we demonstrate that the ensemble approach can
174 significantly improve the reproducibility of clustering outcome compared to the standard
175 Nyström method (**Supplementary Fig. 5d**). In addition, this consensus algorithm
176 naturally fits within the distributed computing environments where their computational
177 costs are roughly the same as that of the standard single sampling method.
178

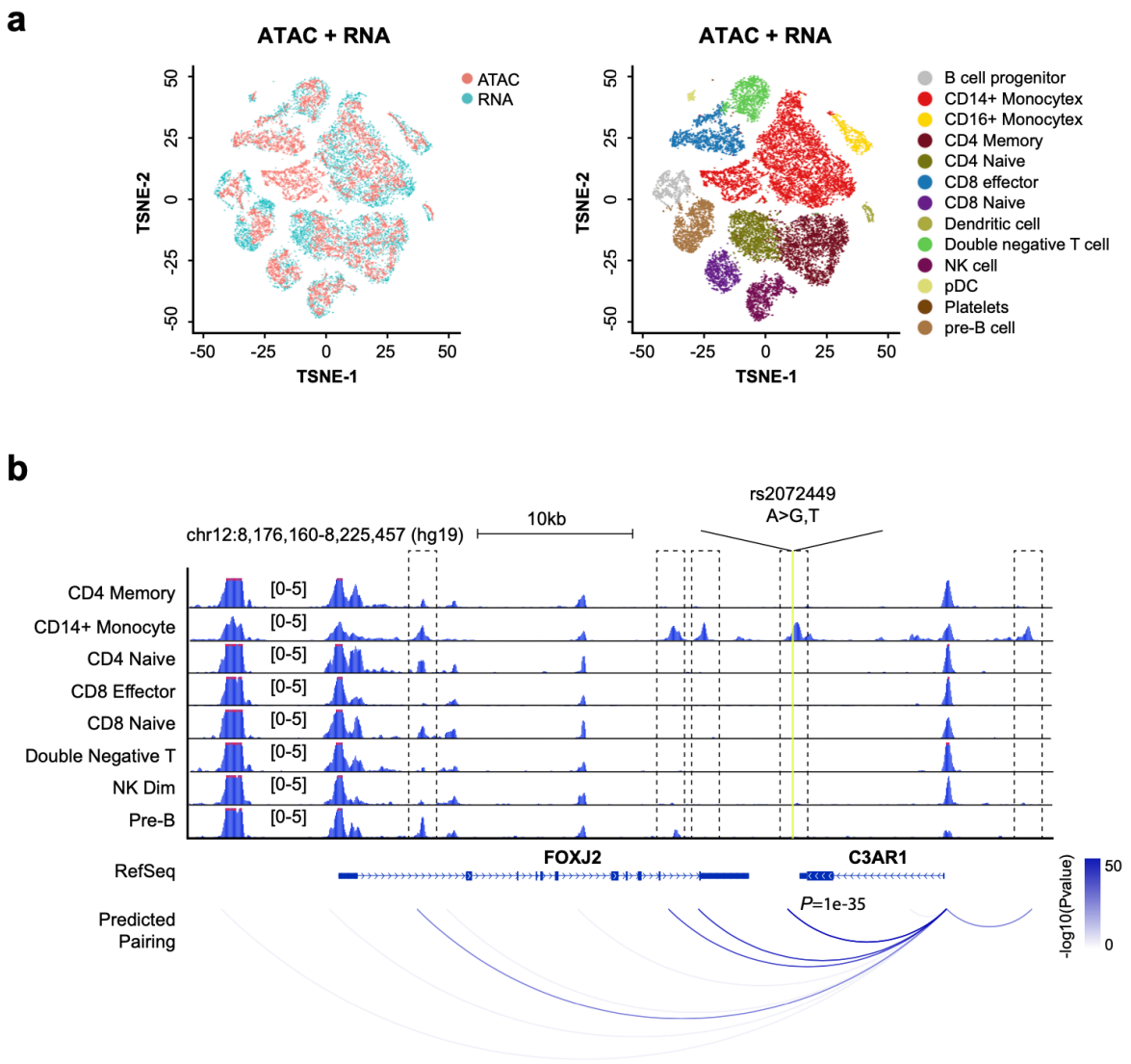


Figure 2. SnapATAC integrates single cell ATAC-seq and RNA-seq data to link enhancers to putative target genes. (**a**) Joint t-SNE visualization of scATAC-seq and scRNA-seq datasets from peripheral blood mononuclear cells (PBMC). Cells are colored by modality (left) and predicted cell types (right). (**b**) Cell-type specific chromatin landscapes are shown together with the association score between gene expression of C3AR1 and accessibility at its putative enhancers. Dash lines highlight the significant enhancer-promoter pairs. Yellow line represents the SNP (rs2072449) that is associated with C3AR1 expression[25].

As a standalone software package, SnapATAC also provides a number of commonly used functions for scATAC-seq analysis by incorporating many existing useful tools, as described below:

First, to facilitate the annotation of resulting cell clusters, SnapATAC provides three different approaches: i) SnapATAC annotates the clusters based on the accessibility score at the canonical marker genes (**Supplementary Methods**); ii) it infers cell type labels by integrating with corresponding single cell RNA-seq datasets[18] (**Supplementary Methods** and **Fig. 2a**); iii) it allows supervised annotation of new single cell ATAC-seq dataset based on an existing cell atlas (**Supplementary Methods**).

Second, SnapATAC allows identification of the candidate regulatory elements in each cluster by applying various peak-calling algorithms[26] to the aggregate chromatin profiles. Differential analysis is then performed to identify cell-type specific regulatory elements[22]. Candidate master transcription factors in each cell cluster are discovered through motif enrichment analysis of the differentially accessible regions in each cluster[27]. SnapATAC further conducts Genomic Regions Enrichment of Annotation Tool (GREAT)[28] analysis to identify the biological pathways active in each cell type.

Third, SnapATAC incorporates a new approach to link candidate regulatory elements to their putative target genes. In contrast to previous method[17] that relies on analysis of co-accessibility of putative enhancers and promoters[29], SnapATAC infers the linkage based on the association between gene expression and chromatin accessibility in single cells where scRNA-seq data is available (**Supplementary Methods**). First, SnapATAC integrates scATAC-seq and scRNA-seq using Canonical Correlation Analysis (CCA) as described in the previous study[30]. Second, for each scATAC-seq profile, a corresponding gene expression profile is imputed based on the weighted average of its *k*-nearest neighboring cells (i.e. k=15) in the scRNA-seq dataset. A "pseudo" cell is created that contains the information of both chromatin accessibility and gene expression. Finally, logistic regression is performed to quantify the association between the gene expression and binarized accessibility state at putative enhancers (**Supplementary Methods**). This new approach is used to integrate ~15K peripheral blood mononuclear cells (PBMC) chromatin profiles and ~10K PBMC transcriptomic profiles (**Fig. 2a**) and represent them in a joint t-SNE embedding space (**Fig. 2a**). Over 98% of the single cell ATAC-seq cells can be confidently assigned to a cell type defined in the scRNA-seq dataset

(**Supplementary Fig. 6a**). Enhancer-gene pairs are predicted for 3,000 genes differentially expressed between cell types in PBMC as determined by scRNA-seq using Seurat[18]. The validity of the prediction is supported by two lines of evidence. First, the association score exhibits a distance decay from the TSS, consistent with the distance decay of interaction frequency observed in chromatin conformation study[31] (**Supplementary Fig. 6b**). Second, the predictions match well with the expression quantitative trait loci (*cis*-eQTLs) derived from interferon-γ and lipopolysaccharide stimulation of monocytes[25] with reasonable prediction power (AUROC=0.66, AUPRC=0.68; **Supplementary Fig. 6c-d** and **Supplementary Methods**). It is important to note that while statistical association between scATAC-seq and scRNA-seq provides another approach to symmetrically link enhancers to their putative target genes, the predictions require further experimental validation.

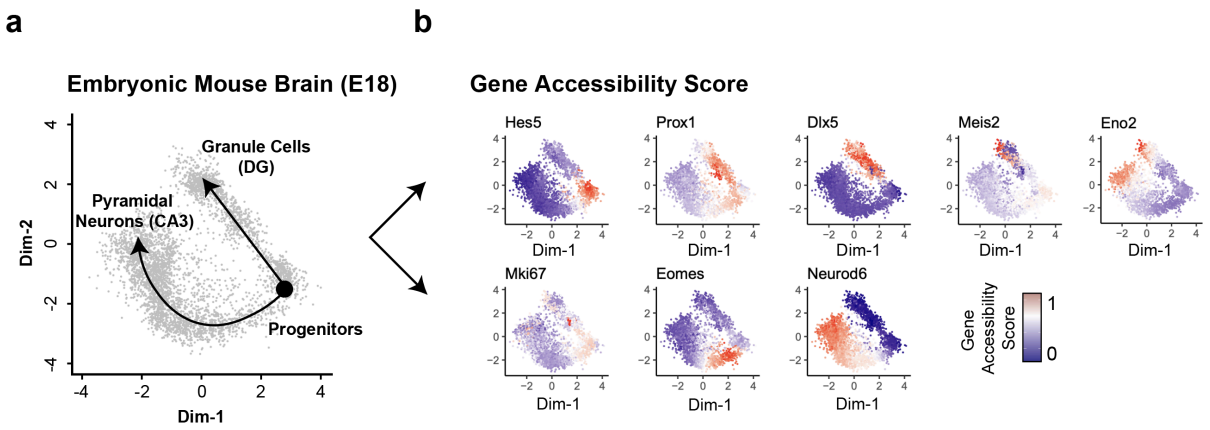


Figure 3. SnapATAC constructs cellular trajectories for the developing mouse brain. (**a**) Two-dimensional visualization of a dataset that contains 4,259 single cell chromatin profiles from the hippocampus and ventricular zone in embryonic mouse brain (E18) reveals two-branch differentiation trajectories from progenitor cells to Granule Cells (DG) and Pyramidal Neurons (CA3) (left). Data source is listed in **Supplementary Table S1**. The cellular trajectory is determined by Slingshot[32]. (**b**) Gene accessibility score of canonical marker genes is projected onto the 2D embedding.

Fourth, SnapATAC has incorporated a function to construct cellular trajectories from single cell ATAC-seq. As a demonstration of this feature, SnapATAC is used to analyze a dataset that contains 4,259 cells from the hippocampus in the fetal mouse brain (E18)

250 (**Supplementary Table S1**). Immature granule cells originating in the dentate gyrus
251 give rise to both mature granule cells (DG) and pyramidal neurons (CA3)[33]. Analysis of
252 4,259 cells reveals a clear branching structure in the first two dimensions (**Fig. 3a**), the
253 pattern of which is remarkably similar to the result previously obtained from single cell
254 transcriptomic analysis[34]. For instance, the DG-specific transcription factor *Prox1* is
255 exclusively accessible in one branch whereas *Neurod6* that is known to be specific to CA3
256 are accessible in the other branch. Markers of progenitors such as *Hes5* and *Mki67*,
257 however, are differentially accessible before the branching point (**Fig. 3b**). Further using
258 lineage inference tool such as Slingshot[32], SnapATAC defines the trajectories of cell states
259 for pseudo-time analysis (**Fig. 3a**). These results demonstrate that SnapATAC can also
260 reveal lineage trajectories with high accuracy.
261
262 **Performance evaluation**
263 To compare the accuracy of cell clustering between SnapATAC and published scATAC-
264 seq analysis methods, a simulated dataset of scATAC-seq profiles are generated with
265 varying coverages, from 10,000 (high coverage) to 1,000 reads per cell (low coverage) by
266 down sampling from 10 previously published bulk ATAC-seq datasets[27]
267 (**Supplementary Table S2** and **Supplementary Methods**). Based on a recent
268 summary of cell ATAC-seq methods[35], LSA[8] and cisTopic[15] outperforms the other
269 methods in separating cell populations of different coverages and noise levels in both
270 synthetic and real datasets. Therefore, we choose to compare SnapATAC with these two
271 methods.
272
273 The performance of each method in identifying the original cell types is measured by both
274 Adjusted Rank Index (ARI) and Normalized Mutual Index (NMI). The comparison shows
275 that SnapATAC is the most robust and accurate method across all ranges of data sparsity
276 (Wilcoxon signed-rank test, $P < 0.01$; **Fig. 4a**; **Supplementary Fig. 7** and
277 **Supplementary Table S3**). Next, a set of 1,423 human cells corresponding to 10
278 distinct cell types generated using C1 Fluidigm platform, where the ground truth is
279 known[14], is analyzed by SnapATAC and other methods. Again, SnapATAC correctly
280 identifies the cell types with high accuracy (**Supplementary Fig. 8**).
281

282 To compare the sensitivity of SnapATAC on detecting cell types to that of previously
283 published methods, we analyzed two scATAC-seq datasets representing different types of
284 bio-samples. First, to quantify the clustering sensitivity, we applied an existing
285 integration method to predict the cell type of 4,792 PBMC cells using corresponding 10X
286 single cell RNA-seq by following the tutorial
287 (https://satijalab.org/seurat/v3.1/atacseq_integration_vignette.html). To obtain the
288 most confident prediction, we only kept single cell ATAC-seq profiles whose cell type
289 prediction score is greater than 0.9. Using the remaining cells, we calculated the
290 connectivity index (CI; **Supplementary Methods**) in the low-dimension manifold for
291 each of the methods (LSA, cisTopic and SnapATAC). Connectivity index estimates the
292 degree of separation between clusters in an unbiased manner and a lower connectivity
293 index represents a higher degree of separation between clusters. SnapATAC exhibits
294 substantially higher sensitivity in distinguishing different cell types compared to the other
295 two methods (**Fig. 4b**). The second is a newly produced dataset that contains 9,529 single
296 nucleus open chromatin profiles generated from the mouse secondary motor cortex.
297 Based on the gene accessibility score at canonical marker genes (**Supplementary Fig.**
298 **9**), SnapATAC uncovers 22 distinct cell populations (**Supplementary Fig. 10**) whereas
299 alternative methods fail to distinguish the rare neuronal subtypes including Sst (Gad2+
300 and Sst+), Vip (Gad2+ and Vip+), L6b (Sulf1- and Tl4e+) and L6.CT (Sulf1+ and Foxp2+).
301 These results suggest that SnapATAC outperforms existing methods in sensitivity of
302 separating different cell types in both synthetic and real datasets.
303
304 To compare the scalability of SnapATAC to that of existing methods, a previous scATAC-
305 seq dataset that contains over 80k cells from 13 different mouse tissues[8] is used
306 (**Supplementary Table S1**). This dataset is down sampled to different number of cells,
307 ranging from 20,000 to 80,000 cells. For each sampling, SnapATAC and other methods
308 are performed, and the CPU running time of dimensionality reduction is monitored
309 (**Supplementary Methods**). The running time of SnapATAC scales linearly and
310 increases at a significantly lower slope than alternative methods (**Fig. 4c**). Using the same
311 computing resource, when applied to 100k cells, SnapATAC is much faster than existing
312 methods (**Fig. 4c**). For instance, when applied to 100k cells, SnapATAC is nearly 10 times
313 faster than LSA and more than 100 times faster than cisTopic. More importantly, because

314 SnapATAC avoids the loading of the full cell matrix in the memory and can naturally fit
315 within the distributed computing environments (**Supplementary Methods**), the
316 running time and memory usage for SnapATAC plateau after 20,000 cells, making it
317 possible for analyzing datasets of even greater volumes. To test this, we simulate one
318 million cells of the same coverage with the above dataset (**Supplementary Methods**)
319 and process it with SnapATAC, LSA and cisTopic. Using the same computing resource,
320 SnapATAC is the only method that is able to process this dataset (**Fig. 4c** and
321 **Supplementary Methods**). These results demonstrate that SnapATAC provides a
322 highly scalable approach for analyzing large-scale scATAC-seq dataset.
323
324 To evaluate the clustering reproducibility, the above mouse scATAC-seq dataset is down-
325 sampled to 90% of the original sequencing depth in five different iterations. Each down
326 sampled dataset is clustered using SnapATAC and other methods. Clustering results are
327 compared between sampled datasets to estimate the stability. SnapATAC has a
328 substantially higher reproducibility of clustering results between different down-sampled
329 datasets than other methods (**Fig. 4d**).
330

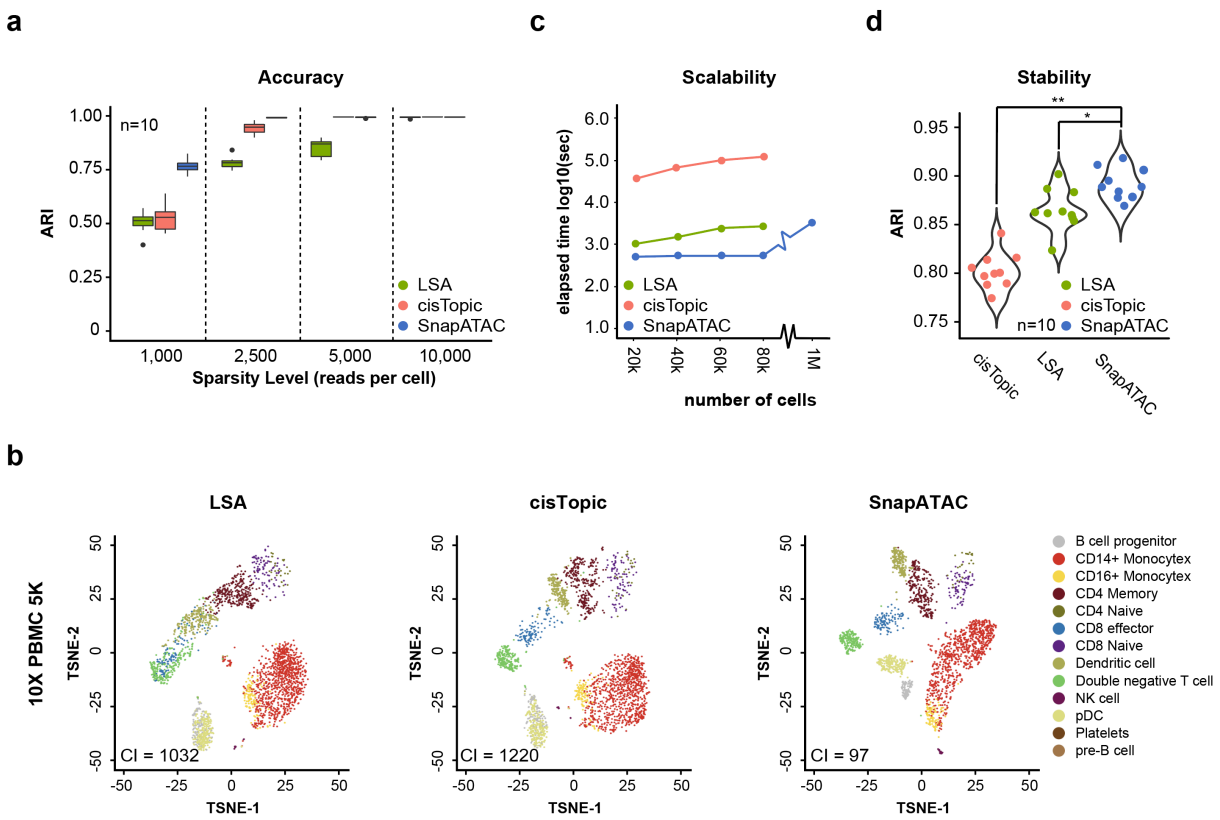


Figure 4. SnapATAC outperforms current methods in accuracy, sensitivity, scalability and stability of identifying cell types in complex tissues. (**a**) A set of simulated datasets are generated with varying coverage ranging from 1,000 to 10,000 reads per cell cells (**Supplementary Methods**). For each coverage, n=10 random replicates are simulated, and clustering accuracy measurement is based on Adjusted Rank Index (ARI). (**b**) T-SNE representation of PBMC single cell ATAC-seq profiles analyzed by LSA (left), cisTopic (middle) and SnapATAC (right). The cell type identification was predicted by 10X PBMC single cell RNA-seq using recent integration method[30]. CI = connectivity index (see **Supplementary Methods**). (**c**) A mouse dataset[8] is sampled to different number of cells ranging from 20k to 1M. For each sampling, we compared the CPU running time of different methods for dimensionality reduction (**Supplementary Methods**). SnapATAC is the only method that is able to process a dataset of one million (1M) cells. (**d**) A set of perturbations (n=5) are introduced to the mouse atlas dataset by down sampling to 90% of the original sequencing depth. Clustering outcomes are compared between different down sampled datasets (n=10) to estimate the reproducibility. One-tailed t-test was performed to estimate the significance level between SnapATAC and each of the other two methods (* < 0.05 and ** < 0.01).

The improved performance of SnapATAC likely results from the fact that it considers all reads from each cell, not just the fraction of reads within the peaks defined in the

352 population. To test this hypothesis, clustering is performed after removing the reads
353 overlapping the predefined peak regions. Although the outcome is worse than the full
354 dataset as expected, it still recapitulates the major cell types obtained from the full dataset
355 (**Supplementary Fig. 11**). This holds true for all three datasets tested
356 (**Supplementary Fig. 11a-c**). One possibility is that the off-peak reads may be enriched
357 for the euchromatin (or compartment A) that strongly correlates with active genes[28] and
358 varies considerably between cell types[29,30]. Consistent with this hypothesis, the density of
359 the non-peak reads in scATAC-seq library is highly enriched for the euchromatin
360 (compartment A) as defined using genome-wide chromatin conformation capture
361 analysis (i.e. Hi-C) in the same cell type[31] (**Supplementary Fig. 12**). These observations
362 suggest that the non-peak reads discarded by existing methods can actually contribute to
363 distinguish different cell types.
364
365 Including the off-peak reads, however, raises a concern regarding whether SnapATAC is
366 sensitive to technical variations (also known as batch effect). To test this, SnapATAC is
367 applied to four datasets generated using different technologies (**Supplementary Table**
368 **S1**). Each dataset contains at least two biological replicates produced by the same
369 technology. In all cases, the biological replicates are well mixed in the t-SNE embedding
370 space showing no batch effect (**Supplementary Fig. 13**), suggesting that SnapATAC is
371 robust to the technical variations.
372
373 To test whether SnapATAC is robust to technical variation introduced by different
374 technological platforms, it is used to integrate two mouse brain datasets generated using
375 plate and droplet-based scATAC-seq technologies (**Supplementary Table S1**). In the
376 joint t-TSNE embedding space, these two datasets are separated based on the
377 technologies (**Supplementary Fig. 14a**). To remove the platform-to-platform
378 variations, Harmony[23], a single cell batch effect correction tool, is incorporated into the
379 SnapATAC pipeline (**Supplementary Methods**). After applying Harmony[23], these two
380 datasets are fully mixed in the joint t-SNE embedding (**Supplementary Fig. 14b**) and
381 clusters are fairly represented by both datasets (**Supplementary Fig. 14c**).
382
383 **A high-resolution *cis*-regulatory atlas of the mouse secondary motor cortex**

384 To demonstrate the utility of SnapATAC in resolving cellular heterogeneity of complex
385 tissues and identify candidate *cis*-regulatory elements in diverse cell type, it is applied to
386 a new single nucleus ATAC-seq dataset generated from the secondary mouse motor cortex
387 in the adult mouse brain as part of the BRAIN Initiative Cell Census Consortium[36]
388 (**Supplementary Fig. 15a**). This dataset includes two biological replicates, each pooled
389 from 15 mice to minimize potential batch effects. The aggregate signals show high
390 reproducibility between biological replicates (Pearson correlation = 0.99;
391 **Supplementary Fig. 15b-d**) and a significant enrichment for transcription start sites
392 (TSS), indicating a high signal-to-noise ratio (**Supplementary Fig. 15e**). After filtering
393 out the low-quality nuclei (**Supplementary Fig. 16a**) and removing putative doublets
394 using Scrublet[37] (**Supplementary Methods**; **Supplementary Fig. 16b**), a total of
395 55,592 nuclear profiles with an average of ~5,000 unique fragments per nucleus remain
396 and are used for further analysis (**Supplementary Table S4**). To our knowledge, this
397 dataset represents one of the largest single cell chromatin accessibility studies for a single
398 mammalian brain region to date.
399
400 SnapATAC identifies initially a total of 20 major clusters using the consensus clustering
401 approach (**Supplementary Fig. 17**). The clustering result is highly reproducible
402 between biological replicates (Pearson correlation=0.99; **Supplementary Fig. 18a**)
403 and is resistant to sequencing depth effect (**Supplementary Fig. 18b**). Based on the
404 gene accessibility score at the canonical marker genes (**Supplementary Fig. 19**), these
405 clusters are classified into 10 excitatory neuronal subpopulations (Snap25+, Slc17a7+,
406 Gad2-; 52% of total nuclei), three inhibitory neuronal subpopulations (Snap25+, Gad2+;
407 10% of total nuclei), one oligodendrocyte subpopulation (Mog+; 8% of total nuclei), one
408 oligodendrocyte precursor subpopulation (Pdgfra+; 4% of total nuclei), one microglia
409 subpopulation (C1qb+; 5% of total nuclei), one astrocyte subpopulation (Apoe+; 12% of
410 total nuclei), and additional populations of endothelial, and smooth muscle cells
411 accounting for 6% of total nuclei (**Fig. 5a**).
412

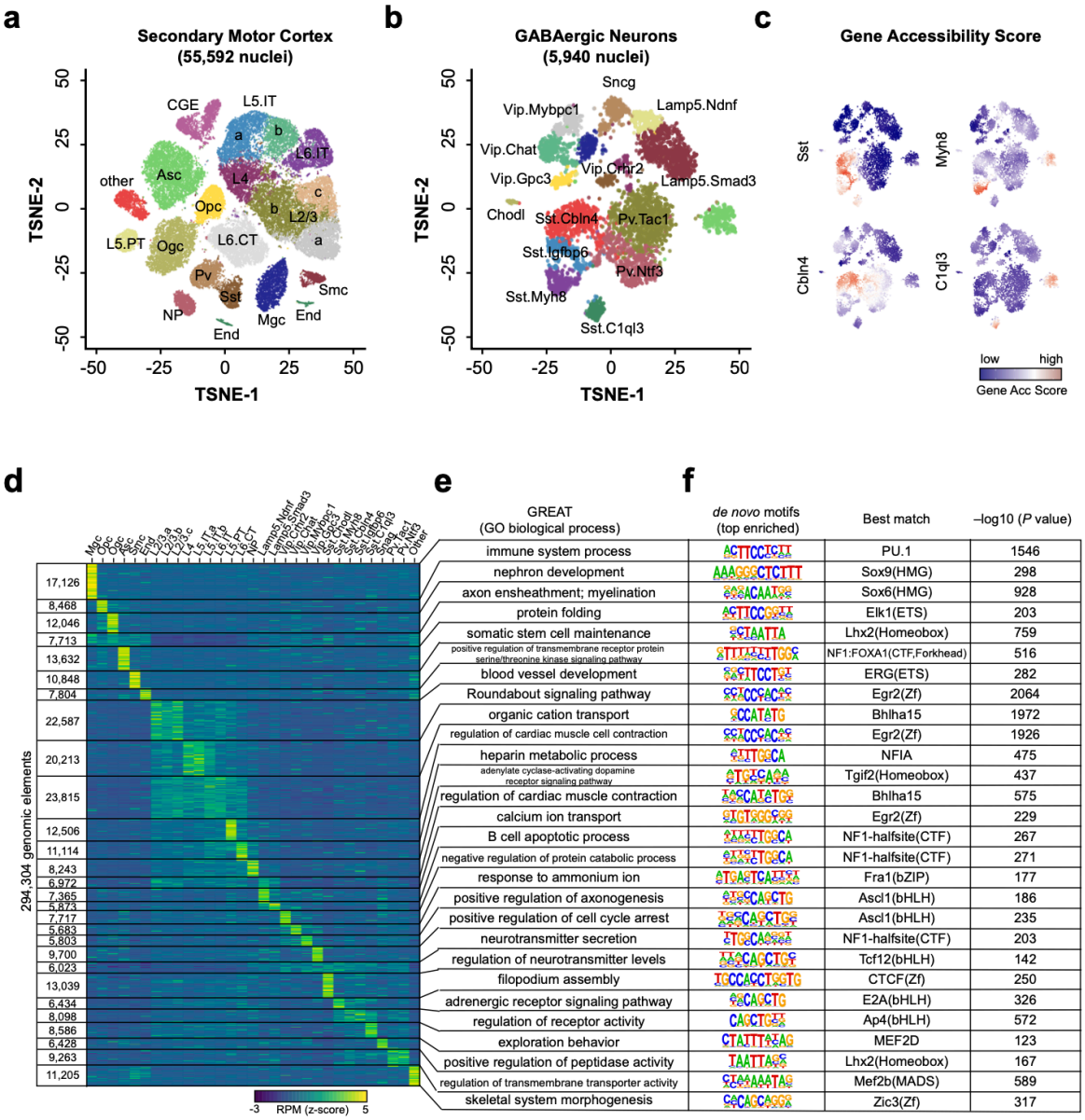
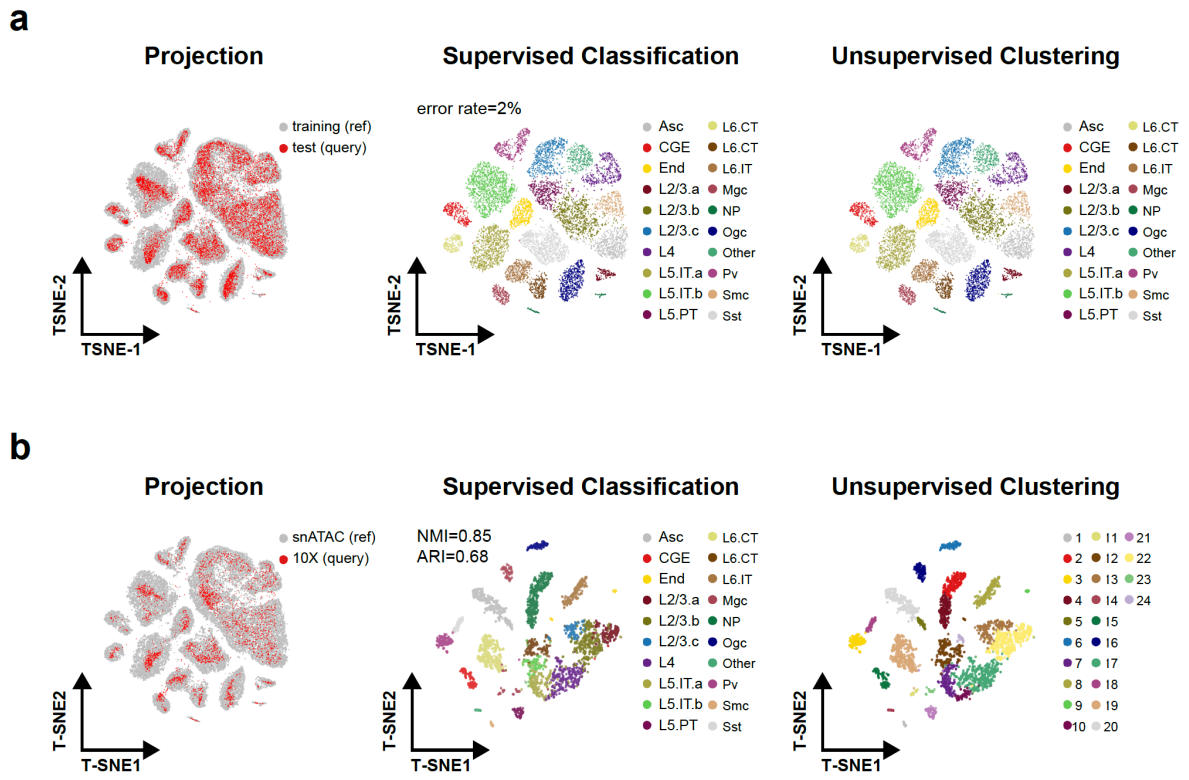


Figure 5. A high-resolution *cis*-regulatory atlas of mouse secondary motor cortex (MOs). (**a**) T-SNE visualization of 20 cell types in MOs identified using SnapATAC. (**b**) Fourteen GABAergic subtypes revealed by iterative clustering of 5,940 GABAergic neurons (Sst, Pv and CGE). (**c**) Gene accessibility score of canonical marker genes for GABAergic subtypes projected onto the t-SNE embedding. Marker genes were identified from previous scRNA-seq analysis[38]. (**d**) *k*-means clustering of 294,304 differentially accessible elements based on chromatin accessibility. (**e**) Gene ontology analysis of each cell type predicted using GREAT analysis[39]. (**f**) Transcription factor motif enriched in each cell group identified using Homer[21].

424 In mammalian brain, GABAergic interneurons exhibit spectacular diversity that shapes
425 the spatiotemporal dynamics of neural circuits underlying cognition[40]. To examine
426 whether iterative analysis could help tease out various subtypes of GABAergic neurons,
427 SnapATAC is applied to the 5,940 GABAergic nuclei (CGE, Sst and Vip) identified above,
428 finding 17 distinct sub-populations (**Supplementary Fig. 20a**) that are highly
429 reproducible between biological replicates (Pearson correlation = 0.99; **Supplementary**
430 **Fig. 20b**). Based on the chromatin accessibility at the marker genes (**Supplementary**
431 **Fig. 21**), these 17 clusters are classified into five Sst subtypes (Chodl+, Cbln4+, Igfbp6+,
432 Myh8+ and C1ql3+), two Pv subtypes (Tac1+ and Ntf3+), two Lamp5 subtypes (Smad3+
433 and Ndnf+), four Vip subtypes (Mybpc1+, Chat+, Gpc3+, Crhr2+), Sncg and putative
434 doublets (**Fig. 5b**). These clusters include a rare type Sst-Chodl (0.1%) previously
435 identified in single cell RNA[38] and single cell ATAC-seq analysis[41]. While the identity and
436 function of these subtypes require further experimental validation, our results
437 demonstrate the exquisite sensitivity of SnapATAC in resolving distinct neuronal
438 subtypes with only subtle differences in the chromatin landscape.
439
440 A key utility of single cell chromatin accessibility analysis is to identify regulatory
441 sequences in the genome. By pooling reads from nuclei in each major cluster (**Fig. 5a**),
442 cell-type specific chromatin landscapes can be obtained (**Supplementary Fig. 22** and
443 **Supplementary Methods**). Peaks are determined in each cell type, resulting in a total
444 of 373,583 unique candidate *cis*-regulatory elements. Most notably, 56%
445 (212,730/373,583) of these open chromatin regions cannot be detected from bulk ATAC-
446 seq data of the same brain region (**Supplementary Methods**). The validity of these
447 additional open chromatin regions identified from scATAC-seq data are supported by
448 several lines of evidence. First, these open chromatin regions are only accessible in minor
449 cell populations (**Supplementary Fig. 23a**) that are undetectable in the bulk ATAC-seq
450 signal. Second, these sequences show significantly higher conservation than randomly
451 selected genomic sequences with comparable mappability scores (**Supplementary Fig.**
452 **23c**). Third, these open chromatin regions display an enrichment for transcription factor
453 (TF) binding motifs corresponding to the TFs that play important regulatory roles in the
454 corresponding cell types. For example, the binding motif for Mef2c is highly enriched in
455 novel candidate *cis*-elements identified from Pvalb neuronal subtype (P-value = 1e-363;

456 **Supplementary Fig. 23d**), consistent with previous report that Mef2c is upregulated
457 in embryonic precursors of Pv interneurons[42]. Finally, the new open chromatin regions
458 tend to test positive in transgenic reporter assays. Comparison to the VISTA enhancer
459 database[43] shows that enhancer activities of 256 of the newly identified open chromatin
460 regions have been previously tested using transgenic reporter assays in e11.5 mouse
461 embryos. Sixty five percent (167/256; 65%) of them drive reproducible reporter
462 expression in at least one embryonic tissue, which was substantially higher than
463 background rates (9.7%) estimated from regions in the VISTA database that lack
464 canonical enhancer mark[44]. Four examples are displayed (**Supplementary Fig. 23e**).
465
466 SnapATAC identifies 294,304 differentially accessible elements between cell types
467 (**Supplementary Methods**; **Fig. 5d**). GREAT analysis (**Fig. 5e**) and motif inference
468 (**Fig. 5f**) identify the master regulators and transcriptional pathways active in each of the
469 cell types. For instance, the binding motif for ETS-factor PU.1 is highly enriched in
470 microglia-specific candidate CREs, motifs for SOX proteins are enriched in Ogc-specific
471 elements, and bHLH motifs are enriched in excitatory neurons-specific CREs (**Fig. 5f**).
472 Interestingly, motifs for candidate transcriptional regulators, including NUCLEAR
473 FACTOR 1 (NF1), are also enriched in candidate CREs detected in two inhibitory neuron
474 subtypes (Lamp5.Ndnf and Lamp5.Smad3). Motif for CTCF, a multifunctional protein in
475 genome organization and gene regulation[45], is highly enriched in Sst-Chodl, indicating
476 that CTCF may play a role in neurogenesis. Finally, motifs for different basic-helix-loop-
477 helix (bHLH) family transcription factors, known determinants of neural differentiation[46],
478 show enrichment for distinct Sst subtypes. For instance, E2A motif is enriched in
479 candidate CREs found in Sst.Myh8 whereas AP4 motif is specifically enriched in peaks
480 found in Sst.Cbln4, suggesting specific role that different bHLH factors might play in
481 different neuronal subtypes.
482
483 **SnapATAC enables reference-based annotation of new scATAC-seq datasets**
484 Unsupervised clustering of scATAC-seq datasets frequently requires manual annotation,
485 which is labor-intensive and limited to prior knowledge. To overcome this limitation,
486 SnapATAC provides a function to project new single cell ATAC-seq datasets to an existing
487 cell atlas to allow for supervised annotation of cells. First, the Nystrom method is used to

488 project the query cells to the low-dimension manifold pre-computed from the reference
489 cells (**Supplementary Methods**). In the joint manifold, a neighborhood-based
490 classifier is used to determine the cell type of each query cell based on the label of its *k*
491 nearest neighboring cells in the reference dataset (**Supplementary Methods**). The
492 accuracy of this method is determined by five-fold cross validation using the mouse motor
493 cortex atlas. On average, 98% (±1%) of the cells can be correctly classified, suggesting a
494 high accuracy of the method (**Fig. 6a**).
495
496 To demonstrate that SnapATAC could be applied to datasets generated from distinct
497 technical platforms, it is used to annotate 4,098 scATAC-seq profiles from mouse brain
498 cells generated using a droplet-based platform (**Supplementary Table 2**). After
499 removing batch effect introduced by different platforms using Harmony[23], the query cells
500 are well mixed with the reference cells in the joint embedding space (**Supplementary**
501 **Fig. 24**). The predicted cluster labels are also consistent with the cell types defined using
502 unbiased clustering analysis (NMI=0.85, ARI=0.68; **Fig. 6b**).
503



504

505 **Figure 6. SnapATAC enables supervised annotation of new scATAC-seq**
506 **dataset using reference cell atlas.** (**a**) MOs snATAC-seq dataset is split into 80% and
507 20% as training and test dataset. A predictive model learned from the training dataset
508 predicts cell types on the test dataset of high accuracy (error rate = 2%) as compared to
509 the original cell type labels (right). (**b**) A predictive model learned from the reference
510 dataset - MOs (snATAC) – accurately predicts the cell types on a query dataset from
511 mouse brain generated using a different technological platform, the 10X scATAC-seq. The
512 t-SNE embedding is inferred from the reference cell atlas (left) or generated by SnapATAC
513 in an unbiased manner from 10X mouse brain dataset (middle and right). Cells are
514 visualized using t-SNE and are colored by the cell types predicted by supervised
515 classification (middle) compared to the cluster labels defined using unsupervised
516 clustering (right).
517
518 To investigate whether SnapATAC could recognize cell types in the query dataset that are
519 not present in the reference atlas, multiple query data sets are sampled from the above
520 mouse motor cortex dataset and a perturbation is introduced to each sampling by
521 randomly dropping a cell cluster. When this resulting query dataset is analyzed by
522 SnapATAC against the original cell atlas, the majority of the cells that are left out from the
523 original atlas are filtered out due to the low prediction score (**Supplementary Fig. 25**),
524 again suggesting that our method is not only accurate but also robust to the novel cell
525 types in the query dataset.
526
527 **Discussion**
528 In summary, SnapATAC is a comprehensive bioinformatic solution for single cell ATAC-
529 seq analysis. The open-source software runs on standard hardware, making it accessible
530 to a broad spectrum of researchers. Through extensive benchmarking, we have
531 demonstrated that SnapATAC outperforms existing tools in sensitivity, accuracy,
532 scalability and robustness of identifying cell types in complex tissues.
533
534 SnapATAC differs from previous methods in at least seven aspects. First, SnapATAC
535 incorporates many useful tools and represents the most comprehensive solution for single
536 cell ATAC-seq data analysis to date. In addition to clustering analysis, SnapATAC
537 provides preprocessing, annotation, trajectory analysis, peak calling[26], differential
538 analysis[22], batch effect correction[23] and motif discovery[27] all in one package. Second,

539 SnapATAC identifies cell types in an unbiased manner without the need for population-
540 level peak annotation, leading to superior sensitivity for identifying rare cell types in
541 complex tissues. Third, SnapATAC utilizes a new algorithm for dimensionality reduction
542 and to identify cell types in heterogeneous tissues and map cellular trajectories. Fourth,
543 with Nyström sampling method[47], SnapATAC significantly reduces both CPU and
544 memory usage, enabling analysis of large-scale dataset of a million cells or more. Fifth,
545 SnapATAC not only incorporates existing method to integrate scATAC-seq with scRNA-
546 seq dataset[30] but also provides a new method to predict promoter-enhancer pairing
547 relations based on the statistical association between gene expression and chromatin
548 accessibility in single cells. Sixth, our method achieves high clustering reproducibility
549 using a consensus clustering approach. Finally, SnapATAC also enables supervised
550 annotation of a new scATAC-seq dataset based on an existing reference cell atlas.
551
552 It is important to note that a different strategy has been used to overcome the bias
553 introduced by population-based peak annotation[8]. This approach involves iterative
554 clustering, with the first round defining the "crude" clusters in complex tissues followed
555 by identifying peaks in these clusters, which are then used in subsequent round(s) of
556 clustering. However, several limitations still exist. First, the strategy of iterative clustering
557 requires multiple rounds of clustering, aggregation, and peak calling, thus hindering its
558 application to large-scale datasets. Second, the "crude" clusters represent the most
559 dominant cell types in the tissues; therefore, peaks in the rare populations may still be
560 underrepresented. Finally, peak-based methods hinder multi-sample integrative analysis
561 where each sample has its own unique peak reference.
562
563 Finally, SnapATAC is applied to a newly generated scATAC-seq dataset including 55,592
564 high quality single nucleus ATAC-seq profiles from the mouse secondary motor cortex,
565 resulting in a single cell atlas consisting of >370,000 candidate *cis*-regulatory elements
566 across 31 cell types in this mouse brain region. The cellular diversity identified by
567 chromatin accessibility is at an unprecedented resolution and is consistent with mouse
568 neurogenesis and taxonomy revealed by single cell transcriptome data[38,48]. Besides
569 characterizing the constituent cell types, SnapATAC identifies candidate *cis*-regulatory
570 sequences in each of the major cell types and infers the likely transcription factors that

regulate cell-type specific gene expression programs. Importantly, a large fraction (56%) of the candidate *cis*-elements identified from the scATAC-seq data are not detected in bulk analysis. While further experiments to thoroughly validate the function of these additional open chromatin regions are needed, the ability for SnapATAC to uncover *cis*-elements from rare cell types of a complex tissue will certainly help expand the catalog of *cis*-regulatory sequences in the genome.

Data availability

Raw and processed data to support the findings of this study have been deposited to NCBI Gene Expression Omnibus with the accession number GSE126724 with the token of srkxoisclpkppcd.

Code availability

The scripts and pipeline for the analysis can be found at https://github.com/r3fang/SnapATAC.

Acknowledgements

We thank D. Gorkin, R. Raviram, and J. Hocker for proofreading and suggestions for the manuscript. We thank S. Kuan for sequencing support. We thank C. Zhang and B. Li for Bioinformatics support. We thank C. O'Connor and C. Fitzpatrick at Salk Institute Flow Cytometry Core for sorting of nuclei. This study was funded by U19MH114831.

Author Contributions

This study was conceived and designed by R.F. and B.R.; Pipeline developed by R.F.; Data analysis performed by R.F.; Tissue collection and nuclei preparation performed by J.L. and M.B.; Single nucleus ATAC-seq experiment performed by S.P., X.H. and X.W.; Tn5 enzymes synthesized and provided by A.M. and A.S.; Manuscript written by R.F. and B.R. with input from all authors.

Competing Financial Interest Statement

The authors declare no competing financial interests.

601 **REFERENCE**

602 1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the
603 human genome. *Nature* **489**, 57–74 (2012).
604 2. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488**,
605 116–120 (2012).
606 3. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition
607 of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-
608 binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
609 4. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across
610 the Genome. *Cell* **132**, 311–322 (2008).
611 5. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by
612 combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
613 6. Cusanovich, D. A. *et al.* The *cis*-regulatory dynamics of embryonic development at single-
614 cell resolution. *Nature* **555**, 538–542 (2018).
615 7. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse
616 forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432 (2018).
617 8. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin
618 Accessibility. *Cell* **174**, 1309-1324.e18 (2018).
619 9. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic states in
620 the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
621 10. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory
622 variation. *Nature* **523**, 486–490 (2015).
623 11. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell resolution.
624 *Nat. Commun.* **9**, 3647 (2018).

625 12. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune
626 cell development and intratumoral T cell exhaustion. *bioRxiv* (2019) doi:10.1101/610550.
627 13. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell
628 chromatin accessibility. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0147-6.
629 14. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
630 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*
631 **14**, 975–978 (2017).
632 15. Bravo González-Blas, C. *et al.* cisTopic: cis-regulatory topic modeling on single-cell ATAC-
633 seq data. *Nat. Methods* **16**, 397–400 (2019).
634 16. de Boer, C. G. & Regev, A. BROCKMAN: deciphering variance in epigenomic regulators
635 by k-mer factorization. *BMC Bioinformatics* **19**, (2018).
636 17. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell
637 Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).
638 18. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21
639 (2019).
640 19. Kumar, S., Mohri, M. & Talwalkar, A. Sampling Methods for the Nystro¨m Method. 26.
641 20. Long, A. W. & Ferguson, A. L. Landmark Diffusion Maps (L-dMaps): Accelerated manifold
642 learning out-of-sample extension. *Appl. Comput. Harmon. Anal.* **47**, 190–211 (2019).
643 21. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime
644 cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–
645 589 (2010).

646 22. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
647 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140
648 (2010).
649 23. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony.
650 *Nat. Methods* **16**, 1289–1296 (2019).
651 24. Rubin, A. J. *et al.* Coupled Single-Cell CRISPR Screening and Epigenomic Profiling
652 Reveals Causal Gene Regulatory Networks. *Cell* **176**, 361-376.e17 (2019).
653 25. Fairfax, B. P. *et al.* Innate Immune Activity Conditions the Effect of Regulatory Variants
654 upon Monocyte Gene Expression. *Science* **343**, 1246949–1246949 (2014).
655 26. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
656 27. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime
657 cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–
658 589 (2010).
659 28. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions.
660 *Nat. Biotechnol.* **28**, 495–501 (2010).
661 29. Wu, J. *et al.* The landscape of accessible chromatin in mammalian preimplantation embryos.
662 *Nature* **534**, 652–657 (2016).
663 30. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21
664 (2019).
665 31. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals
666 Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).
667 32. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell
668 transcriptomics. *BMC Genomics* **19**, (2018).

669 33. Zhao, C., Deng, W. & Gage, F. H. Mechanisms and Functional Implications of Adult
670 Neurogenesis. *Cell* **132**, 645–660 (2008).
671 34. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord
672 with split-pool barcoding. *Science* **360**, 176–182 (2018).
673 35. Chen, H. *et al.* Assessment of computational methods for the analysis of single-cell ATAC-
674 seq data. *Genome Biol.* **20**, 241 (2019).
675 36. Ecker, J. R. *et al.* The BRAIN Initiative Cell Census Consortium: Lessons Learned toward
676 Generating a Comprehensive Brain Cell Atlas. *Neuron* **96**, 542–557 (2017).
677 37. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell
678 Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* **8**, 281-291.e9 (2019).
679 38. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell transcriptomics.
680 *Nat. Neurosci.* **19**, 335–346 (2016).
681 39. McLean, C. Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions.
682 *Nat. Biotechnol.* **28**, 495–501 (2010).
683 40. Huang, Z. J. & Paul, A. The diversity of GABAergic neurons and neural communication
684 elements. *Nat. Rev. Neurosci.* (2019) doi:10.1038/s41583-019-0195-4.
685 41. Graybuck, L. T. *et al. Prospective, brain-wide labeling of neuronal subclasses with*
686 *enhancer-driven AAVs*. http://biorxiv.org/lookup/doi/10.1101/525014 (2019)
687 doi:10.1101/525014.
688 42. Mayer, C. *et al.* Developmental diversification of cortical inhibitory interneurons. *Nature*
689 **555**, 457–462 (2018).
690 43. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a
691 database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).

692 44. Gorkin, D. *et al.* Systematic mapping of chromatin state landscapes during mouse
693 development. *bioRxiv* 166652 (2017) doi:10.1101/166652.
694 45. Phillips, J. E. & Corces, V. G. CTCF: Master Weaver of the Genome. *Cell* **137**, 1194–1211
695 (2009).
696 46. Kageyama, R., Ishibashi, M., Takebayashi, K. & Tomita, K. bHLH Transcription factors and
697 mammalian neuronal differentiation. *Int. J. Biochem. Cell Biol.* **29**, 1389–1399 (1997).
698 47. Erichson, N. B., Mathelin, L., Brunton, S. L. & Kutz, J. N. Diffusion Maps meet Nystr\"om.
699 *ArXiv180208762 Cs Stat* (2018).
700 48. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-
701 1014.e22 (2018).
702 49. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's
703 guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22 (2020).
704 50. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of
705 communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
706 51. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq
707 data using regularized negative binomial regression. *bioRxiv* (2019) doi:10.1101/576827.
708 52. Li, M., Kwok, J. T. & Lu, B.-L. Making Large-Scale Nyström Approximation Possible. 12.
709 53. Kumar, S., Mohri, M. & Talwalkar, A. Ensemble Nystrom Method. 9.
710 54. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory elements in
711 mammalian cortex. *Science* **357**, 600–604 (2017).
712 55. Lister, R. *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development.
713 *Science* **341**, 1237905 (2013).

714 56. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled
715 sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
716 57. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
717 58. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals
718 Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
719

720 **Supplementary Materials**

721

722 **SnapATAC: A Comprehensive Analysis Package for Single Cell ATAC-seq**

723

724 Rongxin Fang[1,2], Sebastian Preissl[3], Yang Li[2], Xiaomeng Hou[3], Jacinta Lucero[4], Xinxin
725 Wang[3], Amir Motamedi[5], Andrew K. Shiau[5], Xinzhu Zhou[6], Fangming Xie[7], Eran A.
726 Mukamel[7], Kai Zhang[2], Yanxiao Zhang[2], M. Margarita Behrens[4], Joseph R. Ecker[4,8], and
727 Bing Ren[2,3,9*]

728

729 *Correspondence to: biren@ucsd.edu

730

731

732 **Outline of the SnapATAC Pipeline**

733 *Barcode Demultiplexing*

734 Using a custom python script, we first de-multicomplex FASTQ files by integrating the
735 cell barcode into the read name in the following format:

736

737 "@"+"barcode"+":"+"original_read_name".

738

739 *Alignment & sorting*

740 Demultiplexed reads are aligned to the corresponding reference genome (i.e. mm10 or
741 hg19) using bwa (0.7.13-r1126) in pair-end mode with default parameter settings. Aligned
742 reads are then sorted based on the read name using samtools (v1.9) to group together
743 reads originating from the same barcodes.

744

745 *Quality Control & Filtering*

746 Pair-end reads are converted into fragments and only those that meet the following
747 criteria are kept: 1) properly paired (according to SMA flag value); 2) uniquely mapped
748 (MAPQ > 30); 3) insert distance within [50-1000bp]. PCR duplicates (fragments sharing
749 exactly the same genomic coordinates) are removed for each cell separately. Given that
750 Tn5 introduces a 9 bp staggered, reads mapping to the positive and negative strand were
751 shifted by +4 / -5bp respectively[49].

752

753 We identify the high-quality cells based on two criteria: 1) total number of unique
754 fragment count [>1,000]; 2) fragments in promoter ratio – the percentage of fragments
755 overlapping with annotated promoter regions [0.2-0.8]. The promoter regions used in
756 this study are downloaded from 10X genomics for hg19 and mm10.

757

758 *Snap File Generation*

759 Using the remaining fragments, we next generate a snap-format (Single-Nucleus
760 Accessibility Profiles) file using snaptools (https://github.com/r3fang/SnapTools). A
761 snap file is a hierarchically structured hdf5 file that contains the following sections:
762 header (HD), cell-by-bin matrix (BM), cell-by-peak matrix (PM), cell-by-gene matrix
763 (GM), barcode (BD) and fragment (FM). HD session contains snap-file version, date,

764 alignment and reference genome information. BD session contains all unique barcodes
765 and corresponding meta data. BM session contains cell-by-bin matrices of different
766 resolutions. PM session contains cell-by-peak count matrix. GM session contains cell-by-
767 gene count matrix. FM session contains all usable fragments for each cell. Fragments are
768 indexed based on barcodes that enables fast retrieval of reads based on the barcodes.
769 Detailed information about snap file can be found in **Supplementary Note 1**.
770

Box1. Generating Snap file using snaptools

```
snaptools snap-pre
        --input-file=demo.srt.bed.gz
        --output-snap=demo.snap
        --genome-name=mm10
        --genome-size=mm10.gs
        --min-mapq=30
        --min-flen=50
        --max-flen=1000
        --keep-single=False
        --keep-secondary=False
        --keep-discordant=False
        --min-cov=0
        --max-num=20000
        --keep-chrm=True
        --overwrite=True
```

771
772 One major utility of the snap file and snaptools is to retrieve reads belonging to a certain
773 group of barcodes. This can be done using snaptools with following command where
774 "barcodes.sel.txt" is a text file that contains the selected barcodes.
775

Box2. Extracting reads using SnapTools

```
snaptools dump-fragment
        --snap-file=demo.snap
```

```
--barcode-file=barcodes.sel.txt
--output-file=demo.sel.bed.gz
```

Creating Cell-by-Bin Count Matrix

Using the resulting snap file, we next create cell-by-bin count matrix. The genome is segmented into uniform-sized bins and single cell ATAC-seq profiles are represented as cell-by-bin matrix with each element indicating number of sequencing fragments overlapping with a given bin in a certain cell. In the below example, a cell-by-bin matrix of 5kb resolution is added to demo.snap file.

Box 3. Generating cell-by-bin matrix using SnapTools
snaptools snap-add-bmat --snap-file=demo.snap --bin-size-list 5000

Optimizing the Bin Size

To evaluate the effect of bin size to clustering performance, we apply SnapATAC to three datasets namely 5K PBMC (10X), Mouse Brain (10X) and MOs-M1 (snATAC). These datasets are generated by both plate and droplet platforms using either cell or nuclei with considerably different depth, allowing us to systematically evaluate the effect of bin size.

For each dataset, we first define the "landmark" cell types in a supervised manner. First, we perform cisTopic[15] for dimensionality reduction and identify cell clusters using graph-based algorithm Louvain[50] with k=15. Second, we manually define the major cell types in each dataset by examining the gene accessibility score at the canonical marker genes (**see Supplementary Fig. 9** as an example for MOs-M1). Third, clusters sharing the same marker genes are manually merged and those failing to show unique signatures are discarded. In total, we define nine cell types in PBMC 5K (10X), 14 types in Mouse Brain 5K (10X) and 14 types in MOs M1 (snATAC). Among these cell types, 14 cell populations that account for less than 2% of the total population are considered as rare cell populations (**Supplementary Fig. 2a**).

802 We next evaluate the performance of each bin size selection using three metrics: 1) cluster
803 connectivity index (CI) which estimate the degree of connectedness of the landmark cell
804 types; a lower CI represents a better separation. The connectivity index is computed in
805 the following manner. For each cell *i*, the K (K=15) nearest neighbors are found and sorted
806 from the closest to furthest. The algorithm checks if those neighbors are assigned to the
807 same cluster with cell *i*. At the beginning connectivity value is equal 0 and increase with
808 value $1/i$ when the *i*-th nearest neighbors is not assigned to the same cluster with cell *i*.
809 This procedure is repeated for all cells in the dataset. In general, the higher the
810 connectivity index is, the less separated the defined landmarks are. The connectivity index
811 is computed using "connectivity" function implemented in R package clv. 2) coverage bias
812 which estimates the read depth distribution in the two-dimensional embedding space; 3)
813 sensitivity to identify rare populations. Through systematic benchmarking, we found that
814 bin size in the range from 1kb to 10kb appeared to work well on the three benchmarks, we
815 selected 5kb as the default bin width for all the analysis in this work (**Supplementary**
816 **Methods** and **Supplementary Fig. 2**).
817
818 *Matrix Binarization*
819 We found that the vast majority of the elements in the cell-by-bin count matrix is "0",
820 indicating either closed chromatin or missing value. Among the non-zero elements, some
821 has abnormally high coverage (> 200) perhaps due to the alignment errors. These items
822 usually account for less than 0.1% of total non-zero items in the matrix. Thus, we change
823 the top 0.1% elements in the matrix to "0" to eliminate potential alignment errors. We
824 next convert the remaining non-zero elements to "1".
825
826 *Bin Filtering*
827 We next filter out any bins overlapping with the ENCODE blacklist downloaded from
828 http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/. Second, we remove
829 reads mapped to the X/Y chromosomes and mitochondrial DNA. We sort the bins based
830 on the coverage and filter out the top 5% to remove the invariant features. Please note
831 that we do not perform coverage-based bin filtering for a dataset that has low coverage

(average fragment number less than 5,000) where the ranking of bin may be fluctuated by the noise.

Dimensionality Reduction

We next apply the following dimensionality reduction method to project the high-dimension data to a low-dimension manifold for clustering and visualization. Now, let us express the algorithm in matrix notation. Let $\mathbf{X} \in \mathcal{R}^{n \times m}$ be a dataset with $\boldsymbol{n}$ cells and $\boldsymbol{m}$ bins and $\boldsymbol{X} = \{\mathbf{0}, \mathbf{1}\}$. The first step is to compute a similarity matrix between the $\boldsymbol{m}$ high-dimensional data points to construct the $\boldsymbol{n}$-by-$\boldsymbol{n}$ pairwise similarity matrix using a kernel function $\boldsymbol{k}$ that is an appropriate similarity metric. A popular choice is gaussian kernel:

$$\boldsymbol{k}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{exp}(-\frac{||\boldsymbol{x_i} - \boldsymbol{x_j}||}{\boldsymbol{\epsilon}})$$

where $||.||$ is a square root of Euclidean distance between observations $\boldsymbol{i}$ and $\boldsymbol{j}$.

Due the binarization nature of single cell ATAC-seq dataset, in this case, we replace the Gaussian kernel with Jaccard coefficient which estimates the similarity between cells simply based on ratio of overlap over the total union:

$$\boldsymbol{jaccard}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \frac{|\boldsymbol{x_i} \cap \boldsymbol{x_j}|}{|\boldsymbol{x_i} \cup \boldsymbol{x_j}|}$$

For instance, given two cells $\boldsymbol{x_i} = \{\mathbf{0}, \mathbf{1}, \mathbf{1}, \mathbf{0}\}$ and $\boldsymbol{x_j} = \{\mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{1}\}$, the Jaccard coefficient is $\boldsymbol{jaccard}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \mathbf{1/4}$. The Jaccard coefficient has the following properties that meet the requirement of being a kernel function:

$$\boldsymbol{jaccard}(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{jaccard}(\boldsymbol{x_j}, \boldsymbol{x_i}) \text{ (symmetric)}$$

$$\boldsymbol{jaccard}(\boldsymbol{x_i}, \boldsymbol{x_j}) \geq \mathbf{0} \text{ (positivity preserving)}$$

Using $\boldsymbol{jaccard}$ as a kernel function, we next form a symmetric kernel matrix $\boldsymbol{J} \in \boldsymbol{\mathcal{R}}^{n \times n}$ where each entry is obtained as $\boldsymbol{J_{i,j}} = \boldsymbol{jaccard}(\boldsymbol{x_i}, \boldsymbol{x_j})$

Theoretically, the similarity $\boldsymbol{J_{i,j}}$ would reflect the true similarity between cell $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. Unfortunately, due to the high-dropout rate, this is not the case. If there is a high sequencing depth for cell $\boldsymbol{x_i}$ or $\boldsymbol{x_j}$, then $\boldsymbol{J_{i,j}}$ tend to have higher values, regardless whether cell $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ is actually similar or not.

This can be proved theoretically. Given 2 cells $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ and corresponding coverage (number of "1"s) $\boldsymbol{C_i} = \sum_k^m \boldsymbol{x_{ik}}$ and $\boldsymbol{C_j} = \sum_k^m \boldsymbol{x_{jk}}$, let $\boldsymbol{P_i} = \boldsymbol{C_i}/\boldsymbol{m}$ and $\boldsymbol{P_j} = \boldsymbol{C_j}/\boldsymbol{m}$ be the probability of observing a signal in cell $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ where $\boldsymbol{m}$ is the length of the vector. Assuming $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are two "random" cells without any biological relevance, in another word, the "1"s in $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are randomly distributed, then the expected Jaccard index between cell $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ can be calculated simply as:

$$\boldsymbol{E_{ij}} = \frac{\boldsymbol{P_i P_j}}{\boldsymbol{P_i} + \boldsymbol{P_j} - \boldsymbol{P_i P_j}}$$

Because $\boldsymbol{P_i} \times \boldsymbol{P_j} > \boldsymbol{0}$ (no empty cells allowed), then

$$\boldsymbol{E_{ij}} = \frac{\boldsymbol{1}}{(\boldsymbol{1}/\boldsymbol{P_i} + \boldsymbol{1}/\boldsymbol{P_j} - \boldsymbol{1})}$$

The increase of either $\boldsymbol{P_i}$ or $\boldsymbol{P_j}$ will result in an increase of $\boldsymbol{E_{ij}}$ which suggests the Jaccard similarity between cells is highly affected by the read depth. Such observation prompts us to develop an *ad hoc* normalization method to eliminate the read depth effect.

To learn the relationship between the $\boldsymbol{E_{ij}}$ and $\boldsymbol{J_{ij}}$ from the data, we next fit a curve to predict the observed Jaccard coefficient $\boldsymbol{J_{ij}}$ as a function of its expected value $\boldsymbol{E_{ij}}$ by fitting a polynomials regression of degree 2 using R function lm. Theoretically, $\boldsymbol{E_{ij}}$ should be

linear with J_{ij} if cells are completely random, but in real dataset, we have observed a non-linearity between E_{ij} and J_{ij} especially among the high-coverage cells. We suspect, to some extent, the degree of randomness of fragment distribution in a single cell is associated with the coverage. To better model the non-linearity, we include a second order polynomial in our model:

$$J_{ij} = \beta_0 + \beta_1 E_{ij} + \beta_2 E_{ij}^2$$

This fitting provided estimators of parameters $\{\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}\}$. As such, we next use it to normalize the observed Jaccard coefficient by:

$$N_{ij} = J_{ij} / (\widehat{\beta_0} + \widehat{\beta_1} E_{ij} + \widehat{\beta_2} E_{ij}^2)$$

The fitting of the linear regression, however, can be very time consuming with a large matrix. Here we test the possibility of performing this step on a random subset of y cells in lieu of the full matrix. When selecting a subset of y cells to speed up the first step, we do not select cells at random with a uniform sampling probability. Instead, we set the probability of selecting a cell i to

$$\frac{1}{d(log_{10}(C_i))}$$

where d is the density estimate of all log10-transformed cell fragment count and C_i is the number of fragments in cell i and $C_i = \sum_k^m x_{ik}$. Similar approach was first introduced in SCTranscform[51] to speed up the normalization of single cell RNA-seq.

We then proceed to normalize the full Jaccard coefficient matrix $J \in \mathcal{R}^{n \times n}$ using the regression model learned from y cells and compared the results to the case where all cells are used in the initial estimation step as well. We use the correlation of normalized Jaccard coefficient to compare this partial analysis to the full analysis. We observe that

using as few as 2000 cells in the estimation gave rise to virtually identical estimates. We therefore use 2,000 cells in the initial model-fitting step. To remove outliers in the normalized similarity, we use the 0.99 quantile to cap the maximum value of the normalized matrix.

Next, using normalized Jaccard coefficient matrix $\boldsymbol{N}$, we next normalize the matrix by:

$$\boldsymbol{A} = \boldsymbol{D}^{-1/2}\boldsymbol{N}\boldsymbol{D}^{-1/2}$$

where $\boldsymbol{D} \in \mathcal{R}^{n\times n}$ is a diagonal matrix which is composed as $\boldsymbol{D}_{i,i} = \sum_j \boldsymbol{N}_{i,j}$. We next perform eigenvector decomposition against $\boldsymbol{A}$.

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$$

The columns $\boldsymbol{\varphi}_i \in \mathcal{R}^n$ of $\boldsymbol{U} \in \mathcal{R}^{n\times n}$ are the eigenvectors. The diagonal matrix $\boldsymbol{\Lambda} \in \mathcal{R}^{n\times n}$ has the eigenvalues $\boldsymbol{\lambda}_1 \geq \boldsymbol{\lambda}_2 \geq \cdots \geq \boldsymbol{0}$ in descending order as its entries. Finally, we report the first $\boldsymbol{r}$ eigenvectors as the final low-dimension manifold.

Evaluation of Ad Hoc Normalization Method

To assess the performance of normalization of SnapATAC we processed three datasets. As shown in **Supplementary Fig. 3**, before normalization, SnapATAC exhibits a strong gradient that is correlated with sequencing depth within the cluster (**Supplementary Fig. 3a**). Although the sequencing depth effect is still observed in some of the small clusters, it is clear that the normalization method has largely eliminated the read depth effect as compared to the unnormalized ones (**Supplementary Fig. 3b**).

To better quality the coverage bias, we next computed the Shannon entropy that estimates the "uniformness" of the distribution of cell coverage in the UMAP embedding space. In detail, we first chose the top 10% cells of the highest coverage as "high-coverage" cells. Second, in the 2D UMAP embedding space, we discretize "high-coverage" cells from a continuous random coordinate (umap1, umap2) into bins (n=50) and returns the

948 corresponding vector of counts. This is done using a function called "discretize2d" in the
949 "entropy" R package. Third, we estimated the Shannon entropy of the random variable
950 from the corresponding observed counts. This is done using function "entropy" in the
951 "entropy" R package. A higher entropy indicates that the "high-coverage" cells are more
952 uniformly distributed in the UMAP embedding space, overall suggesting a better
953 normalization performance.
954
955 We next examine another eight possible sources of biases by projecting to the UMAP
956 embedding space, some metrics show cluster specificity for all three methods perhaps due
957 to biological relevance, but all three methods can reveal significant biological
958 heterogeneity without exhibiting substantial intra-cluster bias for any metrics examined
959 (**Supplementary Fig. 4**).
960
961 *Removing batch effects using Harmony*
962 When the technical variability is at a larger scale than the biological variability, we apply
963 batch effect corrector - Harmony[23] - to eliminate such confounding factor. Given two
964 datasets $\mathbf{X} = \{\mathbf{X^1}, \mathbf{X^2}\}$ generated using different technologies, we first calculate the joint
965 low-dimension manifold $\boldsymbol{U} = \{\mathbf{U^1}, \mathbf{U^2}\}$ as described above. We next apply Harmony to $\boldsymbol{U}$
966 to regress out batch effect, resulting in a new harmonized embedding $\boldsymbol{U^H}$. This is
967 implemented as a function "runHarmony" in SnapATAC package.
968
969 *Selection of Eigenvector and Eigenvalues*
970 We next determine how many eigenvectors to include for the downstream analysis. Here
971 we use an *ad hoc* approach for choosing the optimal number of components. We look at
972 the scatter plot between every two pairs of eigenvectors and choose the number of
973 eigenvectors that start exhibiting "blob"-like structure in which no obvious biological
974 structure is revealed.
975
976 *Nyström Landmark-Extension*
977 The computational cost of the dimensionality reduction scales quadratically with the
978 increase of number of cells. For instance, calculating and normalizing the pair-wise kernel
979 Matrix $\boldsymbol{N}$ becomes computationally infeasible for large-scale dataset. To overcome this

limitation, here we combine the Nyström method[19,52] (a sampling technique) and our dimensionality reduction method to present Nyström landmark-extension method.

A Nyström landmark-extension algorithm includes three major steps: i) sampling $\boldsymbol{O}(\boldsymbol{K})$: sample a subset of $\boldsymbol{K}$ ($\boldsymbol{K} \ll \boldsymbol{N}$) cells from $\boldsymbol{N}$ total cells as "landmarks". Instead of random sampling, here we adopt a density-based sampling approach developed in SCTransform[51] to preserve the density distribution of the $\boldsymbol{N}$ original points; ii) embedding $\boldsymbol{O}(\boldsymbol{K}^2)$: compute the low-dimension embedding for $\boldsymbol{K}$ landmarks; iii) extension $\boldsymbol{O}(\boldsymbol{N}-\boldsymbol{K})$: project the remaining $\boldsymbol{N}-\boldsymbol{K}$ cells onto the low-dimensional embedding as learned from the landmarks to create a joint embedding space for all cells.

This approach significantly reduces the computational complexity and memory usage given that $\boldsymbol{K}$ is considerably smaller than $\boldsymbol{N}$. The out-of-sample extension (step iii) further enables projection of new single cell ATAC-seq datasets to the existing reference single cell atlas. This allows us to further develop a supervised approach to predict cell types of a new single cell ATAC-seq dataset based on an existing reference atlas.

A key aspect of this method is the procedure according to which cells are sampled as landmark cells, because different sampled landmark cells give different approximations of the original embedding using full matrix. Here we employ the density-based sampling as described above which preserves the density distribution of the original points.

Let $\boldsymbol{X} \in \mathcal{R}^{n\times m}$ be a dataset with $\boldsymbol{n}$ cells and $\boldsymbol{m}$ variables (bins) and $\boldsymbol{N} \in \mathcal{R}^{n\times n}$ be a symmetric kernel matrix calculated using normalized Jaccard coefficient. To avoid calculating the pairwise kernel matrix and performing eigen-decomposition against a big matrix $\boldsymbol{N} \in \mathcal{R}^{n\times n}$, we first sample $\boldsymbol{k}$ ($\boldsymbol{k} \ll \boldsymbol{n}$) landmarks without replacement. This breaks down the original kernel matrix $\boldsymbol{N} \in \mathcal{R}^{n\times n}$ into four components.

$$\boldsymbol{N} = \begin{pmatrix} \boldsymbol{N}^{kk} & \boldsymbol{N}^{kv} \\ \boldsymbol{N}^{vk} & \boldsymbol{N}^{vv} \end{pmatrix}$$

in which $\boldsymbol{N^{kk}} \in \boldsymbol{\mathcal{R}^{k\times k}}$ is the pairwise kernel matrix between $\boldsymbol{k}$ landmarks and $\boldsymbol{N^{vk}} \in \boldsymbol{\mathcal{R}^{(n-k)\times k}}$ is the similarity matrix between $(\boldsymbol{n}-\boldsymbol{k})$ cells and $\boldsymbol{k}$ landmarks. Using $\boldsymbol{N^{kk}}$, we perform dimensionality reduction to obtain the $\boldsymbol{r}$-rank manifold $\boldsymbol{U^{kk}} \in \boldsymbol{\mathcal{R}^{k\times r}}$ as described above.

Using $\boldsymbol{N^{vk}}$ which estimates the similarity between $\boldsymbol{n}-\boldsymbol{k}$ cells and $\boldsymbol{k}$ landmark cells, we project the rest of $\boldsymbol{n}-\boldsymbol{k}$ cells to the embedding previously obtained using $\boldsymbol{k}$ landmark:

$$\boldsymbol{A^{vk}} = (\boldsymbol{D^{vv}})^{-\frac{1}{2}}(\boldsymbol{N^{vk}})(\boldsymbol{D^{kk}})^{-\frac{1}{2}}$$

where $\boldsymbol{D^{vv}} \in \boldsymbol{\mathcal{R}^{(n-k)\times(n-k)}}$ is a diagonal matrix which is composed as $\boldsymbol{D^{vk}_{i,i}} = \sum_j \boldsymbol{N^{vk}_{i,j}}$. The projected coordinates of the new points onto the r-dimensional intrinsic manifold defined by the landmarks are then given by,

$$\boldsymbol{U^{vk}} = \boldsymbol{A^{vk}}\boldsymbol{U^{kk}}/\boldsymbol{\Lambda^{kk}}$$

The resulting $\boldsymbol{U^{vk}} \in \boldsymbol{\mathcal{R}^{(n-k)\times r}}$ is the approximate $\boldsymbol{r}$-rank low dimension representation of the rest $\boldsymbol{n}-\boldsymbol{k}$ cells. Combing $\boldsymbol{U^{kk}}$ and $\boldsymbol{U^{vk}}$creates a joint embedding space for all cells:

$$\boldsymbol{\widetilde{U}} = \begin{bmatrix} \boldsymbol{U^{kk}} \\ \boldsymbol{U^{vk}} \end{bmatrix}$$

In the approximate joint $\boldsymbol{r}$-rank embedding space $\boldsymbol{\widetilde{U}}$, we next create a k-nearest neighbor (KNN) graph in which every cell is represented as a node and edges are drawn between cells within k nearest neighbors defined using Euclidean distance. Finally, we apply community finding algorithm such as Louvain (implemented by igraph package in R) to identify the 'communities' in the resulting graph which represents groups of cells sharing similar profiles, potentially originating from the same cell type.

Optimizing the Number of Landmarks

1040 To evaluate the effect of the number of landmarks, we apply our method to a complex
1041 dataset that contains over 80k cells from 13 different mouse tissues. We employ the
1042 following three metrics to evaluate the performance. First, using different number of
1043 landmarks (k) ranging from 1,000 to 10,000, we compare the clustering outcome to the
1044 cell type label defined in the original study. The goal of this is to identify the "elbow" point
1045 that performance drops abruptly. Second, for each sampling, we repeat for five times
1046 using different set of landmarks to evaluate stability between sampling. Third, we spiked
1047 in 1% Patski cells to assess the sensitivity of identifying rare cell types. We choose Patski
1048 cells because these cells were profiled using the same protocol by the same group (Data
1049 source listed in **Supplementary Table S1**) to minimize the batch effect.
1050
1051 We observe that using as few as 5,000 landmarks can largely recapitulate the result
1052 obtained using 10,000 landmarks (**Supplementary Fig. 5a**), and 10,000 landmarks
1053 can achieve highly robust embedding between sampling (**Supplementary Fig. 5b**) and
1054 successfully recover spiked-in rare populations (**Supplementary Fig. 5c**). To obtain a
1055 reliable low-dimensional embedding, we use 10,000 landmarks for all the analysis
1056 performed in this study.
1057
1058 *Ensemble Nyström Method*
1059 Nyström method is stochastic in its nature, different sampling will result in different
1060 embedding and clustering outcome. To improve the robustness of the clustering method,
1061 we next employ Ensemble Nyström Algorithm which combines a mixture of Nyström
1062 approximation to create an ensemble representation[53]. Supported by theoretical analysis,
1063 this Ensemble approach has been demonstrated to guarantee a convergence and in a
1064 faster rate in comparison to standard Nyström method[53]. Moreover, this ensemble
1065 algorithm naturally fits within distributed computing environments, where their
1066 computational costs are roughly the same as that of the standard Nyström single sampling
1067 method.
1068

We treat each approximation generated by the Nyström method using $\boldsymbol{k}$ landmarks as an expert and combined $\boldsymbol{p} \geq 1$ such experts to derive an improved approximation, typically more accurate than any of the original experts[53].

The ensemble set-up is defined as follows. Given a dataset $\boldsymbol{X} \in \mathcal{R}^{n \times m}$ of $\boldsymbol{n}$ cells. Each expert $\boldsymbol{S_j}$ receives $\boldsymbol{k}$ landmarks randomly selected from matrix $\boldsymbol{X}$ using density-based sampling approach without replacement. Each expert $\boldsymbol{S_r}, \boldsymbol{r} \in [\mathbf{1}, \boldsymbol{p}]$ is then used to define the low dimension embedding $\widetilde{\boldsymbol{U}}_{\boldsymbol{j}} \in \mathcal{R}^{n \times r}$ as described above. For each low-dimension embedding $\widetilde{\boldsymbol{U}}_{\boldsymbol{j}} \in \mathcal{R}^{n \times r}$, we create a KNN-graph as $\widetilde{\boldsymbol{G}}_{\boldsymbol{j}}$. Thus, the general form of the approximation, $\widetilde{\boldsymbol{G}}^{\boldsymbol{en}}$, generated by the ensemble Nyström method is

$$\widetilde{\boldsymbol{G}}^{\boldsymbol{en}} = \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{p}} \boldsymbol{\mu}^{\boldsymbol{j}} \widetilde{\boldsymbol{G}}^{\boldsymbol{j}}$$

where $\boldsymbol{\mu}^{\boldsymbol{j}}$ is the mixture weights that can be defined in many ways. Here we choose to use the most straightforward method by assigning an equal weight to each of the KNN-graph obtained from different samplings, $\boldsymbol{\mu}^{\boldsymbol{j}} = \mathbf{1}/\boldsymbol{p}, \boldsymbol{r} \in [\mathbf{1}, \boldsymbol{p}]$. While this choice ignores the relative quality of each Nyström approximation, it is computational efficient and already generates a solution superior to any one of the approximations used in the combination. Using the ensemble weighted KNN graph $\widetilde{\boldsymbol{G}}^{\boldsymbol{en}}$, we next apply community finding algorithm to identify cell clusters. By testing on the mouse atlas dataset[8], we demonstrate that the clustering stability of the ensemble approach is significantly higher than the standard Nystrom method (**Supplementary Fig. 5d**).

Visualization

We use the t-SNE implemented by FI-tsne, Rtsne or UMAP (umap_0.2.0.0) to visualize and explore the dataset.

Gene Accessibility Score

To annotate the identified clusters, SnapATAC calculated the gene-body accessibility matrix $\boldsymbol{G}$ using "calGmatFromMat" function in SnapATAC packge where $\boldsymbol{G_{i,j}}$ is the

number of fragments overlapping with **j**-th genes in **i**-th cell. $\boldsymbol{G_{i,j}}$ is then normalized to CPM (count-per-million reads) as $\widetilde{\boldsymbol{G}}$. The normalized accessibility score is then smoothed using Markov affinity-graph based method:

$$\widehat{\boldsymbol{G}}=\widetilde{\boldsymbol{G}}\boldsymbol{A}^{\boldsymbol{t}}$$

where $\boldsymbol{A}$ is the adjacent matrix obtained from K nearest neighbor graph and $\boldsymbol{t}$ is number of steps taken for Markov diffusion process. We set $\boldsymbol{t} = \mathbf{3}$ in this study. Please note that the gene accessibility score is only used to guide the annotation of cell clusters identified using cell-by-bin matrix. The clusters are identified using cell-by-bin matrix in prior.

Read Aggregation & Peak Calling

After annotation, cells from the same cluster are pooled to create aggregated signal for each of the identified cell types. This allows for identifying *cis*-elements from each cluster. MACS2 (version 2.1.2) is used for generating signal tracks and peak calling with the following parameters: --nomodel --shift 100 --ext 200 --qval 1e-2 -B –SPMR. This can be done by "runMACS" function in SnapATAC package.

Motif Analysis

SnapATAC incorporates chromVAR[14] to estimate the motif variability and Homer[21] for *de novo* motif discovery. This is implemented as function "runChromVAR" and "runHomer" in SnapATAC package.

Identification of differentially accessible peaks

For a given group of cells C_i, we first look for their neighboring cells C_j ($|C_i| = |C_j|$) in the low-dimension manifold as "background" cells to compare to. If C_i accounts for more than half of the total cells, we use the remaining cells as local background. Next, we aggregate C_i and C_j to create two raw-count vectors as V_{ci} and V_{cj} .We then perform differential analysis between V_{ci} and V_{cj} using exact test as implemented in R package edgeR (v3.18.1) with BCV=0.1. P-value is then adjusted into False Discovery Rate (FDR) using Benjamini-Hochberg correction. Peaks with FDR less than 0.01 are selected as

1130 significant DARs. However, the statically significance is under powered for small
1131 clusters.
1132
1133 *GREAT analysis*
1134 SnapATAC incorporates GREAT analysis[39] to infer the candidate biological pathway
1135 active in each cell populations. This is implemented as function "runGREAT" SnapATAC
1136 package.
1137
1138 **Integration with single cell RNA-seq**
1139 We use canonical correlation analysis (CCA) embedded in Seurat V3[18] to integrate
1140 single cell RNA-seq and single cell ATAC-seq. We first calculate the gene accessibility
1141 account at variable genes identified using single cell RNA-seq dataset. This can be done
1142 using a function called "createGmatFromMat" in SnapATAC package. Next, SnapATAC
1143 converts the snap object to a Seurat v3 object using a function called "SnapToSeurat"
1144 in preparation for integration. Different from integration method in Seurat, we use the
1145 our low-dimension manifold as the dimensionality reduction method in the Seurat
1146 object. We next follow the vignette in Seurat website
1147 (https://satijalab.org/seurat/v3.0/atacseq_integration_vignette.html) to integrate
1148 these two modalities. The cell type for scATAC-seq is predicted using function
1149 "TransferData" in Seurat V3.
1150
1151 Finally, for each single cell ATAC profile, we infer its gene expression profile by
1152 calculating the weighted average expression profile of its nearest neighboring cells in
1153 the single cell RNA-seq dataset[18]. By doing so, we create pseudo-cells that contain
1154 information of both chromatin accessibility and gene expression profiles. The
1155 imputation of gene expression profile is done by "TransferData" function in Seurat V3.
1156
1157 **Linking enhancers to putative target genes**
1158 Using the "pseudo" cells, we next sought to predict the putative target genes for regulatory
1159 elements based on the association between expression of a gene and chromatin
1160 accessibility at its enhancer elements. Given a gene $\boldsymbol{G}$, we first identify its surrounding

1161 regulatory elements within 1MB window flanking $\boldsymbol{G}$. Let $\boldsymbol{Y^G}$ be the imputed gene
1162 expression value for gene $\boldsymbol{G}$ among $\boldsymbol{n}$ cells. We perform logistic regression using $\boldsymbol{Y^G}$ as
1163 variable to predict the binary state for each of peaks surrounding $\boldsymbol{G}$. The idea behind using
1164 logistic regression is that if there is a relationship between the gene expression (continuous
1165 variable) and chromatin accessibility (categorical variable), we should be able to predict
1166 chromatin accessibility from the gene expression. Logistic regression does not make many
1167 of the key assumptions such as normality of the continuous variables. In addition, since
1168 we only have one variable (gene expression) for prediction every time, there is no problem
1169 of multicollinearity.
1170
1171 We next fit logistic regression between each of flanking peak and gene expression using
1172 "glm" function in R with binomial(link='logit') as the family function. By doing so, we
1173 obtain the regression coefficient $\boldsymbol{\beta_1}$ and its corresponding P-value for each peak
1174 separately. Here we used 5e-8, a standard P-value cutoff for human genome-wise
1175 association study to determine the significant association. While this cutoff is less sample
1176 or gene specific compared to more complicated methods such as permutation test, it is
1177 computational efficient and already generates a reasonable set of gene-enhancer pairings.
1178
1179 To evaluate the performance of our methods, we compare our prediction with cis-eQTL
1180 derived from interferon-γ and lipopolysaccharide stimulation of monocytes[25]. Significant
1181 cis-eQTL associations are downloaded from supplementary material (Table S2) in Fairfax
1182 (2014)[25]. We filter cis-eQTL based on two criteria: 1) only cis-eQTLs that overlap with the
1183 peaks identified in PBMC dataset are considered; 2) In addition, we only keep the cis-
1184 eQTLs whose genes overlap with the variable genes determined by scRNA-seq. This
1185 filtering reduced the cis-eQTL list to 456 candidates.
1186
1187 Next, we estimate the association for each of cis-eQTLs by preforming logistic regression
1188 test as described above. To make a comparison, we derive a set of negative pairs matched
1189 for the distance. The negative control pairs for cis-eQTL are chosen in the following
1190 manner to control for both distance and chromatin accessibility: for each positive eQTL
1191 pair p_{ij} which connects gene i and enhancer j with a distance of d_{ij}, we look for the

1192 enhancer k on the opposite direction of the gene i that minimizes $|d_{ij} - d_{iz}|$. By doing so,
1193 the negative sets are controlled for distance, chromatin accessibility level and gene
1194 expression level.
1195
1196 **Simulation of scATAC-seq datasets**
1197 First, we download the alignment files (bam files) for ten bulk ATAC-seq experiment from
1198 ENCODE (data source listed in **Supplementary Table S2**). From each bam file, we
1199 simulate 1,000 single cell ATAC-seq datasets by randomly down sampling to a variety of
1200 coverages ranging from 1,000 to 10,000 reads per cells. We next create a cell-by-bin
1201 matrix of 5kb which is used for SnapATAC clustering. Merging peaks identified from each
1202 bulk experiment, we create cell-by-peak matrix used for LSA, Cis-Topic, Cicero and
1203 chromVAR for clustering. We repeat the sampling for n=10 times to estimate the
1204 variability of the clustering.
1205
1206 **Comparison of scalability**
1207 To compare the scalability between SnapATAC to other methods, we next simulate
1208 multiple datasets of different number of cells ranging from 20k to 1M. We simulate these
1209 datasets in the following manner. Using the 80k mouse atlas dataset, we randomly sample
1210 this dataset to different number of cells ranging from 20k to 1M cells. For the sampling
1211 that has cells more than 80K, we sample with replacement and introduce perturbation to
1212 each cell by randomly removing 1% of the "1"s in each of the cells. This removes the
1213 duplicate cells and largely maintains the density of the matrix.
1214
1215 For each sampling, we then perform dimensionality reduction using LSA and cisTopic
1216 and compare their CPU running time. Specifically, we monitor the running time for 1) TF-
1217 IDF transformation and Singular Value Decomposition (SVD) for LSA, 2) function
1218 "runModels" with topics = c(2, 5, 10, 15, 20, 25, 30, 35, 40) and "selectModel" function in
1219 cisTopic. The time for matrix loading is not counted.
1220
1221 All the comparisons were tested on a machine with 5 AMD Operon (TM) Processor 6276
1222 CPUs.

1223 **Doublets Detection Using Scrublet**

1224 To identify doublets from secondary motor cortex single nucleus ATAC-seq datasets, we
1225 use single cell RNA-seq doublets detection algorithm Scrublet[37]. Briefly, Scrublet
1226 identifies doublets in the following manner: 1) Scrublet performs normalization, gene
1227 filtering, and principal components analysis (PCA) to project the high-dimension data to
1228 a low-dimension space; 2) Scrublet simulates doublets by adding the unnormalized
1229 counts from randomly sampled observed transcriptomes; 3) the simulated doublets are
1230 projected to the low dimension embedding computed in step 1. The more neighbors of a
1231 cell are the simulated doublets, the more likely this cell is a "doublet". Based on this idea,
1232 a KNN classifier was then used to estimate the doublet score for each cell.

1233

1234 Since Scrublet was designed for detecting doublets in single cell RNA-seq, it is unclear
1235 whether it can be used for single cell ATAC-seq. To examine this, we applied Scrublet to
1236 a single cell ATAC-seq dataset of mixed human and mouse cells where the "ground-truth"
1237 doublets can be identified based on the alignment ratio to human and mouse genome.
1238 Compared to the ground truth, Scrubet can identify over 90% of the doublets in this
1239 dataset with ~90% accuracy (**Supplementary Fig. 26**). This result suggests that
1240 although Scrubet was not developed for detecting doublets in single cell ATAC-seq, it can
1241 find the doublets in scATAC-seq dataset with reasonable accuracy and sensitivity.

1242

1243 **Projection of single cell ATAC-seq datasets to reference atlas**

1244 We reason that landmark-extension algorithm can also be extended to project new single
1245 cell ATAC-seq datasets to a reference atlas. Given a query dataset $\mathbf{Y} \in \mathcal{R}^{l\times m}$ that contains
1246 $\boldsymbol{l}$ query cells with $\boldsymbol{m}$ bins and a reference dataset $\mathbf{X} \in \mathcal{R}^{n\times m}$ with $\boldsymbol{n}$ reference cells of $\boldsymbol{m}$
1247 bins. We first randomly sample $\boldsymbol{k}$ =10,000 landmarks from $\mathbf{X}$ using density-based
1248 sampling as described above. Next, we compute the pairwise similarity using normalized
1249 jaccard coefficient for $\boldsymbol{k}$ landmarks as $\boldsymbol{N}^{kk} \in \mathcal{R}^{k\times k}$ and obtain the low-dimension
1250 manifold $\boldsymbol{U}^{k} \in \mathcal{R}^{k\times r}$. We then compute $\boldsymbol{N}^{lk} \in \mathcal{R}^{l\times k}$ which estimates the similarity
1251 between $\boldsymbol{l}$ query cells and $\boldsymbol{k}$ landmark cells, and then project the $\boldsymbol{l}$ query cells to the
1252 embedding pre-computed for $\boldsymbol{k}$ landmark cells as following:

1253

$$\boldsymbol{A^l} = (\boldsymbol{D^l})^{-\frac{1}{2}}(\boldsymbol{U^k})(\boldsymbol{D^k})^{-\frac{1}{2}}$$

where $\boldsymbol{D^l} \in \boldsymbol{\mathcal{R}^{l\times l}}$ is a diagonal matrix which is composed as $\boldsymbol{D^l_{i,i}} = \sum_j \boldsymbol{N^l_{i,j}}$ and $\boldsymbol{D^k} \in \boldsymbol{\mathcal{R}^{k\times k}}$ is a diagonal matrix which is composed as $\boldsymbol{D^k_{i,i}} = \sum_j \boldsymbol{N^k_{i,j}}$

$$\boldsymbol{U^l} = \boldsymbol{A^l U^k / \Lambda^k}$$

The resulting $\boldsymbol{U^l} \in \boldsymbol{\mathcal{R}^{l\times r}}$ is the predicted low-dimension manifold for $\boldsymbol{l}$ query cells.

In the joint embedding space $[\boldsymbol{U^k}, \boldsymbol{U^l}]$, we next identify the mutual nearest neighbors between query and landmark cells. For each cell $i_1 \in \mathbf{X}^{\boldsymbol{k}}$ belonging to the landmarks, we find the $\boldsymbol{k.nearest}$ **(5)** cells in the query dataset with the smallest distances to $\boldsymbol{i_1}$. We do the same for each cell in query cell dataset to find its $\boldsymbol{k.nearest}$ (5) neighbors in the landmark dataset. If a pair of cells from each dataset is contained in each other's nearest neighbors, those cells are considered to be mutual nearest neighbors or MNN pairs (or "anchors"). We interpret these pairs as containing cells that belong to the same cell type or state despite being generated in both landmark and query cells. Thus, any differences between cells in MNN pairs should theoretically represent the non-overlapping cell types. Here we removed any query cells that failed to identify an MNN pair correspondence in the reference dataset.

To make a classification of the remaining query cells according to the reference dataset, we next apply the neighborhood-based classifier and wish to highlight the pioneering work by Seurat V3[18]. First, we score each anchor (or MNN pair) using shared nearest neighbor (SNN) graph by examining the consistency of edges between cells in the same local neighborhood as described in the original study[18]. Second, we define a weight matrix that estimates the strength of association between each query cell $\boldsymbol{c}$, and each landmark $\boldsymbol{i}$. For each query cell $\boldsymbol{c}$, we identify the nearest $\boldsymbol{s}$ landmarks in the reference dataset in the joint embedding space. Nearest anchors are then weighted based on their distance to the

cell $\boldsymbol{c}$ over the distance to the $\boldsymbol{s}$-th anchor cell. For each cell $\boldsymbol{c}$ and anchor $\boldsymbol{i}$, we compute the weighted similarities as:

$$\boldsymbol{D_{c,i}} = (\mathbf{1} - \frac{\boldsymbol{dist(c, a_i)}}{\boldsymbol{dist(c, a_s)}})\boldsymbol{S_{ai}}$$

Where $\boldsymbol{dist(c, i)}$ is the Euclidean distance in the joint embedding space and S_{ai} is the weight for the corresponding MNN pair (anchor). We then normalize the similarity using exponential function:

$$\widetilde{\boldsymbol{D_{c,i}}} = \mathbf{1} - \boldsymbol{e}^{\frac{-\boldsymbol{D_{c,i}}}{\mathbf{2}}(\frac{}{\boldsymbol{sd}})^{\mathbf{2}}}$$

where sd is set to 1 by default. Finally, we normalize across all $\boldsymbol{s}$ anchors:

$$\boldsymbol{W_{c,i}} = \frac{\widetilde{\boldsymbol{D_{c,i}}}}{\sum_{\mathbf{1}}^{\boldsymbol{j=s}} \widetilde{\boldsymbol{D_{c,J}}}}$$

Here we set $\boldsymbol{s} = \mathbf{50}$. Please note that the similarity to cells beyond the $\boldsymbol{s^{th}}$ anchor neighbor is set to be zero.

Let $\boldsymbol{L} \in \boldsymbol{\mathcal{R}^{k \times t}}$ be the binary label matrix for $\boldsymbol{k}$ landmarks with $\boldsymbol{t}$ clusters. $\boldsymbol{L_{i,j}} = \mathbf{1}$ indicates the class label for $\boldsymbol{i}$-th landmark cell is $\boldsymbol{j}$-th cluster. The row sum of $\boldsymbol{L}$ must be 1, suggesting each landmark cell can only be assigned to one cluster label. We then compute label predictions for query cells as $\boldsymbol{P^l}$:

$$\boldsymbol{P^l} = \boldsymbol{WL}$$

The resulting $\boldsymbol{P^l}$ is a probability matrix within 0 and 1, $\boldsymbol{P^l_{i,j}}$ indicates the probability of a cell $\boldsymbol{i}$ belong to $\boldsymbol{j}$ cluster. Similarly, we infer the t-SNE position of query cells by replacing $\boldsymbol{L}$ with t-SNE coordinates of reference points. It is important to note that the distance

1310 between cells in the inferred t-SNE coordinate does not neccessarily reflect the cell-to-cell
1311 relationship.
1312
1313 **Tissue collection & nuclei isolation**
1314 Adult C57BL/6J male mice were purchased from Jackson Laboratories. Brains were
1315 extracted from P56-63 old mice and immediately sectioned into 0.6 mm coronal sections,
1316 starting at the frontal pole, in ice-cold dissection media. The secondary motor cortex
1317 (MOs) region was dissected from the first three slices along the anterior-posterior axis
1318 according to the Allen Brain reference Atlas (http://mouse.brain-map.org/, see
1319 **Supplementary Fig. 15a** for depiction of posterior view of each coronal slice; dashed
1320 line highlights the MOs regions on each slice). Slices were kept in ice-cold dissection
1321 media during dissection and immediately frozen in dry ice for posterior pooling and
1322 nuclei production. For nuclei isolation, the MOs dissected regions from 15-23 animals
1323 were pooled, and two biological replicates were processed for each slice. Nuclei were
1324 isolated as described in previous studies[54,55], except no sucrose gradient purification was
1325 performed. Flow cytometry analysis of brain nuclei was performed as described in Luo et
1326 al[54].
1327
1328 **Tn5 transposase purification & loading**
1329 Tn5 transposase was expressed as an intein chitin-binding domain fusion and purified
1330 using an improved version of the method first described by Picelli et al[56]. T7 Express
1331 lysY/I (C3013I, NEB) cells were transformed with the plasmid pTXB1-ecTn5 E54K L372P
1332 (#60240, Addgene)[56]. An LB Ampicillin culture was inoculated with three colonies and
1333 grown overnight at 37°C. The starter culture was diluted to an OD of 0.02 with fresh
1334 media and shaken at 37°C until it reached an OD of 0.9. The culture was then immediately
1335 chilled on ice to 10°C and expression was induced by adding 250 µM IPTG (Dioxane Free,
1336 CI8280-13, Denville Scientific). The culture was shaken for 4 hours at 23°C after which
1337 cells were harvested in 2 L batches by centrifugation, flash frozen in liquid nitrogen and
1338 stored at -80°C. Cell pellets were resuspended in 20 ml of ice cold lysis buffer (20 mM
1339 HEPES 7.2-KOH, 0.8 M NaCl, 1 mM EDTA, 10% Glycerol, 0.2% Triton X-100) with
1340 protease inhibitors (Complete, EDTA-free Protease Inhibitor Cocktail Tablets,
1341 11873580001, Roche Diagnostics) and passed three times through a Microfluidizer (lining

1342 covered with ice water, Model 110L, Microfluidics) with a 5 minute cool down interval in
1343 between each pass. Any remaining sample was purged from the Microfluidizer with an
1344 additional 25 ml of ice-cold lysis buffer with protease inhibitors (total lysate volume
1345 ~50ml). Samples were spun down for 20 min in an ultracentrifuge at 40K rpm (L-80XP,
1346 45 Ti Rotor, Beckman Coulter) at 4°C. ~45 ml of supernatant was combined with 115 ml
1347 ice cold lysis buffer with protease inhibitors in a cold beaker (total volume = 160 ml) and
1348 stirred at 4°C. 4.2ml of 10% neutralized polyethyleneimine-HCl (pH 7.0) was then added
1349 dropwise. Samples were spun down again for 20 min in an ultracentrifuge at 40K rpm (L-
1350 80XP, 45 Ti Rotor, Beckman Coulter) at 4°C. The pooled supernatant was loaded onto
1351 ~10ml of fresh Chitin resin (S6651L, NEB) in a chromatography column (Econo-Column
1352 (1.5 × 15 cm), Flow Adapter: 7380015, Bio-Rad). The column was then washed with 50-
1353 100 ml lysis buffer. Cleavage of the fusion protein was initiated by flowing ~20ml of
1354 freshly made elution buffer (20 mM HEPES 7.2-KOH, 0.5 M NaCl, 1 mM EDTA, 10%
1355 glycerol, 0.02% Triton X-100, 100mM DTT) onto the column at a speed of 0.8ml/min for
1356 25 min. After the column was incubated for 63 hrs at 4°C, the protein was recovered from
1357 the initial elution volume and a subsequent 30 ml wash with elution buffer. Protein-
1358 containing fractions were pooled and diluted 1:1 with buffer [20 mM HEPES 7.2-KOH,1
1359 mM EDTA, 10% glycerol, 0.5mM TCEP) to reduce the NaCl concentration to 250mM. For
1360 cation exchange, the sample was loaded onto a 1ml column HiTrap S HP (17115101, GE),
1361 washed with Buffer A (10mM Tris 7.5, 280 mM NaCl, 10% glycerol, 0.5mM TCEP) and
1362 then eluted using a gradient formed using Buffer A and Buffer B (10mM Tris 7.5, 1M NaCl,
1363 10% glycerol, 0.5mM TCEP) (0% Buffer B over 5 column volumes, 0-100% Buffer B over
1364 50 column volumes, 100% Buffer B over 10 column volumes). Next, the protein-
1365 containing fractions were combined, concentrated via ultrafiltration to ~1.5 mg/mL and
1366 further purified via gel filtration (HiLoad 16/600 Superdex 75 pg column (28989333,
1367 GE)) in Buffer GF (100mM HEPES-KOH at pH 7.2, 0.5 M NaCl, 0.2 mM EDTA, 2mM
1368 DTT, 20% glycerol). The purest Tn5 transposase-containing fractions were pooled and 1
1369 volume 100% glycerol was added to the preparation. Tn5 transposase was stored at -20°C.
1370
1371 To generate Tn5 transposomes for combinatorial barcoding assisted single nuclei
1372 ATAC-seq, barcoded oligos were first annealed to pMENTs oligos (95 °C for 5 min,
1373 cooled to 14 °C at a cooling rate of 0.1 °C/s) separately. Next, 1 µl barcoded transposon

1374 (50 µM) was mixed with 7 ul Tn5 (~7 µM). The mixture was incubated on the lab bench
1375 at room temperature for 30 min. Finally, T5 and T7 transposomes were mixed in a 1:1
1376 ratio and diluted 1:10 with dilution buffer (50 % Glycerol, 50 mM Tris-HCl (pH=7.5),
1377 100 mM NaCl, 0.1 mM EDTA, 0.1 % Triton X-100, 1 mM DTT). For combinatorial
1378 barcoding, we used eight different T5 transposomes and 12 distinct T7 transposomes,
1379 which eventually resulted in 96 Tn5 barcode combinations per sample[7]
1380 (**Supplementary Table S6**).
1381
1382 **Bulk ATAC-seq data generation**
1383 ATAC-seq was performed on 30,000-50,000 nuclei as described previously with
1384 modifications[3]. Nuclei were thawed on ice and pelleted for 5 min at 500 x g at 4 °C. Nuclei
1385 pellets were resuspended in 30 µl tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8),
1386 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF) and counted on a hemocytometer.
1387 30,000-50,000 nuclei were used for tagmentation and the reaction volume was adjusted
1388 to 19 µl using tagmentation buffer. After addition of 1 µl TDE1 (Illumina FC-121-1030),
1389 tagmentation was performed at 37°C for 60 min with shaking (500 rpm). Tagmented
1390 DNA was purified using MinElute columns (Qiagen), PCR-amplified for 8 cycles with
1391 NEBNext® High-Fidelity 2X PCR Master Mix (NEB, 72°C 5 min, 98°C 30 s, [98°C 10 s,
1392 63°C 30 s, 72°C 60 s] x 8 cycles, 12°C held). Amplified libraries were purified using
1393 MinElute columns (Qiagen) and SPRI Beads (Beckmann Coulter). Sequencing was
1394 carried out on a NextSeq500 using a 150-cycle kit (75 bp PE, Illumina).
1395
1396 **Bulk ATAC-seq data analysis**
1397 ATAC-seq reads were mapped to reference genome mm10 using BWA and *samtools*
1398 version 1.2 to eliminate PCR duplicates and mitochondrial reads. The paired-end read
1399 ends were converted to fragments. Using fragments, MACS2[57] version 2.1.2 was used for
1400 generating signal tracks and peak calling with the following parameters: --nomodel --shift
1401 100 --ext 200 --qval 1e-2 -B –SPMR. Library quality control for bulk ATAC-seq can be
1402 found in **Supplementary Table S7**.
1403
1404 **Single-nucleus ATAC-seq data generation**

1405 Combinatorial ATAC-seq was performed as described previously with modifications[5,7].
1406 For each sample two biological replicates were processed. Nuclei were pelleted with a
1407 swinging bucket centrifuge (500 x g, 5 min, 4°C; 5920R, Eppendorf). Nuclei pellets were
1408 resuspended in 1 ml nuclei permeabilization buffer (5 % BSA, 0.2 % IGEPAL-CA630, 1mM
1409 DTT and cOmpleteTM, EDTA-free protease inhibitor cocktail (Roche) in PBS) and
1410 pelleted again (500 x g, 5 min, 4°C; 5920R, Eppendorf). Nuclei were resuspended in
1411 500 µL high salt tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM
1412 potassium-acetate, 11 mM Mg-acetate, 17.6% DMF) and counted using a hemocytometer.
1413 Concentration was adjusted to 4500 nuclei/9 µl, and 4,500 nuclei were dispensed into
1414 each well of a 96-well plate. Glycerol was added to the leftover nuclei suspension for a
1415 final concentration of 25 % and nuclei were stored at -80°C. For tagmentation, 1 µL
1416 barcoded Tn5 transposomes[7,56] (**Supplementary Table S6**) were added using a
1417 BenchSmart™ 96 (Mettler Toledo), mixed five times and incubated for 60 min at 37 °C
1418 with shaking (500 rpm). To inhibit the Tn5 reaction, 10 µL of 40 mM EDTA were added
1419 to each well with a BenchSmart™ 96 (Mettler Toledo) and the plate was incubated at 37
1420 °C for 15 min with shaking (500 rpm). Next, 20 µL 2 x sort buffer (2 % BSA, 2 mM EDTA
1421 in PBS) were added using a BenchSmart™ 96 (Mettler Toledo). All wells were combined
1422 into a FACS tube and stained with 3 µM Draq7 (Cell Signaling). Using a SH800 (Sony),
1423 20 nuclei were sorted per well into eight 96-well plates (total of 768 wells) containing
1424 10.5 µL EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma)[7]. Preparation of
1425 sort plates and all downstream pipetting steps were performed on a Biomek i7 Automated
1426 Workstation (Beckman Coulter). After addition of 1 µL 0.2% SDS, samples were
1427 incubated at 55 °C for 7 min with shaking (500 rpm). We added 1 µL 12.5% Triton-X to
1428 each well to quench the SDS and 12.5 µL NEBNext High-Fidelity 2× PCR Master Mix
1429 (NEB). Samples were PCR-amplified (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72
1430 °C 60 s) × 12 cycles, held at 12 °C). After PCR, all wells were combined. Libraries were
1431 purified according to the MinElute PCR Purification Kit manual (Qiagen) using a vacuum
1432 manifold (QIAvac 24 plus, Qiagen) and size selection was performed with SPRI Beads
1433 (Beckmann Coulter, 0.55x and 1.5x). Libraries were purified one more time with SPRI
1434 Beads (Beckmann Coulter, 1.5x). Libraries were quantified using a Qubit fluorimeter (Life
1435 technologies) and the nucleosomal pattern was verified using a Tapestation (High

1436 Sensitivity D1000, Agilent). The library was sequenced on a HiSeq2500 sequencer
1437 (Illumina) using custom sequencing primers, 25% spike-in library and following read
1438 lengths: 50 + 43 + 40 + 50 (Read1 + Index1 + Index2 + Read2)[7].
1439
1440

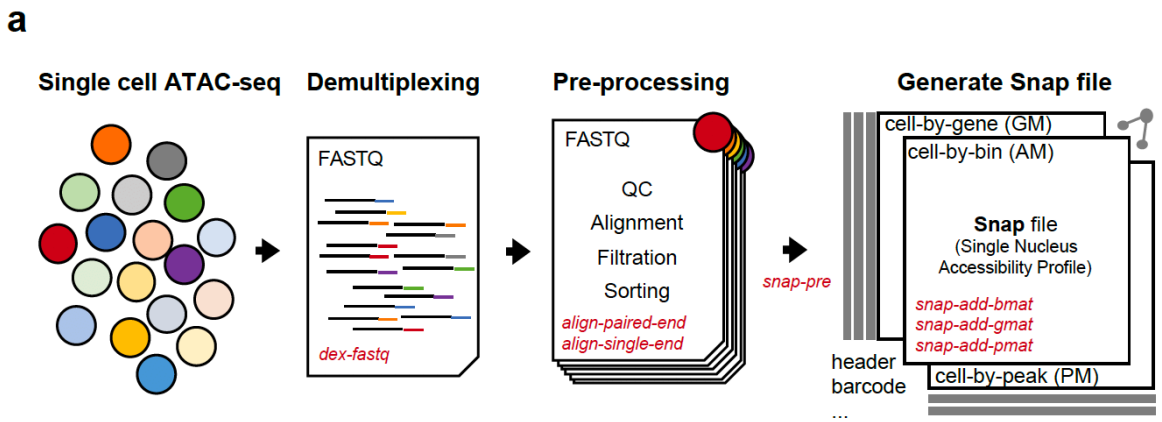
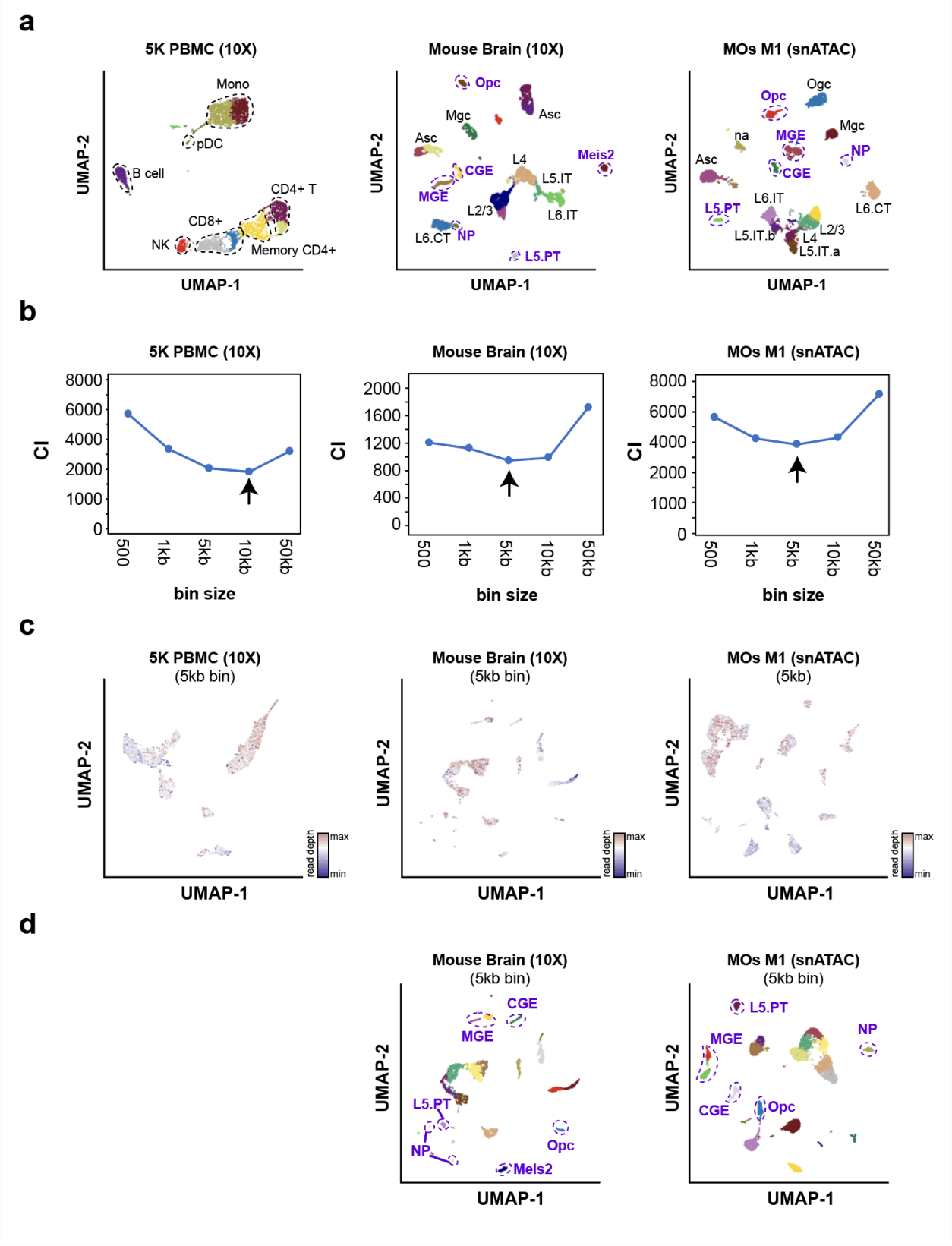


Figure S1. Overview of SnapTools workflow. (**a**) Demultiplexing: SnapTools first demultiplexed the fastq files by adding the cell barcodes to the beginning of each read name; Pre-processing: raw sequencing reads were aligned to the reference genome using BWA followed by filtration of erroneous alignments. A snap file was generated to store indexed reads and multiple cell matrices including cell-by-peak, cell-by-gene and cell-by-bin matrix.

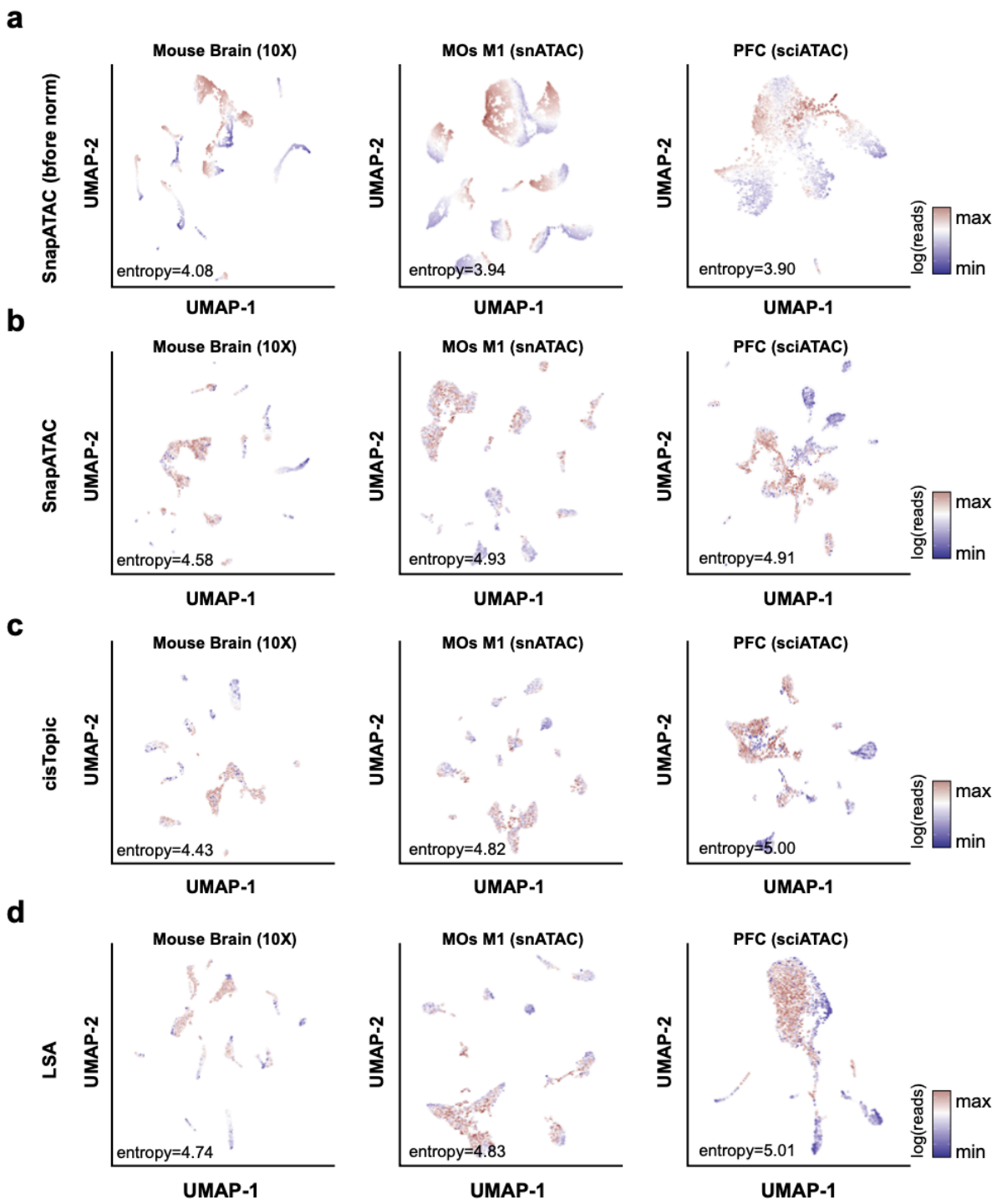

a
5K PBMC (10X)
Mono
pDC
B cell
CD4+ T
CD8+
NK
Memory CD4+
UMAP-2
UMAP-1
Mouse Brain (10X)
Opc
Mgc
Asc
Asc
CGE
MGE
L4
L5.IT
Meis2
L2/3
L6.IT
L6.CT
NP
L5.PT
MOs M1 (snATAC)
Opc
Ogc
Mgc
na
MGE
NP
Asc
CGE
L6.IT
L6.CT
L5.PT
L5.IT.b
L2/3
L4
L5.IT.a
b
CI
bin size
500
1kb
5kb
10kb
50kb
c
(5kb bin)
(5kb)
read depth
max
min
d

1450

1451 **Figure S2. Choosing the optimal bin size.** (**a**) UMAP visualization of landmark cell
1452 types identified in three benchmarking datasets. UMAP embedding was computed using
1453 cisTopic and cell types were manually annotated based on the gene accessibility score at
1454 canonical marker genes (**Supplementary Methods**). Blue dash line highlights the rare
1455 cell populations that account for less than 2% of the total population. (**b**) Relationship
1456 between connectivity index (CI) and bin sizes. Connectivity index were calculated
1457 between landmark cell types in the reduced dimension using function "connectivity" in R
1458 package "clv". A lower CI indicates a better separation of landmark cell types. (**c**) UMAP
1459 representation of three benchmarking datasets generated using SnapATAC using 5kb bin
1460 size. Cells colored by read depth to illustrate the sequencing depth effect. (**d**) Cells are
1461 colored by cluster labels identified by SnapATAC. Data source are listed in
1462 **Supplementary Table S1**. Note that blue circles highlight rare cell populations account
1463 for less than 2% of total population.

1464

1465


a
Mouse Brain (10X)
MOs M1 (snATAC)
PFC (sciATAC)
SnapATAC (bfore norm)
UMAP-2
UMAP-1
entropy=4.08
entropy=3.94
entropy=3.90
log(reads)
max
min
b
SnapATAC
entropy=4.58
entropy=4.93
entropy=4.91
c
cisTopic
entropy=4.43
entropy=4.82
entropy=5.00
d
LSA
entropy=4.74
entropy=4.83
entropy=5.01

1471 **Figure S3. SnapATAC is robust to sequencing depth.** Two dimensional UMAP
1472 representation of three benchmarking datasets analyzed by four methods (**a**) SnapATAC
1473 without normalization; (**b**) SnapATAC with normalization; (**c**) cisTopic and (**d**) Latent
1474 Sematic Analysis (LSA). Cells are color by log-scaled read depth. Read depth bias is
1475 quantified by entropy as described in the **Supplementary Methods**. Data source is
1476 listed in **Supplementary Table S1**.

1477

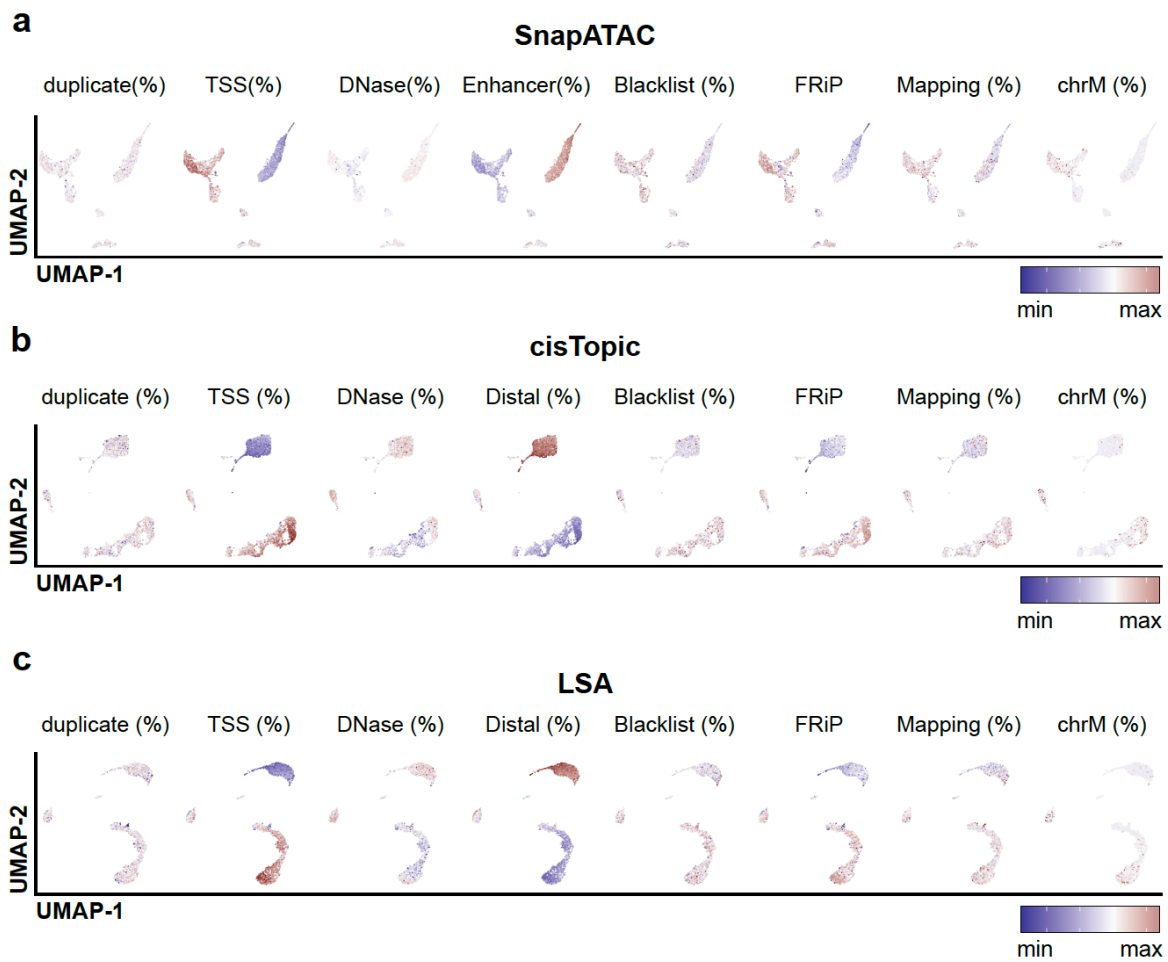
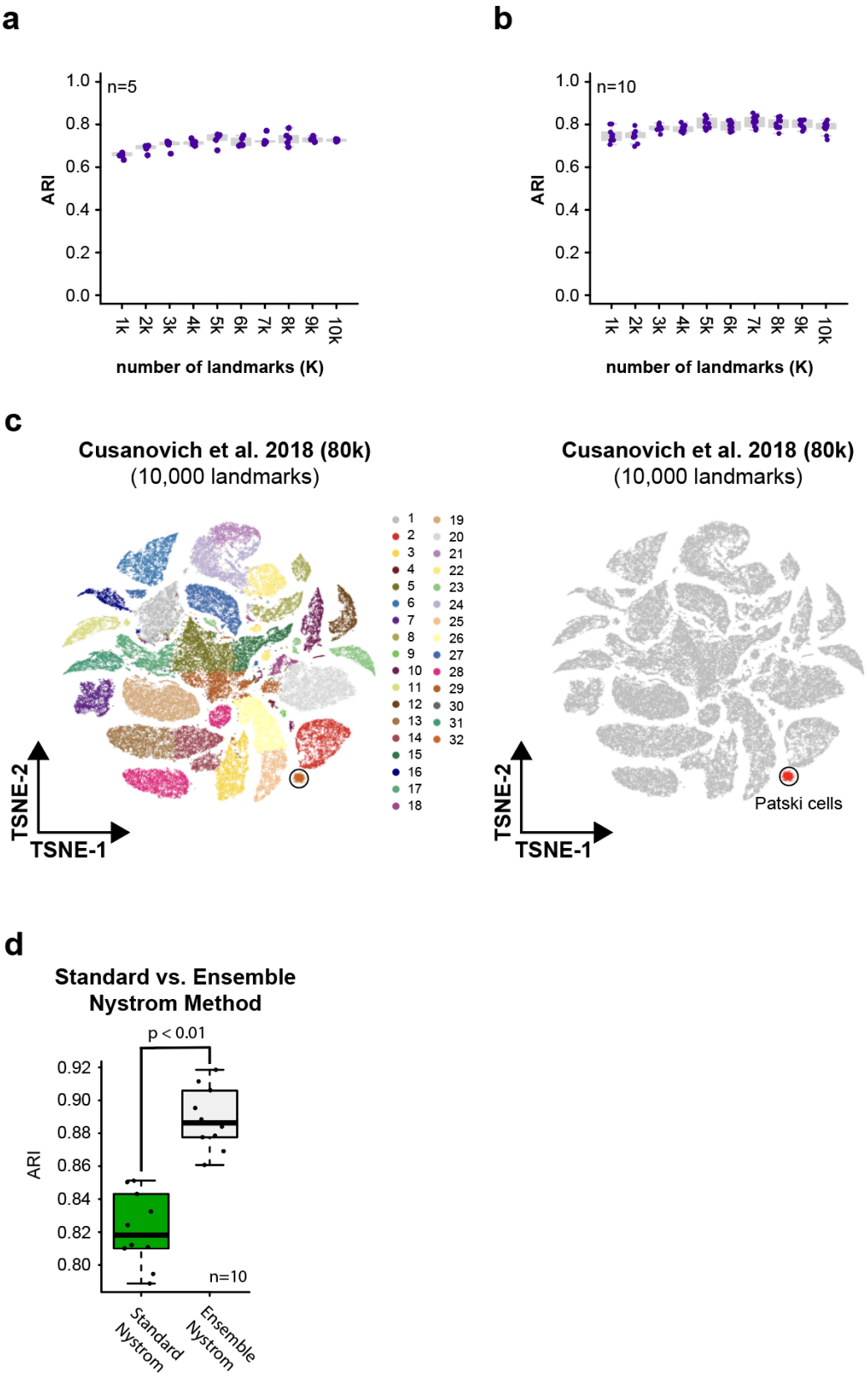


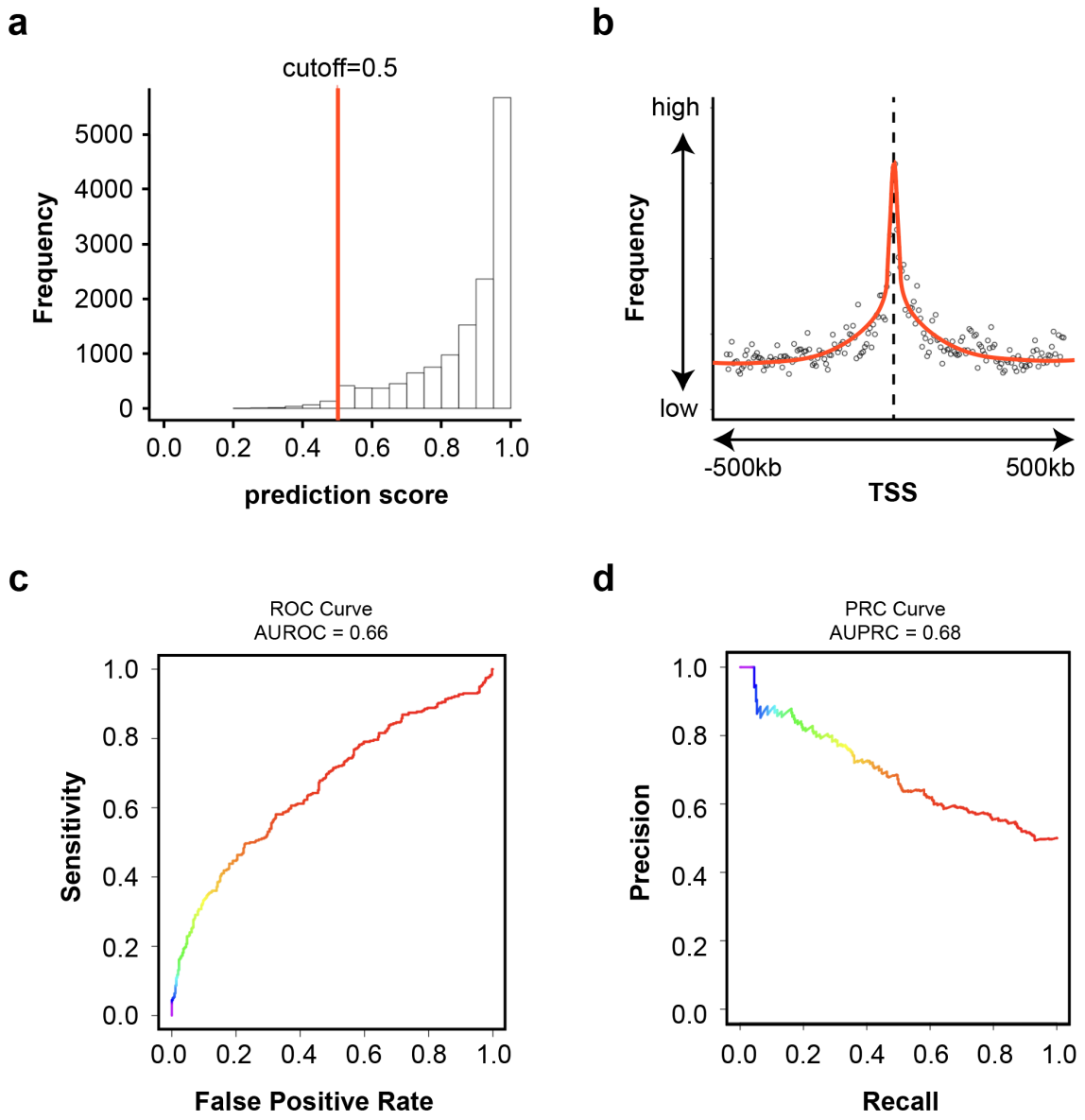
Figure S4. SnapATAC is robust to other biases. Potential bias in single cell ATAC-seq dataset projected onto the UMAP visualization generated using different analysis methods (**a**) SnapATAC (**b**) cisTopic and (**c**) LSA. Duplicate: percentage of fragments that are PCR duplicates. TSS: percentage of fragments overlapping or are within 1kb of a TSS. TSS position is based on the GENECODE V28 (Ensemble 92). DNase: the percentage of fragments overlapping a master DNase peak list. The DNase peak list is created by combining all ENCODE[1] DNase peaks from hg19. Blacklist: the percentage of fragments overlapping with the ENCODE blacklist. FRiP: the percentage of fragments overlapping with the peaks defined from the aggregate signal. Mapping: the percentage of fragments that are uniquely mapped. chrM: the percentage of fragments mapped to mitochondria DNA. Dataset used in this plot is 5k PBMC (10X) as listed in **Supplementary Table S1**.


a
n=5
ARI
1.0
0.8
0.6
0.4
0.2
0.0
1k 2k 3k 4k 5k 6k 7k 8k 9k 10k
number of landmarks (K)
b
n=10
ARI
number of landmarks (K)
c
Cusanovich et al. 2018 (80k)
(10,000 landmarks)
TSNE-2
TSNE-1
Cusanovich et al. 2018 (80k)
(10,000 landmarks)
Patski cells
TSNE-2
TSNE-1
d
Standard vs. Ensemble
Nystrom Method
p < 0.01
ARI
0.92
0.90
0.88
0.86
0.84
0.82
0.80
n=10
Standard Nystrom
Ensemble Nystrom

1491

1492 **Figure S5. Ensemble Nyström sampling improves the scalability and stability**
1493 **without sacrificing the performance.** (**a**) A line plot comparing the performance of
1494 clustering using various sampling parameters. The performance is evaluated using
1495 Adjusted Rank Index (ARI). SnapATAC was applied to the mouse atlas dataset that
1496 contained over 80k cells using different number of landmark cells (*k*) ranging from 1k to
1497 10k. For each *k*, we performed clustering for n=5 times using different sets of randomly
1498 selected landmarks. (**b**) A line plot comparing the stability of clustering results between
1499 five samplings (pairwise comparison n=10). (**c**) To evaluate the sensitivity of identifying
1500 rare cell types, we spiked in 1% mouse Pastki cells generated using the same protocol in
1501 Cusanovich 2015[5] and this rare cell population was recapitulated using 10,000 landmarks
1502 (right). (**d**) To compare the clustering reproducibility between standard and ensemble
1503 Nystrom sampling method, we performed clustering using both methods on Cusanovich
1504 2018[8] for five times with different randomly selected landmark cells. The clustering
1505 reproducibility quantified by ARI (adjusted rank index) between random trails is
1506 significantly higher for the ensemble Nystrom method than the standard Nystrom
1507 method (two-tailed t-test $P < 0.01$).
1508

1509



1510

1511

1512

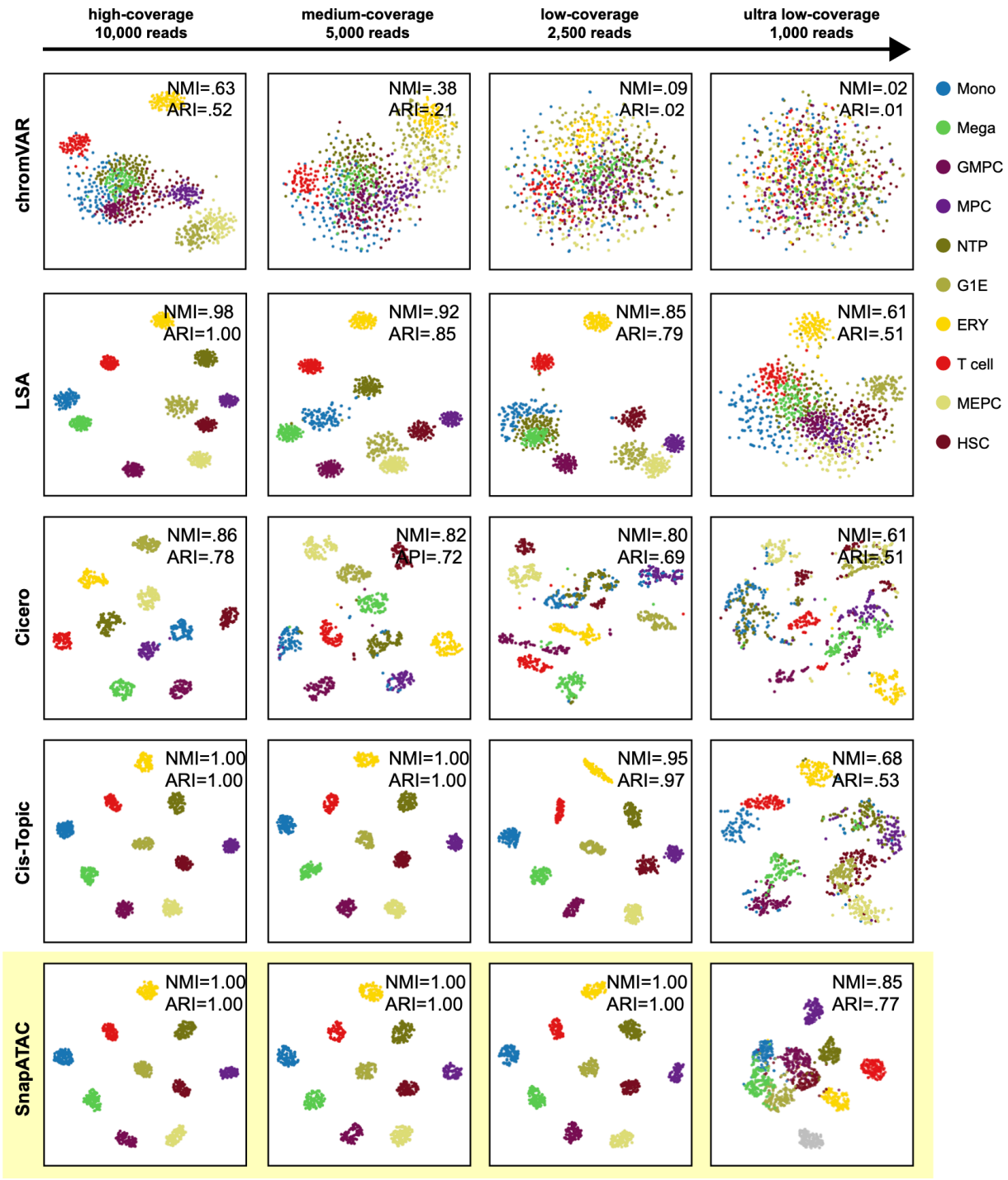
1513

1514

1515

1516 **Figure S6. SnapATAC predicts gene and enhancer pairing by integrating**
1517 **scATAC-seq and scRNA-seq.** (**a**) Prediction score distribution for single cell ATAC-
1518 seq (5K PBMC 10X) by SnapATAC. When predicting the cell type for scATAC-seq using
1519 corresponding scRNA-seq dataset (10X PBMC scRNA-seq), each cell in scATAC-seq was
1520 assigned with a prediction score indicating the confidence of the prediction. It ranges
1521 from 0 to 1, a higher score indicates a higher confidence. Using 0.5 as cutoff as suggested
1522 in Seurat, over 98% of cells in scATAC-seq are confidently assigned to a cell type defined
1523 in scRNA-seq. (**b**) Distance decay curve for the association (-logPvalue) between
1524 regulatory elements and the TSS of their putative target genes. (**c-d**) AUROC and AUPRC
1525 between cis-eQTL pairs and negative control sets. See **Supplementary Methods** for
1526 how the control sets selected.

1527



1528
1529
1530
1531

1532 **Figure S7. Evaluation of clustering accuracy of SnapATAC relative to**
1533 **alternative methods on simulated datasets.** T-SNE visualization of clustering
1534 results on 1,000 simulated cells sampled from 10 bulk ATAC-seq datasets (see
1535 **Supplementary Methods** for the simulation) analyzed by five different methods –
1536 chromVAR[14], LSA[8], Cicero[17], Cis-Topic[15] and SnapATAC. Clustering results are compared
1537 to the original cell type label and the accuracy is estimated using Normalized Mutual
1538 Index (nmi). Mono: monocyte; Mega: megakaryocyte; GMPC: granulocyte monocyte
1539 progenitor cell; MPC: megakaryocyte progenitor cell; NPT: neutrophil; G1E: G1E; T cell:
1540 regulatory T cell; MEPC: megakaryocyte-erythroid progenitor cell; HSC: hematopoietic
1541 stem cell.
1542

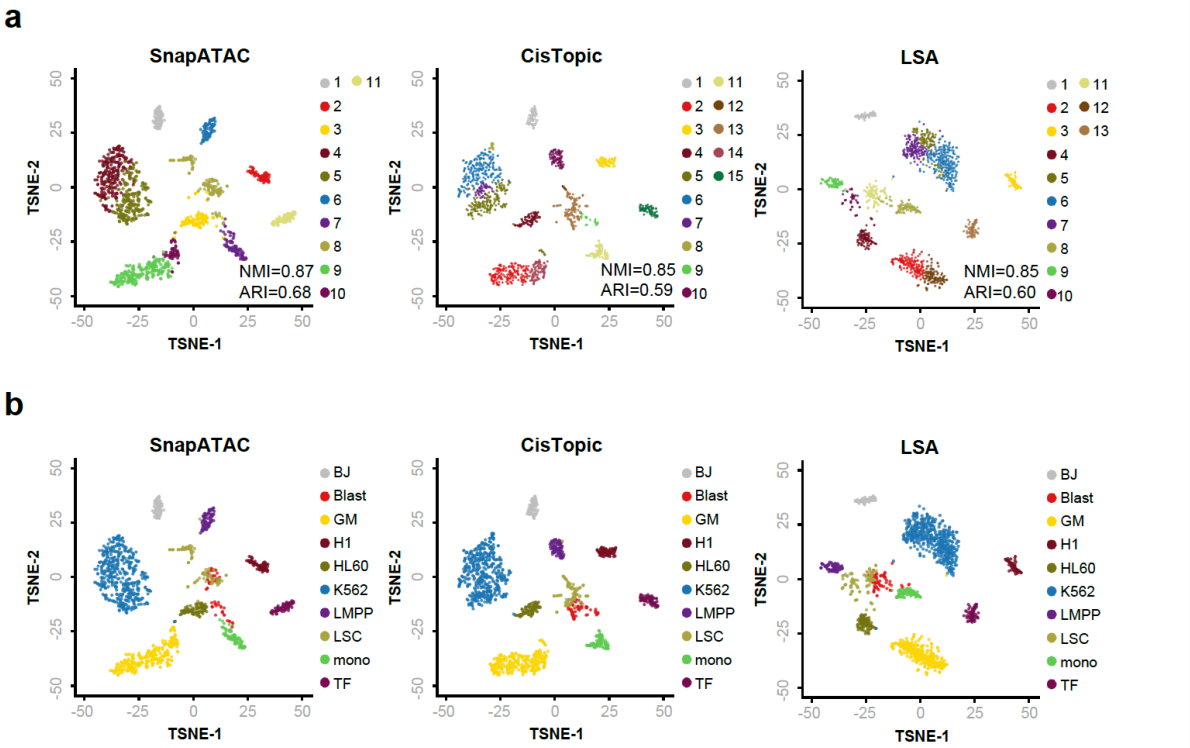
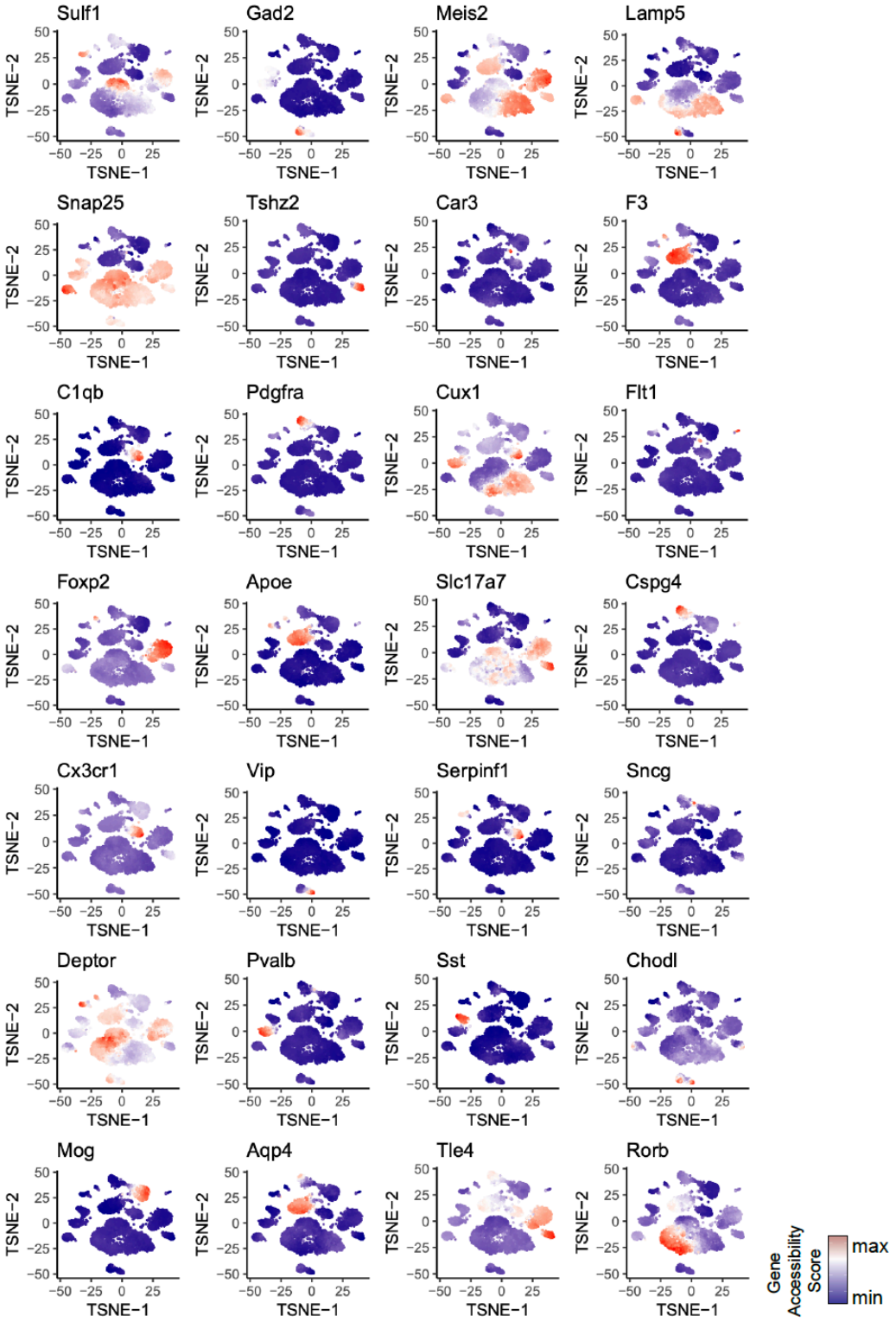


Figure S8. Evaluation of clustering accuracy relative to alternative methods on published single cell ATAC-seq datasets. SnapATAC (left), CisTopic (middle) and LSA (right) clustering performance on single cell ATAC-seq dataset from ten human cell lines generated using Fluidigm C1 platform[10,14]. (**a**) Clustering results are visualized using t-SNE and cells are colored by cluster labels identified by each of analysis methods. (**b**) T-SNE visualization of the human cells colored by the cell type labels. Clustering accuracy of each method is estimated by comparing the predicted clustering labels to the cell type labels. Blast: acute myeloid leukemia blast cells; LSC: acute myeloid leukemia leukemic stem cells; LMPP: lymphoid-primed multipotent progenitors; Mono: monocyte; HL60: HL-60 promyeloblast cell line; TF1: TF-1 erythroblast cell line; GM: GM12878 lymphoblastoid cell line; BJ: human fibroblast cell line; H1: H1 human embryonic stem cell line.


Sulf1
Gad2
Meis2
Lamp5
Snap25
Tshz2
Car3
F3
C1qb
Pdgfra
Cux1
Flt1
Foxp2
Apoe
Slc17a7
Cspg4
Cx3cr1
Vip
Serpinf1
Sncg
Deptor
Pvalb
Sst
Chodl
Mog
Aqp4
Tle4
Rorb
TSNE−1
TSNE−2
Gene Accessibility Score
max
min

1558

1559 **Figure S9. Gene accessibility score of canonical marker genes projected onto**
1560 **t-SNE embedding for snATAC-seq dataset from mouse secondary motor**
1561 **cortex.** T-SNE is generated using SnapATAC; cell type specific marker genes were
1562 defined from previous single cell transcriptomic analysis in the adult mouse brain[38]; gene
1563 accessibility score is calculated using SnapATAC (**Supplementary Methods**). Data
1564 source is listed in **Supplementary Table S1**.

1565

1566

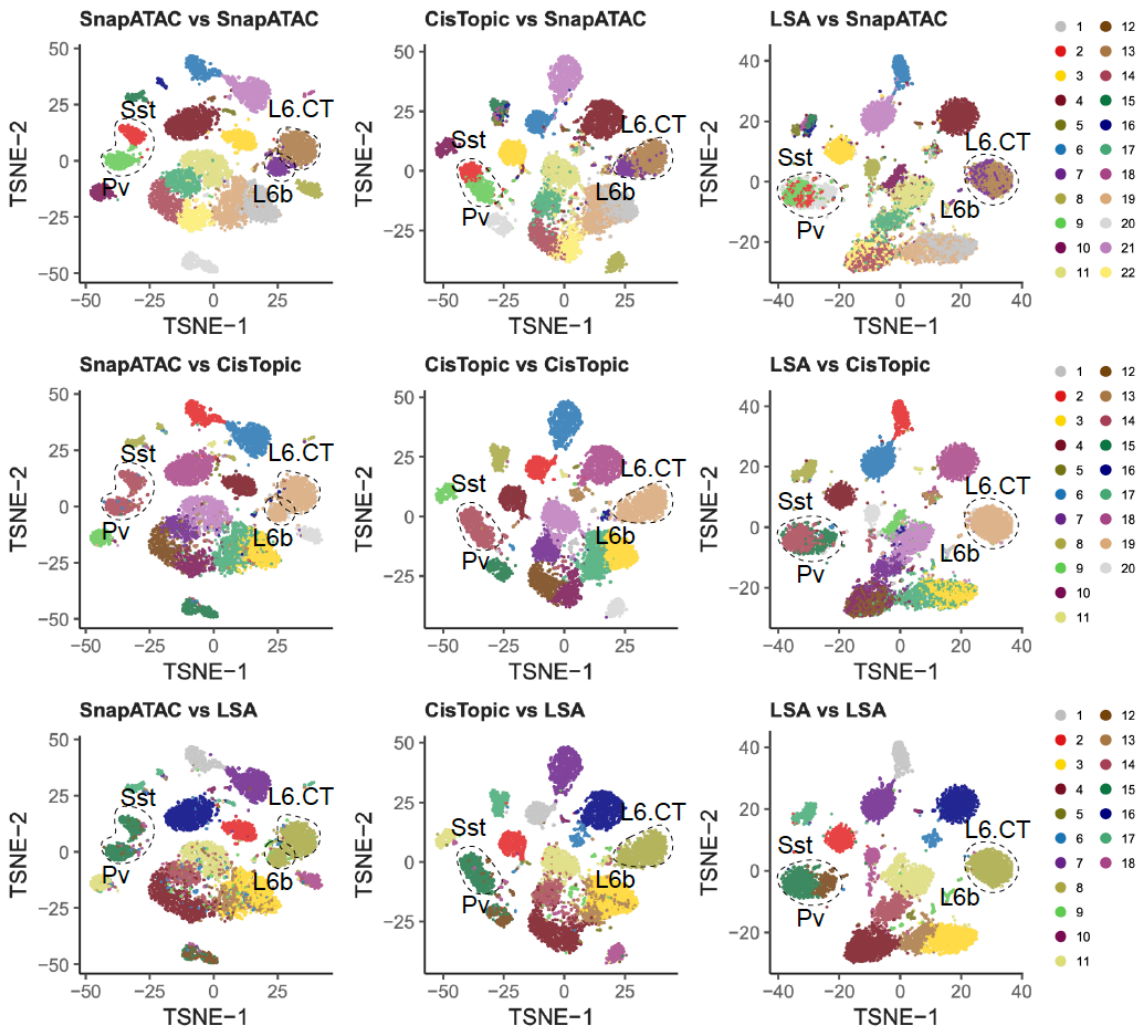


Figure S10. Evaluation of clustering sensitivity of SnapATAC relative to alternative methods on mouse secondary motor cortex snATAC-seq. Three methods (cisTopic, LSA and SnapATAC) were used to analyze a dataset that contained ~10k single nucleus ATAC-seq profiles from the mouse secondary motor cortex. Pairwise comparison of the clustering results is shown by projecting the cluster label identified using one method onto the t-SNE visualization generated by another method (cluster vs. visualization). Black dash line circles highlight the rare pollutions (Sst, Pv, L6b and L6.CT) that were only identified by SnapATAC. Data source is listed in **Supplementary Table S1**.

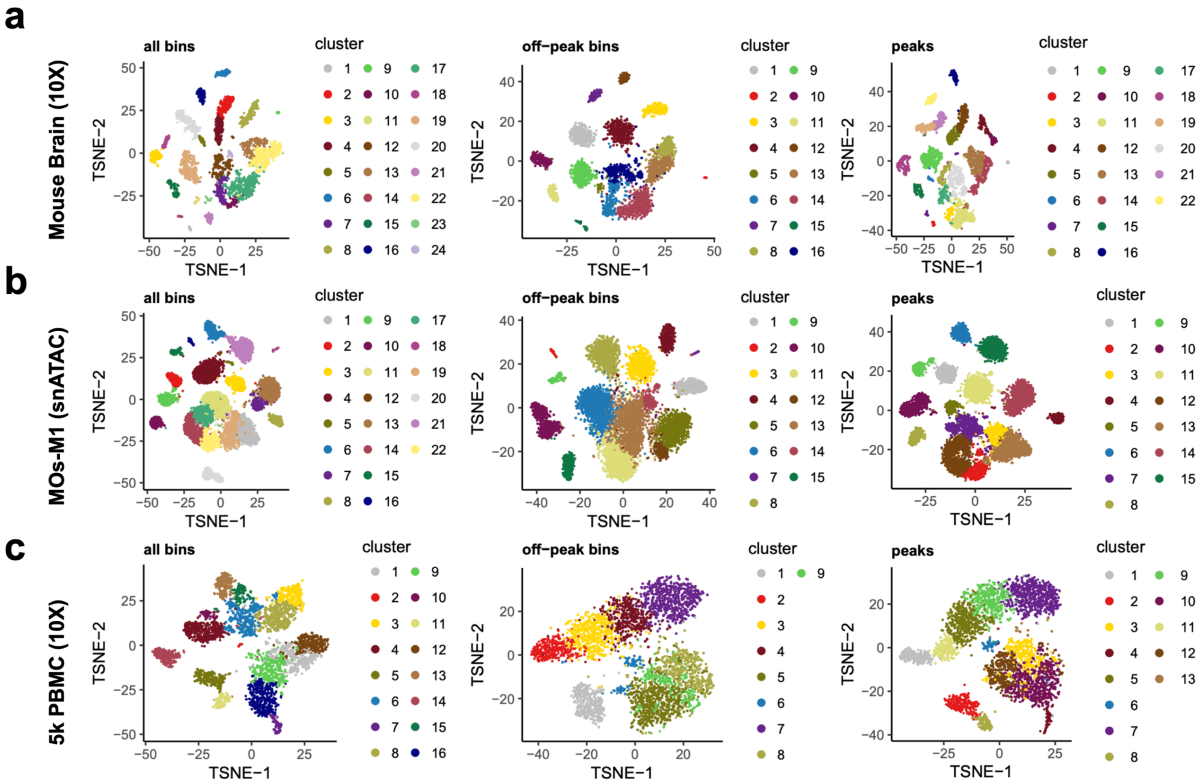


Figure S11. Off-peak reads distinguish major cell types in heterogenous samples. (**a-c**) SnapATAC clustering result on three benchmarking datasets using all bins versus clustering result only using bins that are not overlapped with peaks. Data source is listed in **Supplementary Table S1**.

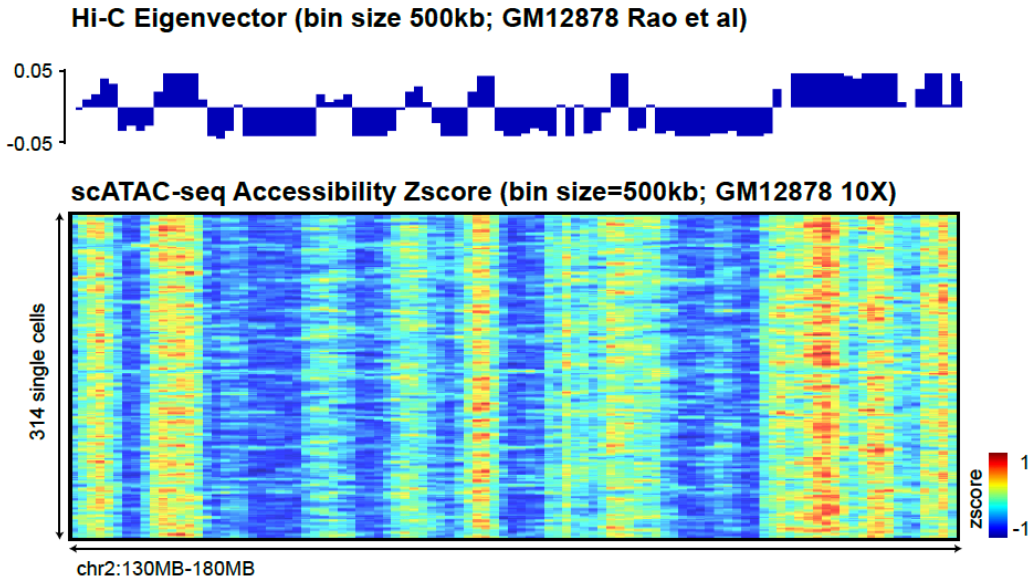


Figure S12. Off-peak reads reflect higher-order chromatin structure. At 500kb bin resolution, profile of compartments identified using Hi-C[58] in GM12878 overlaid the density of "off-peak" reads for 314 cells from GM12878 10X scATAC-seq library. Data source is listed in **Supplementary Table S1**.

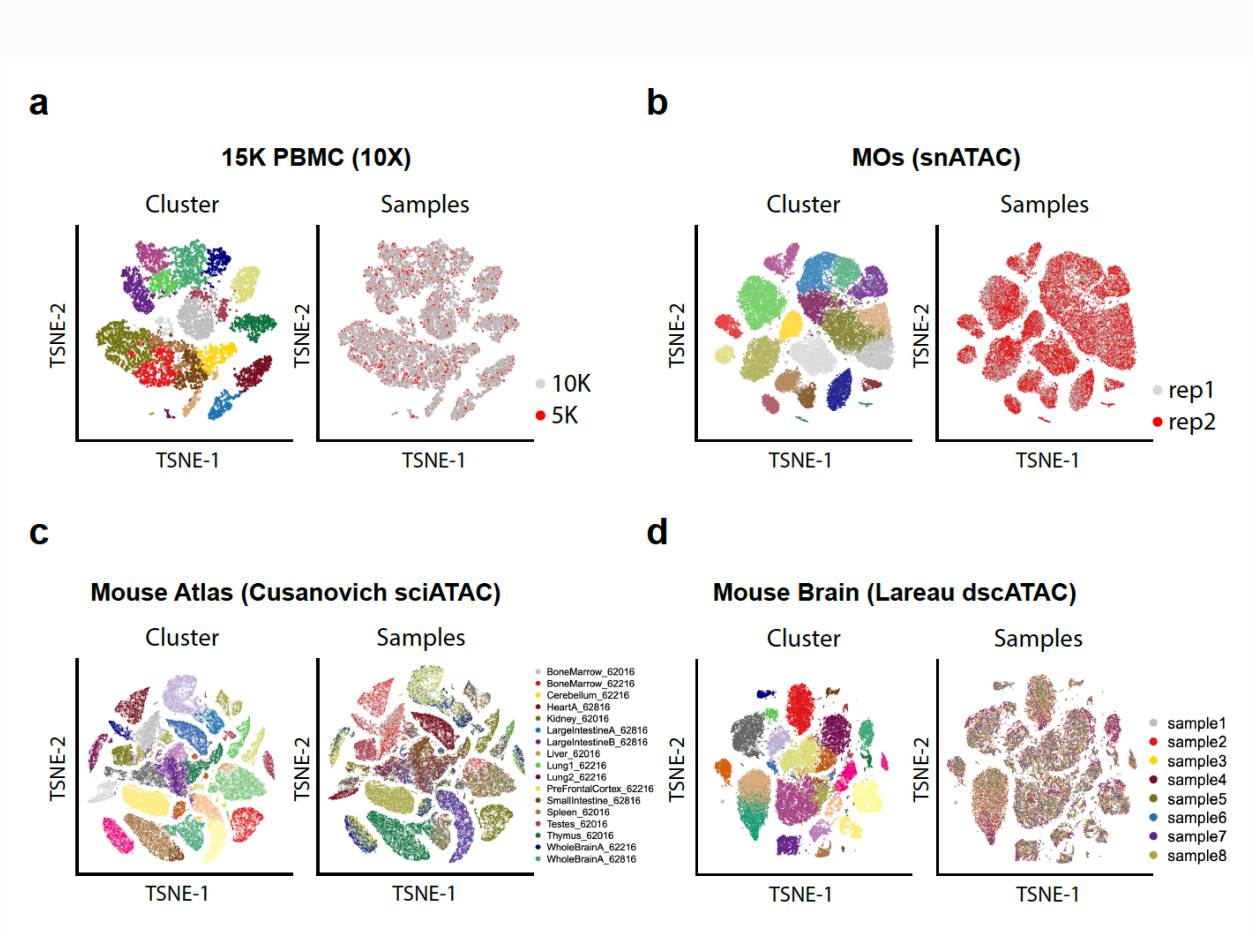


Figure S13. SnapATAC is robust to technical variation. Two-dimensional t-SNE visualization of four benchmarking datasets generated using SnapATAC. Cells are color by cluster label (left) and sample label (right). (**a**) 15k PBMC (10X) – a combination of two datasets (PBMC 5k and 10k) publicly available from 10X genomics. (**b**) MOs (snATAC) – an in-house dataset that contains two biological replicates from secondary motor cortex in the adult mouse brain generated using single nucleus ATAC-seq. (**c**) Mouse Atlas (Cusanovich 2018) – a published dataset that contains over 80K cells from 13 different mouse tissues generated using multiplexing single cell ATAC-seq. (**d**) Mouse Brain (Lareau dscATAC) – a published dataset that contains 46,652 cells from 8 samples in the adult mouse brain generated using BioRad droplet-based single cell ATAC-seq. Data source is listed in **Supplementary Table S1**.

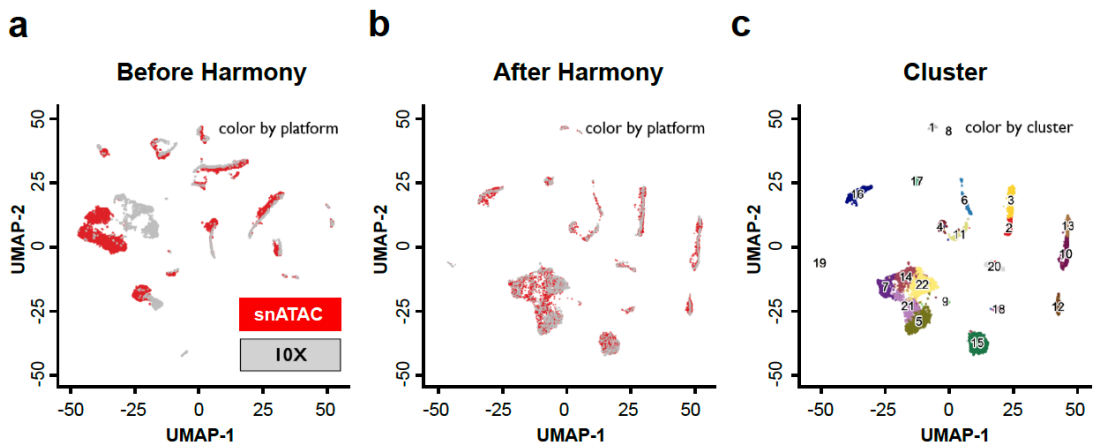
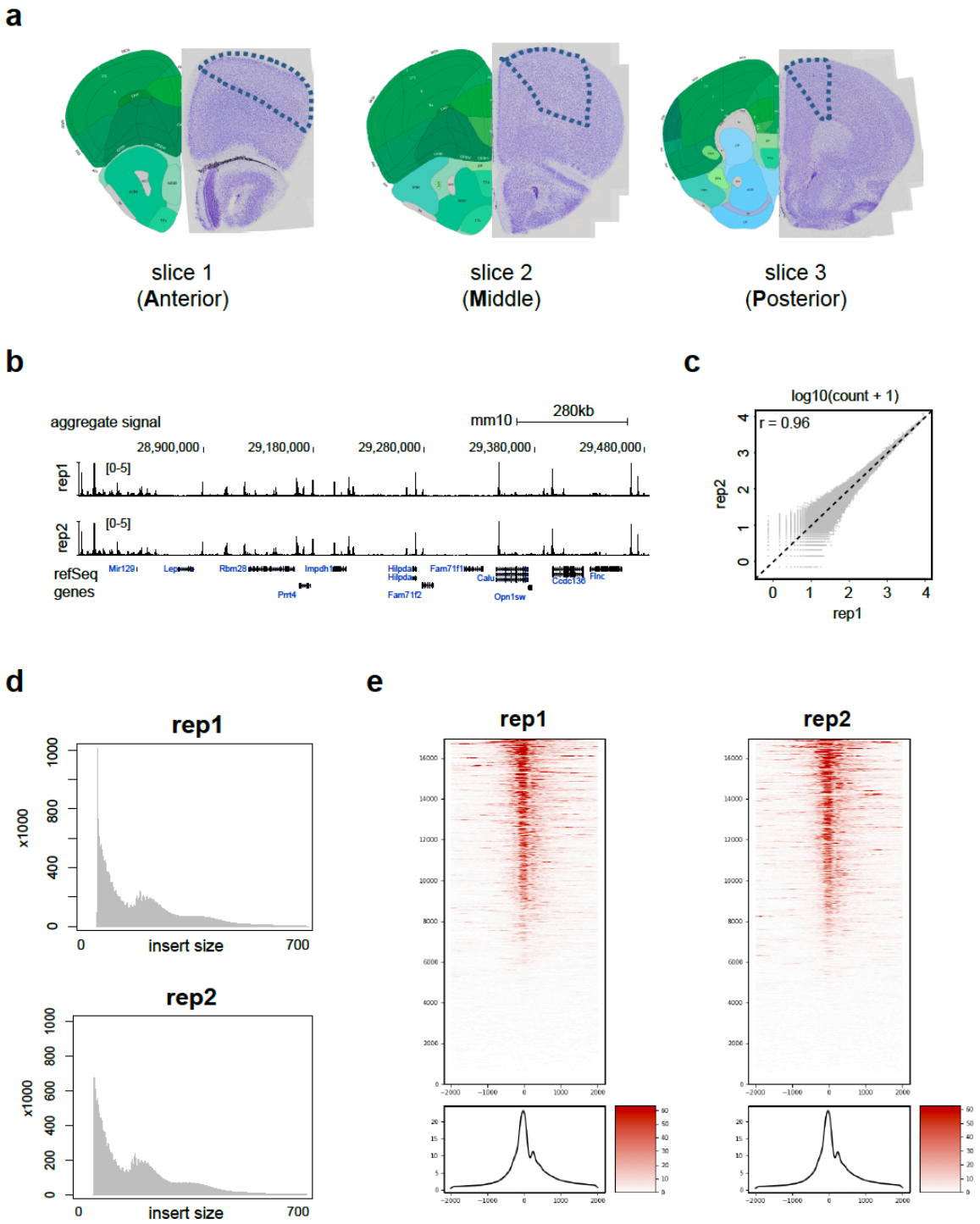


Figure S14. SnapATAC eliminates batch effect using Harmony[23]. The joint UMAP visualization of two datasets of mouse brain generated using combinatorial indexing single nucleus ATAC-seq (MOs-M1 snATAC) and droplet-based platform (Mouse Brain 10X) before (**a**) and after (**b**) performing batch effect correction using Harmony. Data source is listed in **Supplementary Table S1**.


a
slice 1
(Anterior)
slice 2
(Middle)
slice 3
(Posterior)
b
aggregate signal
mm10
280kb
28,900,000
29,180,000
29,280,000
29,380,000
29,480,000
rep1
[0-5]
rep2
[0-5]
refSeq
genes
Mir129
Lepr
Rbm28
Impdh1
Prnt4
Hilpda
Hilpda
Fam71f1
Fam71f2
Calu
Opn1sw
Cedc136
Flnc
c
log10(count + 1)
r = 0.96
rep2
rep1
d
rep1
x1000
insert size
rep2
x1000
insert size
e
rep1
rep2

1622 **Figure S15. Single nucleus ATAC-seq datasets are reproducible between**
1623 **biological replicates.** (**a**) Illustration of dissection. Posterior view of three 0.6 mm
1624 coronal slices from which the secondary motor cortex (MOs) was dissected. The right side
1625 on each image depicts the corresponding view from the Allen Brain Atlas. The left side
1626 correspond to the Nissl staining of the posterior side of each slice. The MOs region was
1627 manually dissected according to the dashed lines on each slice and following the MOs as
1628 depicted in plates 27, 33, and 39 of the Allen Brain Atlas (left side images in figure). Each
1629 slice contains two biological replicates named as A1, A2, M1, M2, P1 and P2 (A: Anterior;
1630 M: Middle; P: Posterior). In this study, A1, M1 and P1 is combined as replicate 1 and A2,
1631 M2 and P2 are combined as replicate 2. (**b**) Genome-browser view of aggregate signal for
1632 two biological replicates. (**c**) Pearson correlation of count per million (CPM) at peaks
1633 between two replicates. (**d**) Insert size distribution and (**e**) TSS enrichment score for two
1634 biological replicates.
1635

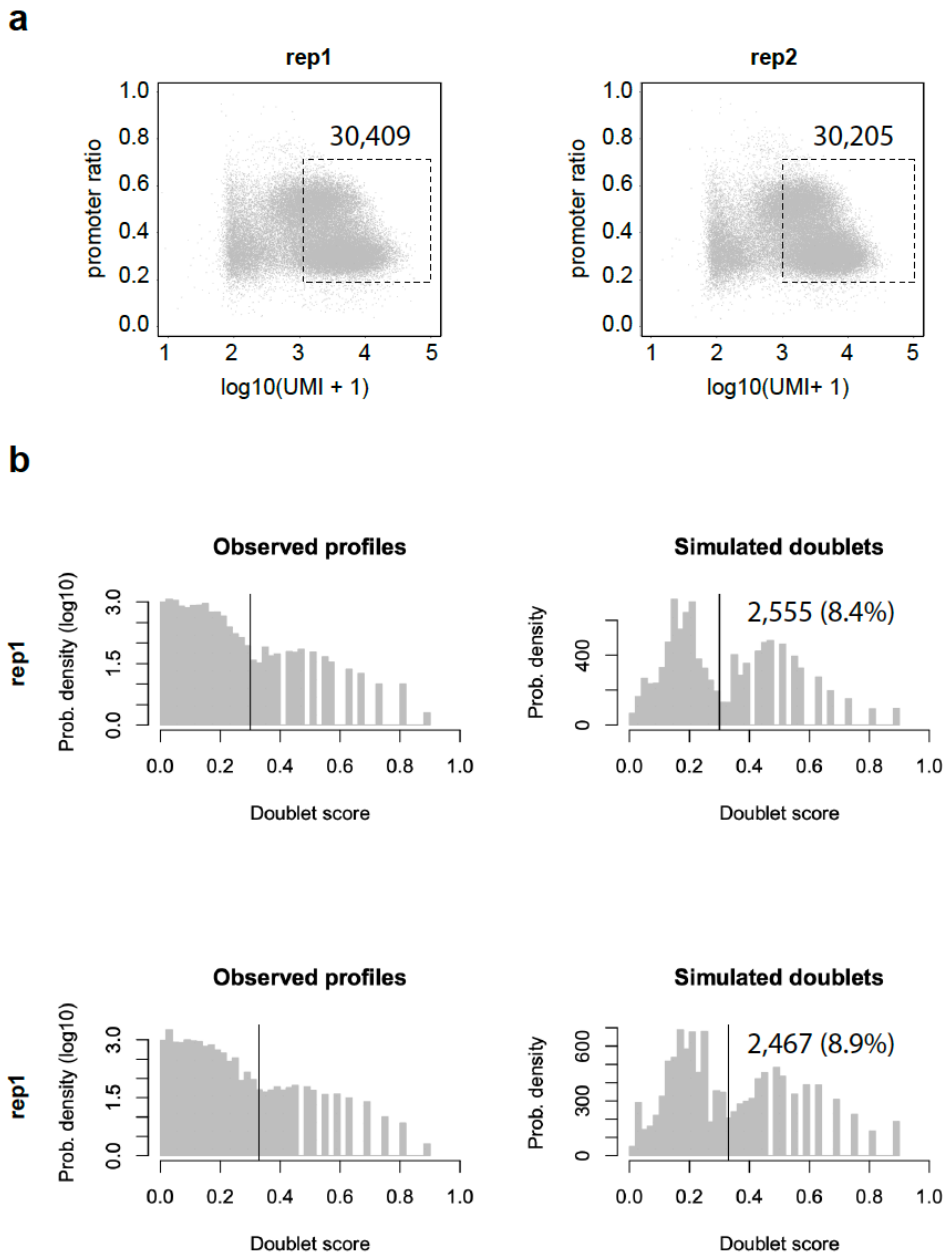


Figure S16. Barcode selection of MOs. (**a**) Cells of unique fragments within the range of 1,000-100,000 and fragments in promoter ratio within the range of 0.2-0.7 were selected. This resulted in 30,409 and 30,205 nuclei for two replicates. (**b**) With 5kb cell-by-bin matrix as input matrix, putative doublets were identified using Scrublets[37], which predicted 2,555 (8.4%) and 2,467 (8.9%) nuclei to be doublets for each replicate. The predicted doublet ratio is similar to the theoretical calculation of doublet ratio for multiplexing single cell ATAC-seq experiment[5,7].

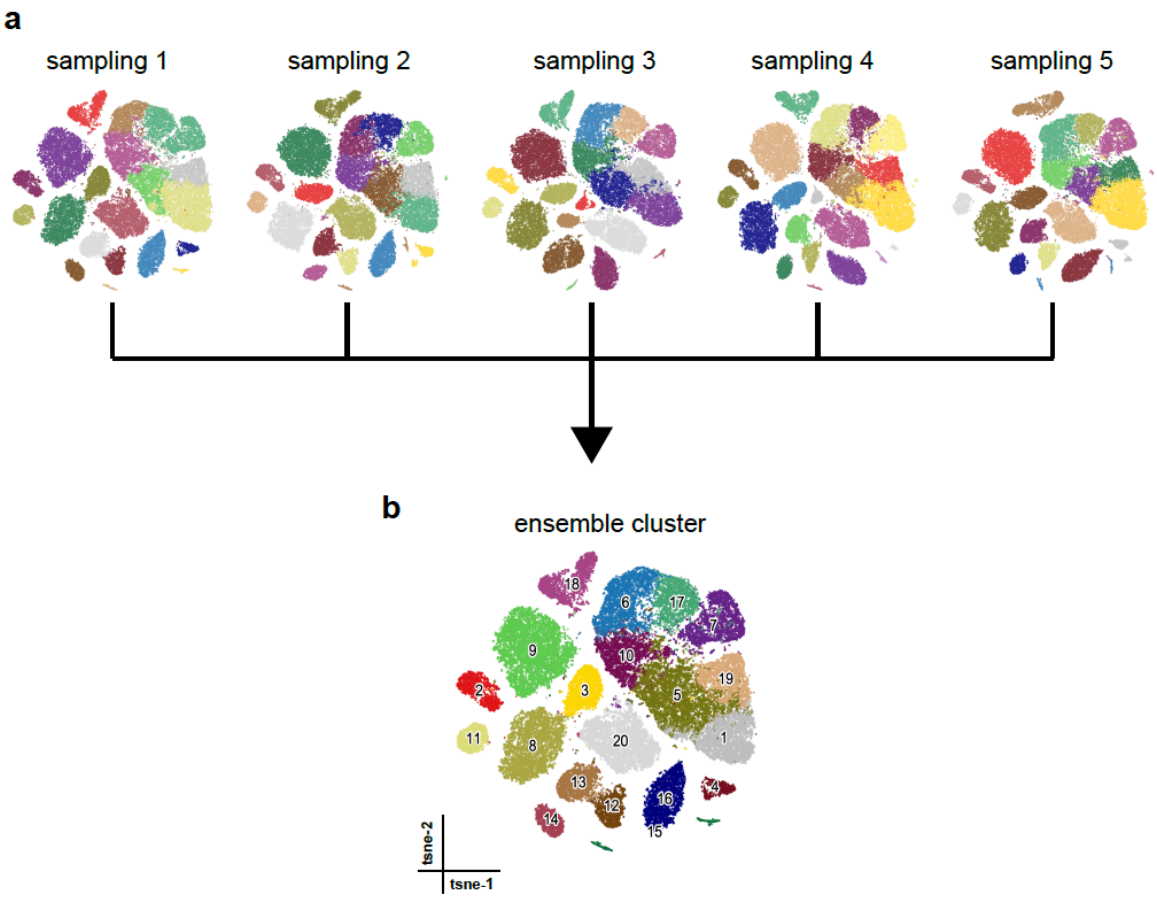


Figure S17. Consensus clustering of MOs. (**a**) Five clustering results were generated using SnapATAC with different set of landmarks (10,000). (**b**) These five clustering solutions were combined to create a consensus clustering which identified 20 clusters in MOs (**Supplementary Methods**).

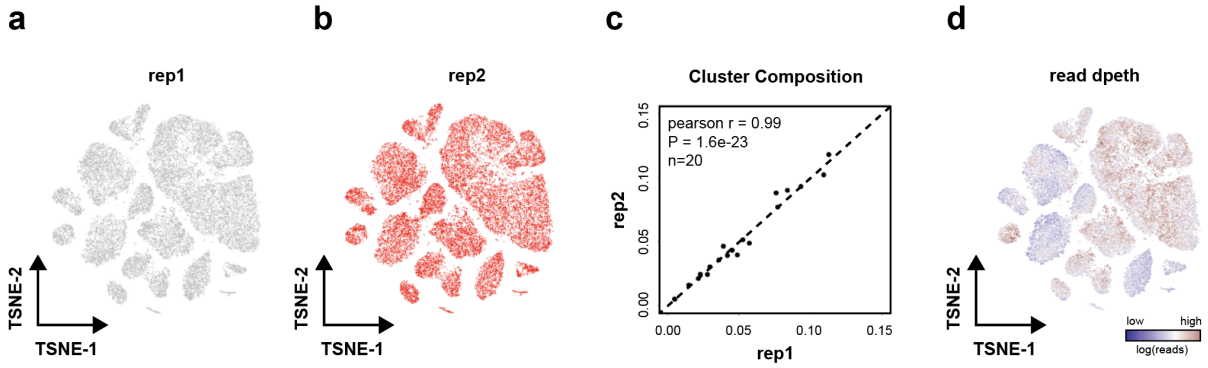


Figure S18. MOs clustering result is reproducible between biological replicates. (**a-b**) T-SNE visualization of cells from two biological replicates. (**c**) The cluster composition is highly reproducible between two biological replicates (r=0.99; P-value = 1.6e-23); (**d**) T-SNE visualization of cells with color scaled by sequencing depth.

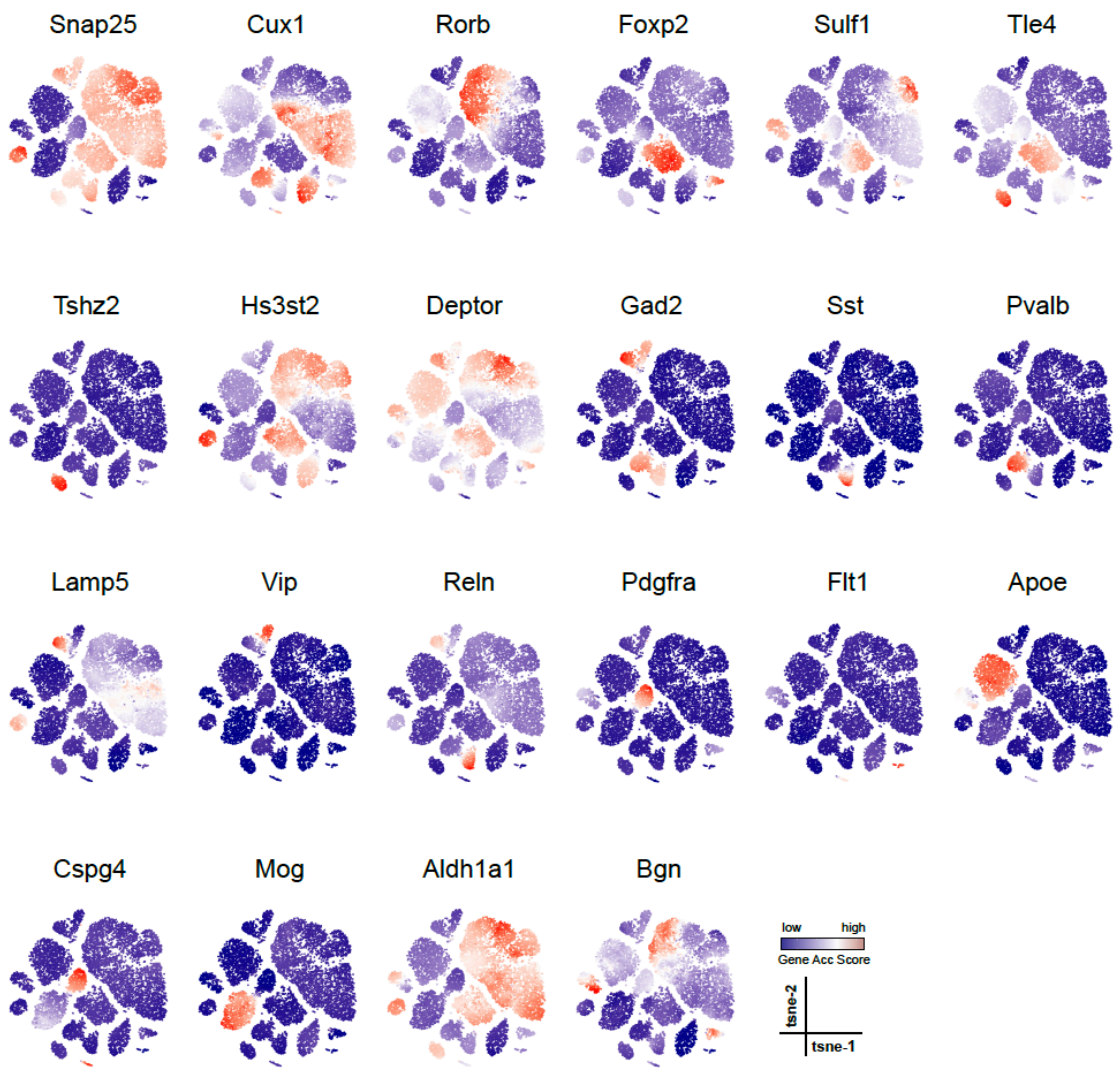


Figure S19. Gene accessibility score of canonical marker genes projected onto MOs t-SNE embedding to guide the cluster annotation. T-SNE is generated using SnapATAC for MOs; cell type specific marker genes was defined from previous single cell transcriptomic analysis in adult mouse brain[38]; gene accessibility score is calculated using SnapATAC (**Supplementary Methods**) and projected to the t-SNE embedding.

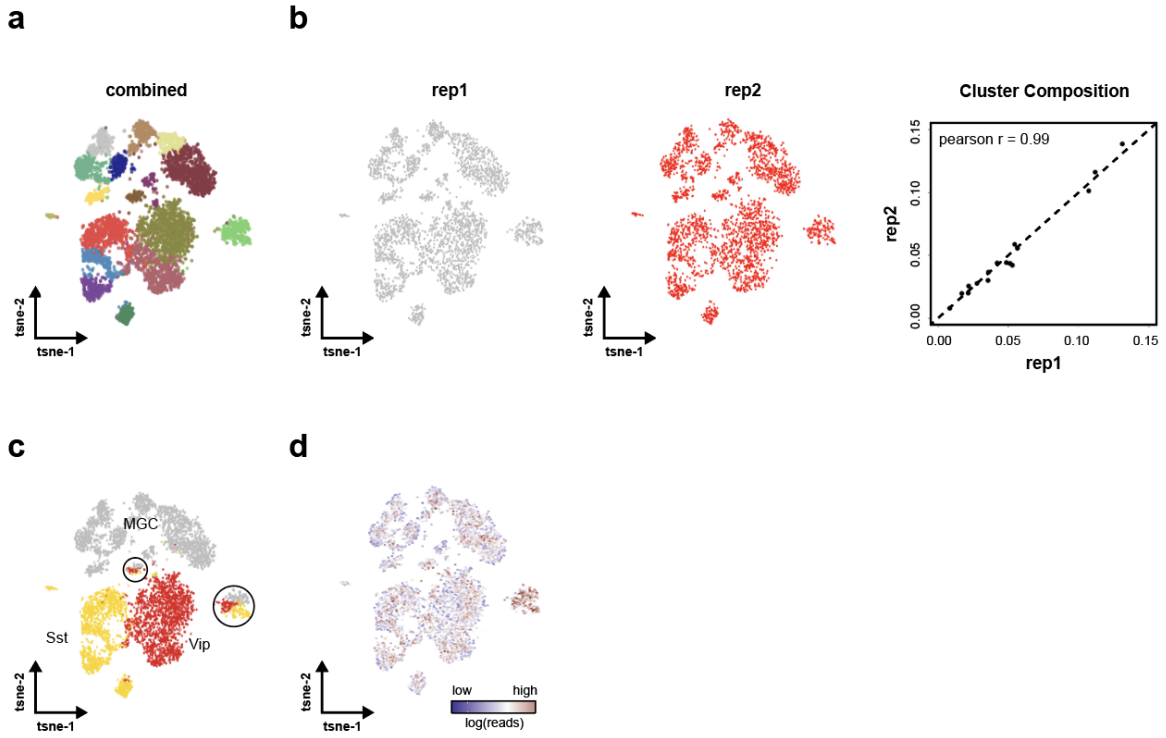
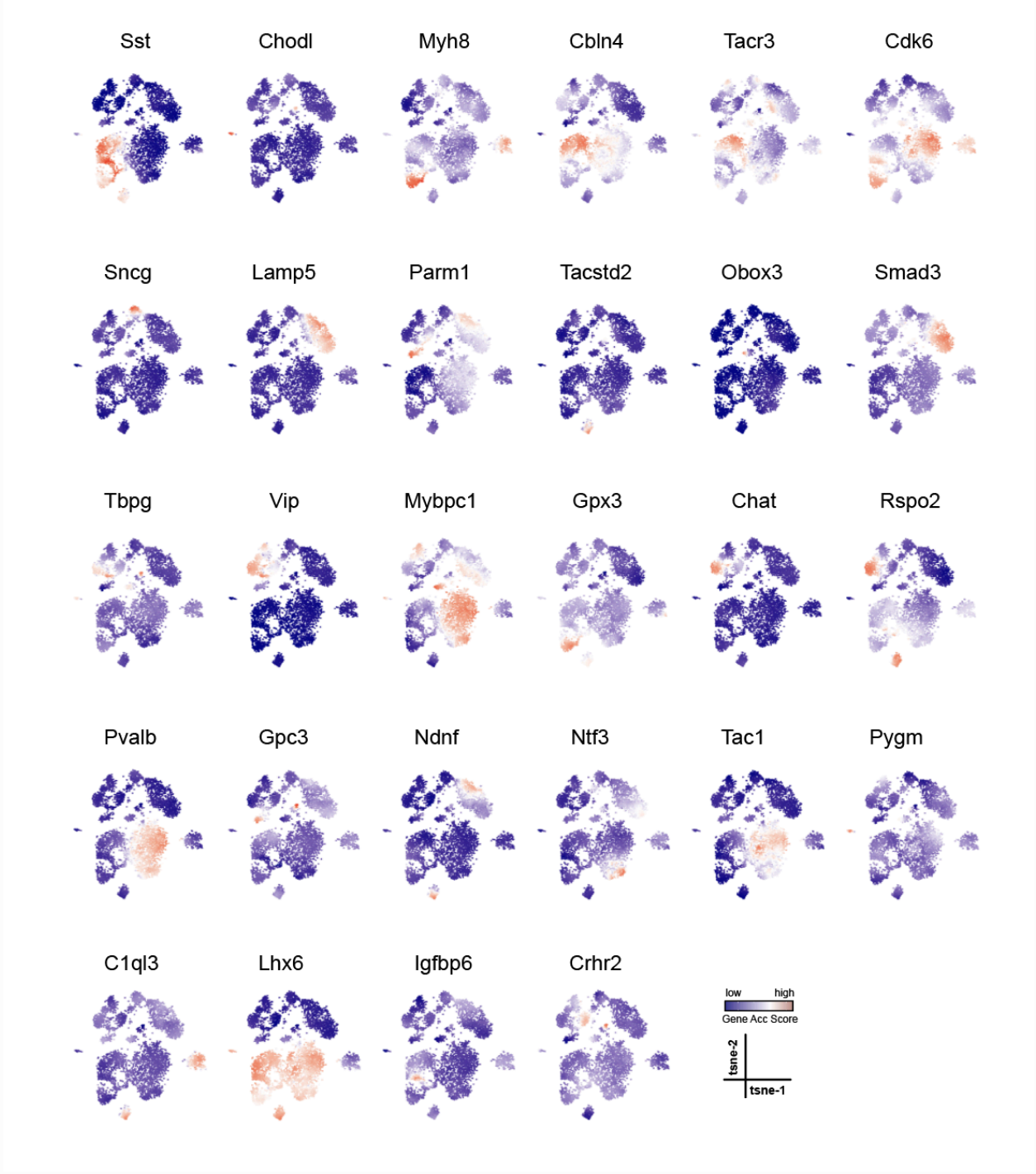


Figure S20. Iterative clustering identifies 17 GABAergic neuronal subtypes. (**a**) Sub-clustering of 5,940 GABAergic neurons identified 17 distinct cell clusters. (**b**) Cluster composition was highly reproducible between two biological replicates. (**c**) TSNE visualization of 5,940 GABAergic neurons colored by cell types identified in the initial clustering (shown in **Fig. 5a**). Black circles mark clusters that are potential doublets, a mixture of multiple cell types. (**d**) TSNE plot of GABAergic neurons colored by sequencing depth.


Sst
Chodl
Myh8
Cbln4
Tacr3
Cdk6
Sncg
Lamp5
Parm1
Tacstd2
Obox3
Smad3
Tbpg
Vip
Mybpc1
Gpx3
Chat
Rspo2
Pvalb
Gpc3
Ndnf
Ntf3
Tac1
Pygm
C1ql3
Lhx6
Igfbp6
Crhr2
low
high
Gene Acc Score
tsne-2
tsne-1

1680 **Figure S21. Gene accessibility score of marker genes projected onto t-SNE**
1681 **embedding from GABAergic neurons to guide the cluster annotation.** Iterative
1682 clustering is performed against GABAergic neurons to identify subtypes. Twenty eight cell
1683 type specific marker genes were defined from previous single cell transcriptomic analysis
1684 in adult mouse brain[38]; gene accessibility score is calculated using SnapATAC
1685 (**Supplementary Methods**).
1686
1687

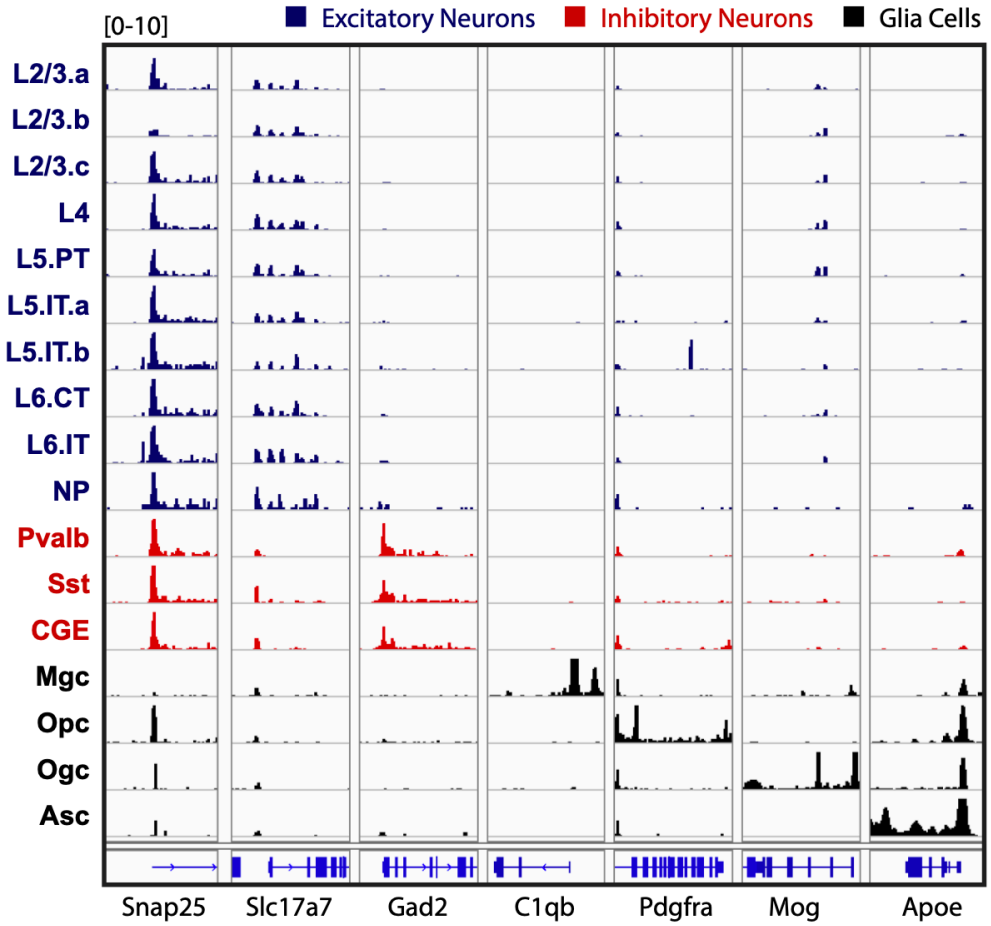


Figure S22. Genome browser view of aggregate signal for each of the major cell populations identified in the adult mouse brain (Fig. 5a).

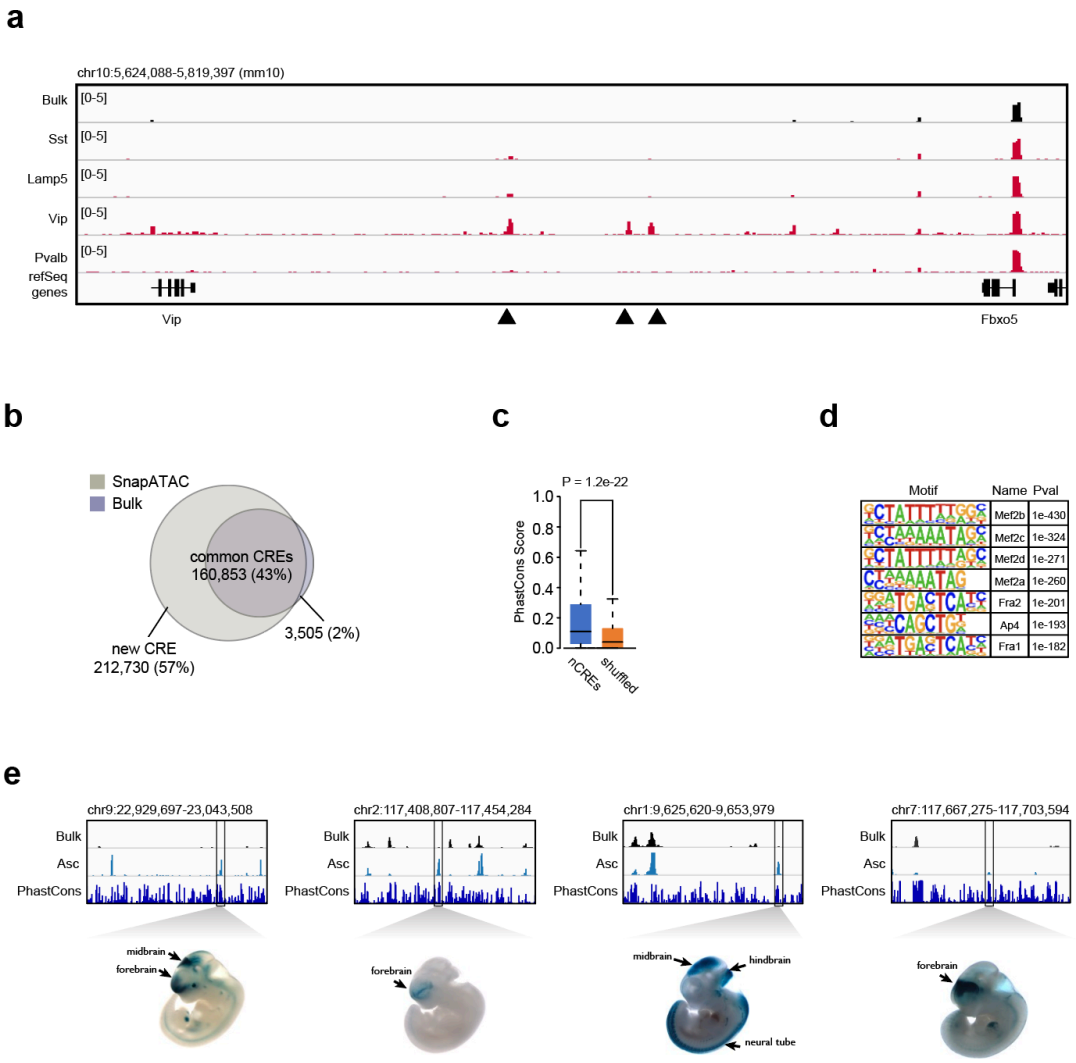


Figure S23. SnapATAC uncovers novel candidate *cis*-regulatory elements in rare cell types. (**a**) Genome browser view of 20Mb region flanking gene *Vip*. Dash line highlight five regulatory elements specific to Vip subtypes that are under-represented in the conventional bulk ATAC-seq signal. (**b**) Over fifty percent of the regulatory elements identified from 20 major cell populations are not detected from bulk ATAC-seq data. (**c**) Sequence conservation comparison between the new elements and randomly chosen genomic regions. (**d**) Top seven motifs enriched in Pv-specific new elements. (**e**) Examples of four new elements that were previously tested positive in transgenic mouse assays (from VISTA database). Bulk: Bulk ATAC-seq; Asc: aggregated signal from astrocyte population (ASC) in the adult mouse brain as shown in **Fig. 5a**.

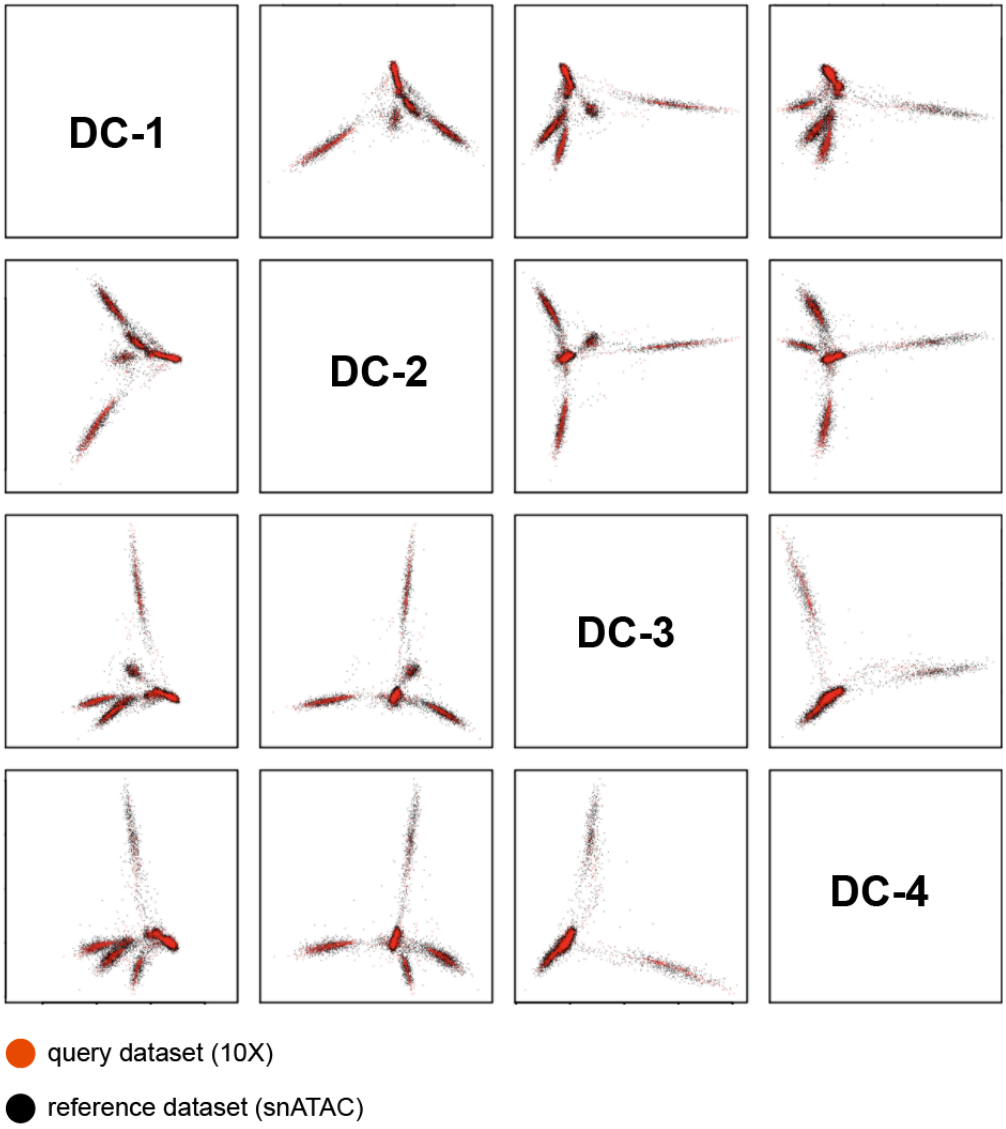


Figure S24. Joint embedding for query (Mouse Brain 10X) and reference dataset (MOs snATAC). The query dataset (10X) is projected onto the low dimension embedding space precomputed for the reference dataset (snATAC). Batch effect is corrected using Harmony. Pairwise plot of the first four dimentions in which cells are colored by dataset - red for query cells (Mouse Brain 10X) and black for reference cells (MOs snATAC). Data source as listed in **Supplementary Table S1**.

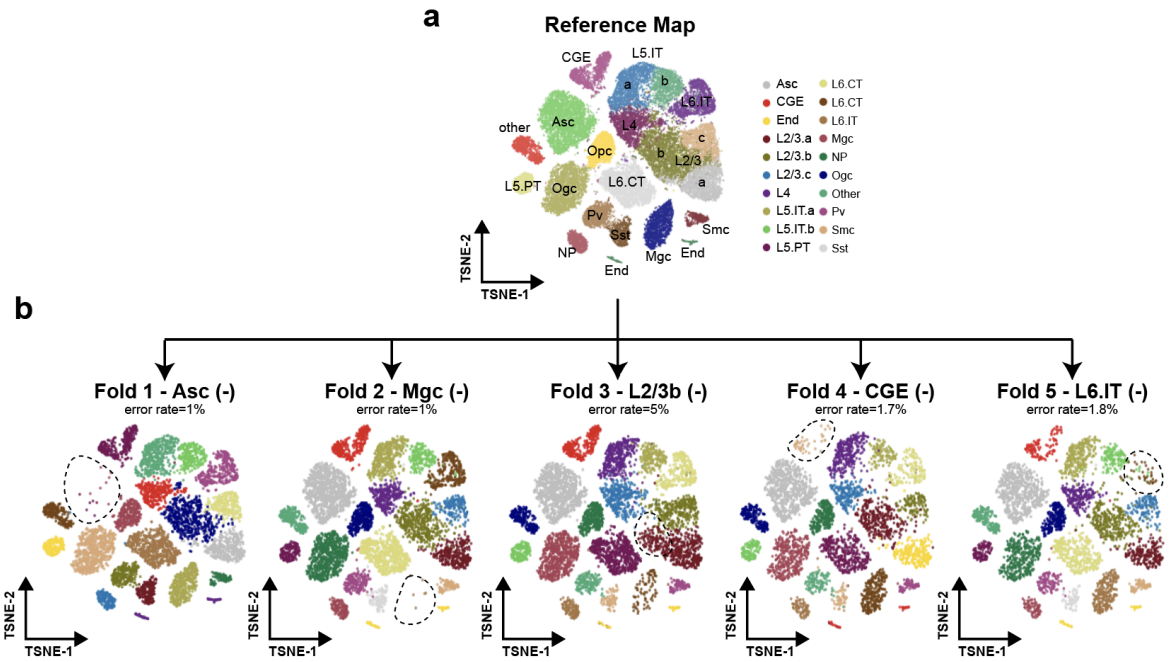


Figure S25. SnapATAC is robust for supervised annotation of datasets containing cell types missing in the reference atlas. (**a**) Two-dimensional t-SNE visualization of the reference dataset MOs (snATAC). (**b**) A five-fold cross validation is performed to this reference dataset. For each fold, we introduce perturbation to the 80% training dataset by randomly dropping one cell type (Asc, Mgc, L2/3b, CGE and L6.IT). We then predict on the 20% test dataset using the model learned from the perturbed training dataset. The prediction accuracy for each fold is shown in (**b**) and cell type removed from the training dataset are highlighted by the dash-line circles.

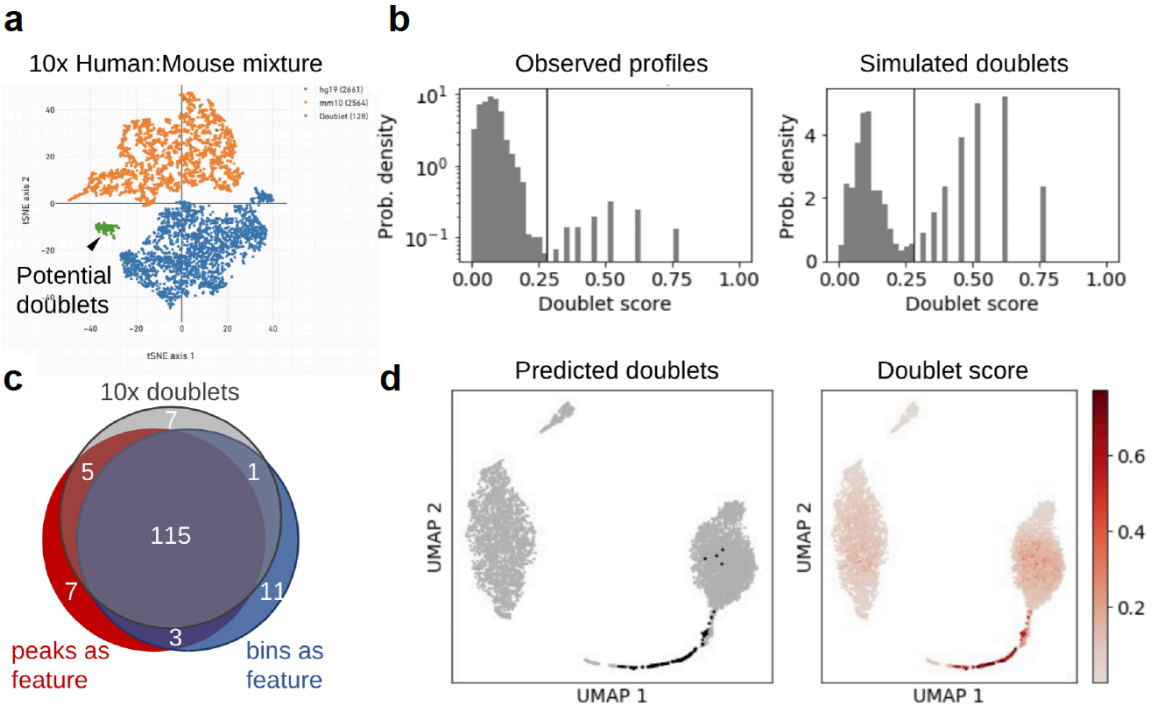


Figure S26. Doublets detection using Scrublet. (**a**) T-SNE representation of a dataset (hgmm_1k 10X) that contained 1,000 human (GM12878) and mouse (A20) cells. Cells are colored by species determined based on the alignment ratio between human and mouse genome. Orange: A20; blue: GM12878; green: putative doublets. (**b**) Distribution of doublet score for putative doublets and simulated doublets estimated using Scrublet[37]. (**c**) Doublets are predicted using cell-by-peak and cell-by-bin matrix separately. Venn diagram show the overlap between Scrublet-predicted doublets using peak or bin matrix and doublets identified based on alignment ratio. (**d**) Doublets scores projected onto the UMAP embedding.