Fast and Accurate Clustering of Single Cell Epigenomes Reveals *Cis*-Regulatory Elements in Rare Cell Types

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

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. Department of Cognitive Science, University of California, San Diego, La Jolla, CA 92037, USA
7. Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
8. Molecular Medicine, Institute of Genomic Medicine, UCSD Moores Cancer Center, La Jolla, CA 92093, USA

*Correspondence to: biren@ucsd.edu

Abstract:

Mammalian tissues are composed of highly specialized cell types defined by distinct gene expression patterns. Identification of *cis*-regulatory elements responsible for cell-type specific gene expression is essential for understanding the origin of the cellular diversity. Conventional assays to map *cis*-elements via open chromatin analysis of primary tissues fail to resolve their cell type specificity and lack the sensitivity to identify *cis*-elements in rare cell types. Single nucleus analysis of transposase-accessible chromatin (ATAC-seq) can overcome this limitation, but current analysis methods begin with pre-defined genomic regions of accessibility and are therefore biased toward the dominant population of a tissue. Here we report a method, Single Nucleus Analysis Pipeline for ATAC-seq (SnapATAC), that can efficiently dissect cellular heterogeneity in an unbiased manner using single nucleus ATAC-seq datasets and identify candidate regulatory sequences in constituent cell types. We demonstrate that SnapATAC outperforms existing methods in both accuracy and scalability. We further analyze 64,795 single cell chromatin profiles from the secondary motor cortex of mouse brain, creating a chromatin landscape atlas with unprecedent resolution, including over 300,000 candidate *cis*-regulatory elements in nearly 50 distinct cell populations. These results demonstrate a systematic approach for comprehensive analysis of *cis*-regulatory sequences in the mammalian genomes.

Introduction

Mammalian tissues comprise of various cell types highly specialized to carry out distinct functions. Cellular identity and function are established and maintained through programs of gene expression that are specific to each cell type and state[1]. Gene regulation is carried out by sequence-specific transcription factors that interact with *cis*-regulatory sequences, such as promoters, enhancers and insulators[2]. Identifying *cis*-regulatory elements in the genome is an essential step towards understanding the cell type specific gene regulatory programs in mammalian tissues.

Since the activity of *cis*-elements often arises from the binding of transcription factors to accessible chromatin, approaches such as ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing)[3] and DNase-seq (DNase I hypersensitive sites sequencing)[4] that identify regions of open chromatin have been widely used to map candidate regulatory sequences in the genomes. However, these conventional assays have limited ability to resolve the diverse cell type-specific chromatin landscapes present in heterogeneous tissues, providing only an average map dominated by signals from the most common cell populations.

Recently, a number of methods have been developed for measuring chromatin accessibility in single cells. One approach involves combinatorial indexing to simultaneously process tens of thousands of cells[5]. This strategy has been successfully applied to embryonic tissues in *D. melanogaster*[6], developing mouse forebrains[7] and multiple adult mouse tissues[8]. A related method, called scTHS-seq (single-cell transposome hypersensitive site sequencing), has also been developed and used to study chromatin landscapes at single cell resolution in the adult human brains[9]. Another approach relies on isolation of single cell using microfluidic devices (Fluidigm, C1)[10] or within individually indexable wells of a nano-well array (Takara Bio, ICELL8)[11]. Whereas fewer cells are processed per experiment compared to the combinatorial indexing approach, the library complexity per single cell is considerably higher with this method[12]. Recently, 10X Genomics and Bio-Rad Laboratories have enabled single cell ATAC-seq on droplet-based microfluidic platform, producing data of similar quality to that of nano-well capture technique[12]. Despite these experimental advances,

data from single cell chromatin accessibility experiments still presents unique computational challenges largely due to the sparsity and high-level noise of the data from single cells.

Existing computational methods rely on pre-defined regions of transposase accessibility identified from the aggregate signals. For instance, chromVAR[13] estimates similarity between cells based on transcription factor occurrence frequency in the peak regions. Alternatively, techniques developed for natural language processing have been applied to scATAC-Seq data by treating each single cell profile as a document, composed of regions of chromatin accessibility which play the role of words. In this framework, Latent Semantic Analysis (LSA)[8] and Latent Dirichlet Allocation (Cis-Topic)[14] infer the relationships between cells. A third approach, Cicero, clusters cells based on the gene activity scores predicted by linking distal or proximal peaks to the gene[15]. Relying on gene activity scores predicted by Cicero, a recent approach attempts to classify individual nuclei from a scATAC-seq dataset based on a reference of transcriptomic states[16].

The use of pre-defined accessibility peaks based on bulk data has at least three key limitations. First, it requires sufficient number of single cell profiles to create robust aggregate signal for peak calling. Second, the cell type identification is biased toward the most abundant cell types in the tissues. Finally, these techniques lack the ability to reveal regulatory elements in the rare cell populations which are underrepresented in the aggregate signal. This concern is critical, for example, in brain tissue, where key neuron types may represent less than 1% of all cells while still playing a critical role in the neural circuit[17].

To overcome these limitations, we developed a bioinformatic package, Single Nucleus Analysis Pipeline for ATAC-seq (SnapATAC), for analyzing single cell ATAC-seq (scATAC-seq) datasets. SnapATAC does not require population-level peak annotation, and instead assembles chromatin landscapes by directly clustering cells based on the similarity of their genome-wide accessibility profile. Using a regression-based normalization procedure, SnapATAC adjusts for differing read depth between cells. With a fast dimensionality reduction technique, it can easily process data from millions of cells. In a

109 battery of tests using simulated and published datasets, SnapATAC outperforms existing
110 tools in both clustering accuracy and scalability. To demonstrate the utility of SnapATAC,
111 we apply it to a dataset of over 60,000 single cell ATAC-seq profiles from the mouse
112 secondary motor cortex that we generated. We detect nearly 50 subtypes including some
113 rare types that account for less than 0.1% of the total population. We also uncover
114 337,932 candidate *cis*-elements in these different cell types, more than twice as many as
115 were identified from bulk analysis. These results suggest that SnapATAC, together with
116 scATAC-seq, can greatly enhance our ability to annotate and characterize the *cis*-
117 regulatory elements in the mammalian genomes.
118

119 **Results**

120 **SnapATAC achieves a new standard for scATAC-seq analysis**

121 A schematic diagram of SnapATAC is shown in **Fig. 1**. Briefly, after pre-processing
122 (**Methods**), the chromatin accessibility profile of each single cell is represented as a
123 binary vector, the length of which corresponds to the number of uniform-sized bins that
124 segmented the genome. A bin with value "1" indicates that one or more reads fall within
125 that bin, and the value "0" indicates otherwise. Next, the set of binary vectors from all the
126 cells is converted into a Jaccard index matrix, with the value of each element calculated
127 from fraction of overlapping bins between every two cells. Since the number of cells is
128 usually far smaller than the number of bins, this operation effectively reduces the
129 dimensions of the matrix therefore significantly improves the scalability of the pipeline
130 (**Methods**). Because the value of Jaccard Index can be influenced by differing sequencing
131 depth between cells (**Supplementary Fig. 1**), therefore, a normalization method is
132 developed to remove such confounding factor (**Methods**; **Supplementary Fig. 1-2**).
133 Next, the normalized matrix is subject to Principal Component Analysis (PCA) and the
134 significant components are selected to create a K-nearest neighbor (KNN) graph, with
135 edges drawn between cells with similar ATAC-seq profiles. The highly interconnected
136 'communities' (or 'clusters') of cells in the resulting graph are identified using Louvain
137 algorithm[18]. Cells belonging to each cluster are pooled to assemble a consensus chromatin
138 landscape for identification of regulatory elements *de novo*. Finally, using candidate
139 regulatory elements in each cluster, the master regulators for each cell cluster are inferred
140 by motif analysis[19] and the potential function of the cluster is predicted with Genomic

141 Regions Enrichment of Annotation Tool (GREAT) analysis[20]. Thus, SnapATAC provides
142 an end-to-end solution for analysis of single cell ATAC-seq datasets.
143
144

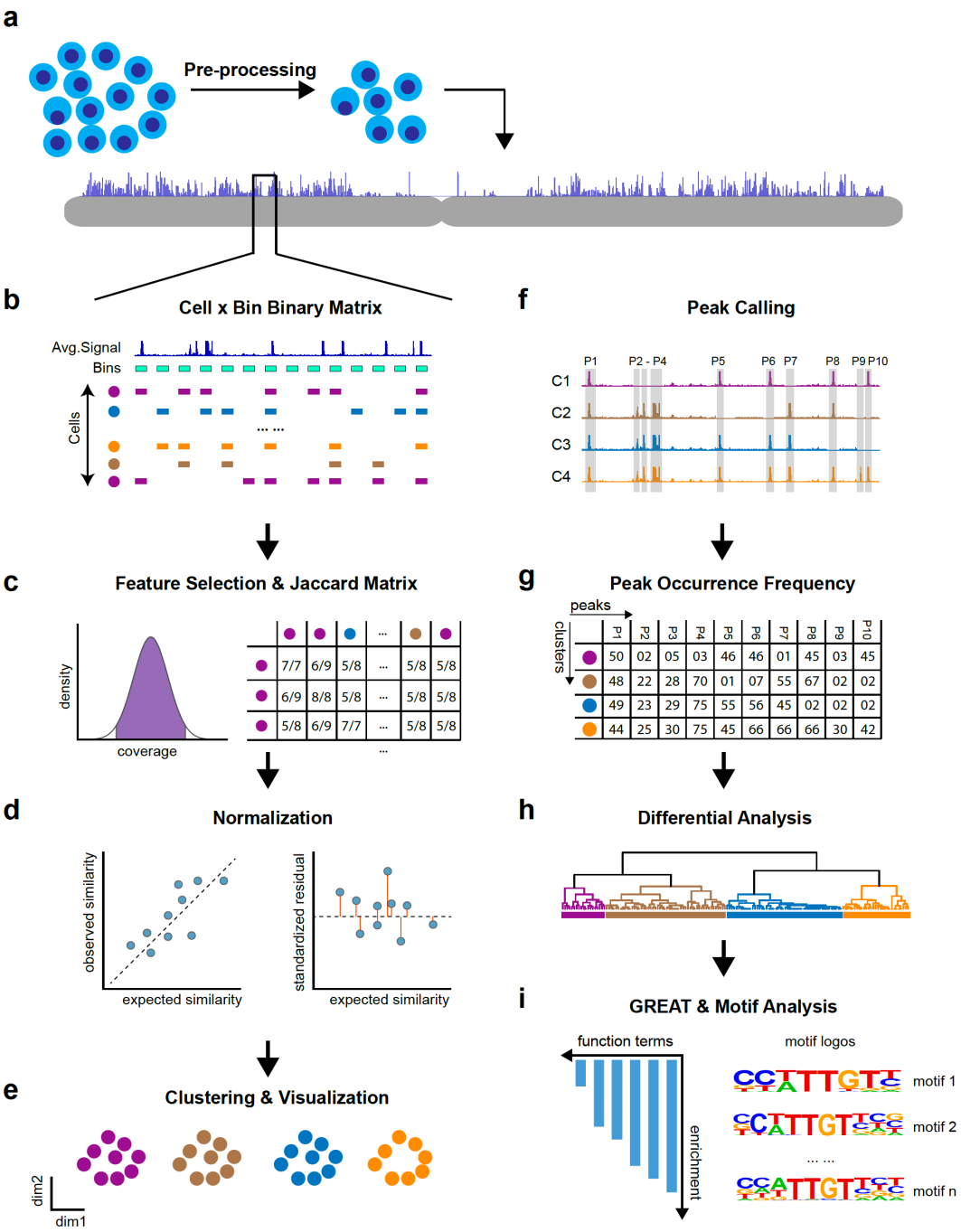


Figure 1. Overview of SnapATAC workflow. (**a**) Pre-processing: SnapATAC takes raw sequencing reads as input and aligns them to the reference genome followed by filtration of low-quality cells. (**b**) Cell-by-Bin Binary Matrix: the genome is segmented into uniform-sized bins and single cell profiles are represented as a binary matrix with "1" indicating a specific bin is accessible in a given cell and "0" denoting inaccessible

151 chromatin or missing data. (**c**) Feature Selection & Jaccard Index Matrix: after filtering
152 undesirable bins, the genome-wide cell-by-bin matrix is converted into a Jaccard index
153 matrix by estimating similarity between cells in the basis of profile overlaps. (**d**)
154 Normalization: Jaccard similarity matrix is normalized using a regression-based
155 method to eliminate the read depth effect. (**e**) Clustering: using normalized matrix, cells
156 of similar accessibility profiles are clustered together and visualized using t-SNE (t-
157 Distributed Stochastic Neighbor Embedding) or UMAP (Uniform Manifold
158 Approximation and Projection for Dimensionality Reduction). (**f**) Peak Calling: cells
159 belonging to the same cluster are aggregated to create a representation of cell-type
160 specific regulatory landscape for identification of candidate *cis*-regulatory elements *de*
161 *novo*. (**g**) Peak Occurrence Frequency Matrix: the frequency (number of cells out of the
162 total) of a peak occurring in each cluster is calculated. (**h**) Differential Analysis:
163 differential analysis performed to identify cell-type specific regulatory elements. (**i**)
164 GREAT & Motif Analysis: using cell-type specific regulatory elements, GREAT (Genomic
165 Region Enrichment of Annotation Tool) analysis performed to predict the potential
166 function of each cluster and motif analysis to reveal candidate master regulators that
167 controls gene expression in each cell type.
168
169 The performance of SnapATAC is benchmarked against a variety of published scATAC-
170 seq analysis methods, including chromVAR[13], LSA[8], Cicero[15] and Cis-Topic[14]. To allow for
171 evaluation of the clustering performance as a function of data sparsity, a set of simulated
172 single-cell ATAC-seq datasets were generated by down sampling from 10 previously
173 published bulk ATAC-seq datasets[21] (**Supplementary Table 1**) with varying coverages,
174 from 10,000 reads per cell (high coverage), to 1,000 reads per cell (low coverage)
175 (**Methods**). The performance of each method in identifying the original cell types was
176 measured by the normalized mutual index (NMI), which ranges from 0 for a level of
177 similarity expected by chance to 1 for perfect clustering. This analysis shows that
178 SnapATAC is the most robust and accurate method across all ranges of data sparsity (**Fig.**
179 **2**) (Wilcoxon signed-rank test, $P < 0.01$; **Supplementary Fig. 3**; **Supplementary**
180 **Table 2**).
181

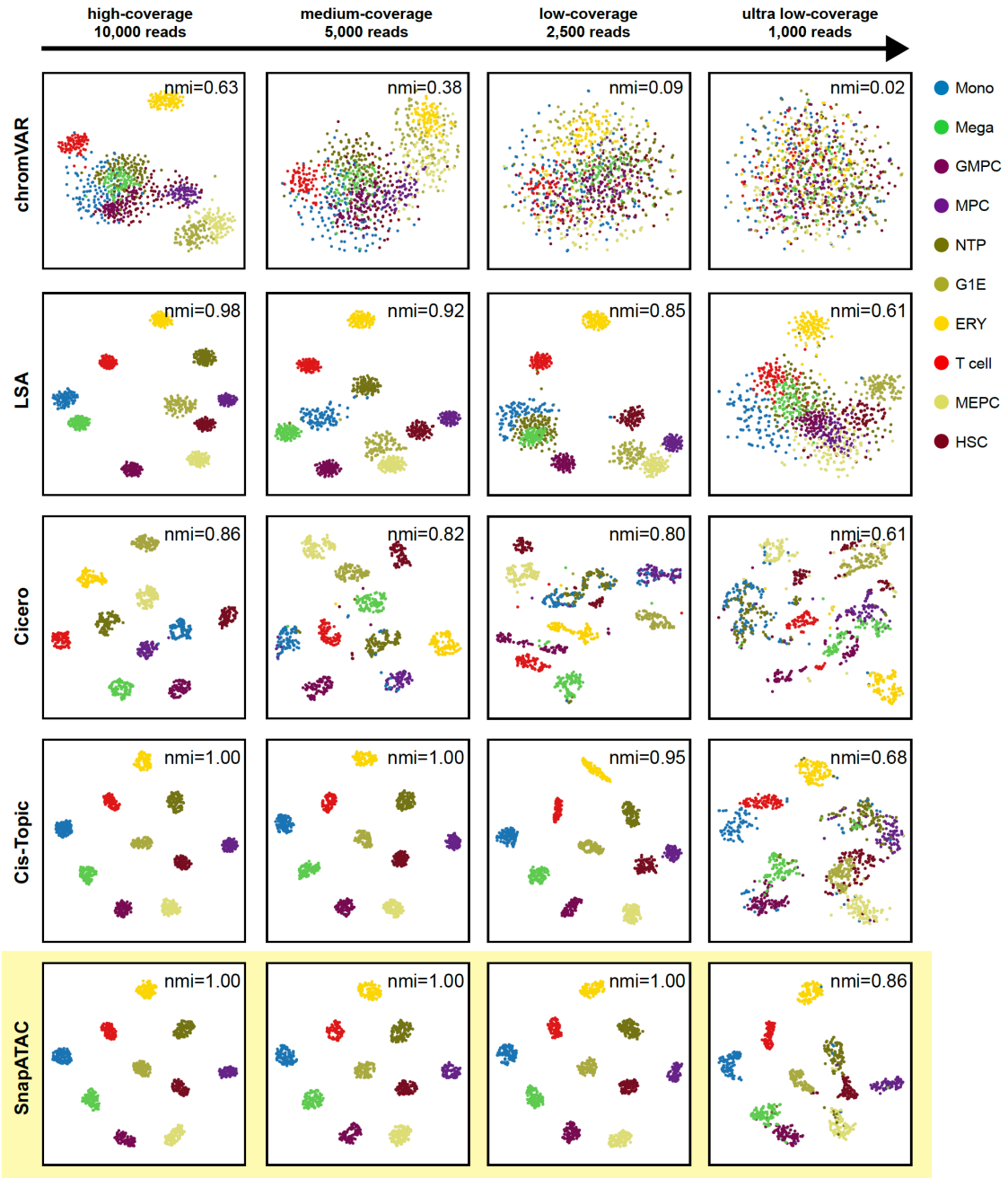


Figure 2. Validation of SnapATAC performance relative to alternative methods on simulated datasets. Method comparison on 2,000 simulated single cell ATAC-seq data down sampled from 10 bulk ATAC-seq datasets with varying coverages. Mono: monocyte; Mega: megakaryocyte; GMPC: granulocyte monocyte progenitor cell; MPC: megakaryocyte progenitor cell; NPT: neutrophil; G1E: G1E; T cell: regulatory T cell; MEPC: megakaryocyte-erythroid progenitor cell; HSC: hematopoietic stem cell.

189 The superior performance of SnapATAC likely results from the fact that it considers all
190 reads from each cell, not just a small fraction of reads that fall within peaks identified
191 from aggregate signals. To test this hypothesis, the above analysis was repeated but only
192 using off-peak reads. Consistent with this hypothesis, all cells were clustered perfectly
193 (nmi=1.0; **Supplementary Fig. 4**). It is likely that these off-peak reads 1) overlap with
194 "weak" elements that are not identified from the aggregate signals; 2) may be enriched for
195 the euchromatin, which strongly correlate with active genes[22] and vary considerably
196 between cell types[23]. Supporting this hypothesis, the density of 70% off-peak reads
197 correlates strongly with compartment A defined in the particular cell types through
198 genome-wide chromatin conformation capture analysis (i.e. Hi-C) (**Supplementary Fig.**
199 **5**). These observations suggest that the superior performance of SnapATAC with low-
200 coverage datasets is, at least in part, due to that the off-peak sequencing reads in the
201 scATAC-seq library contribute significantly for cell clustering.
202
203 To further assess the performance of SnapATAC, it is used to analyze a series of published
204 single cell chromatin accessibility datasets representing a variety of sample types, and the
205 results are compared to the original analysis. When applied to a set of 1,452 human cells
206 corresponding to 10 distinct cell types[13], SnapATAC successfully uncovered distinct cell
207 populations with an accuracy of 0.95 (normalized mutual index) according to the cell
208 labels, while the original method (chromVAR) failed to fully distinguish several blood cell
209 subtypes (HL60, Blast, LMPP, LSC and Mono) (**Fig. 3a; Supplementary Fig. 6a**).
210 Interestingly, SnapATAC divided K562 cells into two sub-clusters (**Fig. 3a**) (labeled as
211 K562.2015, K562.2017), corresponding to the years (2015 and 2017, respectively) when
212 the cells were grown and profiled (**Supplementary Fig. 6b-c**). In addition, GM12878
213 cells were also split into two separate clusters (GM12878.a and GM12878.b)
214 (**Supplementary Fig. 6b**) that represent previously identified subtypes associated with
215 differential NF-kB activity and B cell signaling[5] (**Supplementary Fig. 6d**). Taken
216 together, these results indicate that SnapATAC is a sensitive and accurate method to
217 distinguish different cell types.
218
219 When applied to datasets from more complex tissues, SnapATAC also exhibits much
220 improved performance over previous methods. Reanalyzing single-cell open chromatin

profiles (scTHS-seq) from 6,008 human Occipital Lobe[9], SnapATAC uncovered two additional inhibitory neuron subpopulations and four more excitatory subtypes corresponding to different layers (**Fig. 3b**) that were previously undetectable without incorporating single cell RNA sequencing data[9]. Similarly, when applied to a sci-ATAC-seq dataset comprising ~100,000 single cells from 13 adult mouse tissues[8], SnapATAC revealed almost twice as many additional cell clusters as originally reported (**Fig. 3c, Supplementary Fig. 7-10**). For example, when applied to prefrontal cortex, SnapATAC identified 22 different populations including 12 excitatory neurons representing layer-specific subtypes, 5 inhibitory neuronal clusters including Sst, Vip, Pvalb, Sncg and Lamp5 subtypes, 1 oligodendrocyte cluster, 1 oligodendrocyte precursor cluster, 1 astrocyte cluster, 1 microglia cluster, 1 endothelial cell cluster (**Fig. 3c, Supplementary Fig. 7-8**). This improved cell-type resolution also holds true for other tissue samples tested (**Supplementary Fig. 9-10**). These results indicate that SnapATAC outperforms existing methods on complex single cell accessibility datasets.

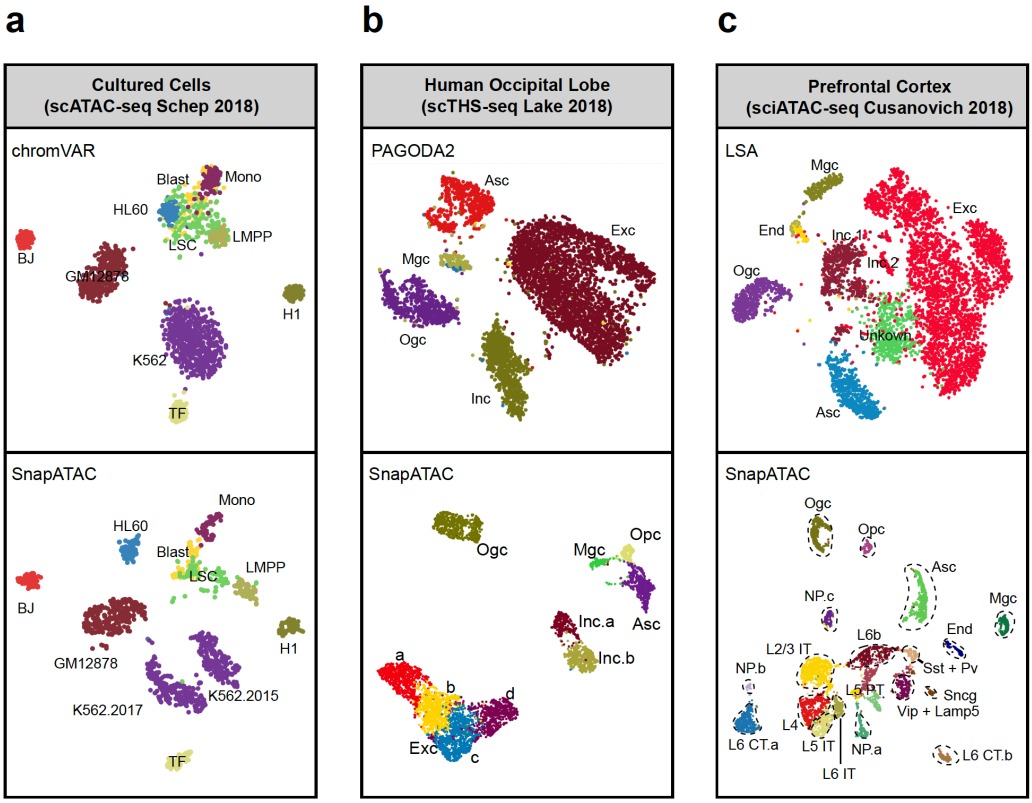


Figure 3. Validation of SnapATAC performance relative to alternative methods on published single cell ATAC-seq datasets. (**a**) ChromVAR (top) versus SnapATAC (bottom) on scATAC-seq data from 11 human cell lines. Points are colored by cell types. Blast: acute myeloid leukemia blast cells; LSC: acute 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. (**b**) POGODA2 (top) versus SnapATAC (bottom) on human Occipital Lobe scTHS-seq. Points are colored by identified cell types. (**c**) LSA (top) versus SnapATAC (bottom) on mouse prefrontal cortex sci-ATAC-seq dataset. Exc: excitatory neuron cells; Inc: inhibitory neuron cells; Opc: Oligodendrocyte precursor cells; Asc: astrocyte cells; Ogc: oligodendrocyte cells; Mgc: microglia cells. End: endothelial cells.

In addition to the clustering performance, SnapATAC also demonstrates high computational efficiency and scalability. Benchmarked using simulated scATAC-seq data sets from 1,000 to 100,000 cells, the CPU-time of SnapATAC scales linearly and at a significantly lower slope than other methods. This difference is especially pronounced relative to topic modeling methods such as Cis-Topic, a probabilistic method that requires extensive parameter optimization for large dataset. Using the same computing resource, when applied to 100,000 cells, SnapATAC is nearly 160 times faster than Cis-Topic, reducing the time from 30 hours to 10 minutes (**Table 1**; **Supplementary Fig. 11**).

Table 1 | Time taken for clustering using Cis-Topic and SnapATAC, as compared for different cell number N

N	*Cis-Topic*	*SnapATAC*
1,000	30min	6 sec
10,000	3h 16min	1min 20sec
100,000	31h 18min	11min 30sec

Tested on a machine with 5 AMD Opteron(TM) Processor 6276 CPUs

262 **A high-resolution *cis*-regulatory atlas of the secondary mouse motor cortex**
263 The mammalian brain is composed of myriad highly specialized cell types and
264 subtypes[17,24–27], which presents a unique challenge for single cell chromatin accessibility
265 analysis. As part of the BRAIN Initiative Cell Census Consortium[28], we have generated
266 single nucleus ATAC-seq profiles from >60,000 individual cells from the secondary
267 motor cortex (MOs) in the adult mouse brain (**Fig. 4a**). To our knowledge, this represents
268 the largest single cell chromatin accessibility dataset yet published from a single tissue
269 type. This dataset includes 2 biological replicates (**Fig. 4b; Supplementary Table 3**),
270 each generated from pooled tissue of at least 15 mice to prevent potential artifacts such as
271 dissection or batch effect. The aggregate signal showed high reproducibility between
272 biological replicates (R > 0.95**; Fig. 4b-c**, **Supplementary Fig. 12a-c**) and significant
273 enrichment for transcription start sites (TSS) indicating a high signal-to-noise ratio
274 (**Supplementary Fig. 12d-e**). After filtering out the poor-quality nuclei using stringent
275 criteria (**Supplementary Fig. 13-14**), we obtained a total of 64,795 nuclear profiles with
276 an average of ~5,000 sequencing fragments per nucleus (**Supplementary Table 4**).
277
278 SnapATAC identified and annotated the same 20 major clusters (**Fig. 4d**) from each
279 biological replicate (**Supplementary Fig. 15**), indicating the robustness of the method.
280 Based on gene body accessibility levels at canonical marker genes (**Fig. 4e**;
281 **Supplementary Fig. 16-17**), the 20 clusters were classified into eight excitatory
282 neuronal subpopulations (Snap25+, Slc17a7+, Gad1-; 50% of total nuclei), four inhibitory
283 neuronal subpopulations (Snap25+, Slc17a7-, Gad2+; 10% of total nuclei), one
284 oligodendrocyte subpopulation (Mog+; 9% of total nuclei), one oligodendrocyte
285 precursor subpopulation (Pdgfra+; 5% of total nuclei), one microglia subpopulation
286 (C1qb+; 7% of total nuclei), one astrocyte subpopulation (Apoe+; 13% of total nuclei), and
287 additional populations of endothelial, somatic, and somatic muscle cells accounting for
288 6% of total nuclei.
289
290 The accuracy of these cell-type classification is supported by several lines of evidence.
291 First, measurements of neuronal vs non-neuronal cell type abundance by Fluorescence-
292 activated cell sorting (FACS) from the same samples are highly consistent with estimates
293 from SnapATAC analysis (**Fig. 4f-g; Supplementary Fig. 28**). Second, the excitatory

294 neuron subpopulations we identify show specificity for known cortical layer-specific
295 marker genes and gradient transition between layers (**Fig. 4h**). Third, neuronal
296 classification for each of the major cell population based on snATAC-seq data was in
297 excellent agreement with previous annotations based on scRNA-seq[26] (**Fig. 5a**). All the
298 major neuronal subpopulations identified from snATAC-seq can be matched to the
299 scRNA-seq based classification of cell types in the mouse visual cortex. In addition, gene
300 body accessibility for marker genes in each cluster correlated well with expression levels
301 for corresponding genes and clusters (**Fig. 5b** and **Supplementary Fig. 18**). Taken
302 together, these data show that snATAC-seq can dissect the cellular heterogeneity of
303 mouse brain and classify cells in a way consistent with previous knowledge.
304
305 Notably, one rare Sst neuronal subtype previously identified from scRNA-seq (Sst-Chodl
306 in **Fig. 5b**) was not initially detected from snATAC-seq dataset. To examine whether
307 iterative analysis could help tease out this rare population, SnapATAC was applied to
308 1,577 Sst nuclei, finding 9 distinct sub-populations including the Sst subtype (Sst.9),
309 which accounts for less than 0.1% (52/64,795) of the total population profiled
310 (**Supplementary Fig. 19a-b**). Based on gene accessibility and analysis of enriched
311 transcription factor motifs (**Supplementary Fig. 19d**), Sst.9 most likely corresponds to
312 Nos1 type I neurons (also known as Sst-Chodl). Applying SnapATAC to each of the other
313 major neuronal cell types identified a total of 41 subtypes (**Supplementary Fig. 20**). To
314 our knowledge, this represents the highest resolution of scATAC-seq analysis of a
315 mammalian brain region. While the identity and function of these subtypes require
316 further experimental validation, our results demonstrate the exquisite sensitivity of
317 SnapATAC in resolving distinct neuronal subtypes with only subtle differences in
318 chromatin landscape and gene expression patterns.
319
320

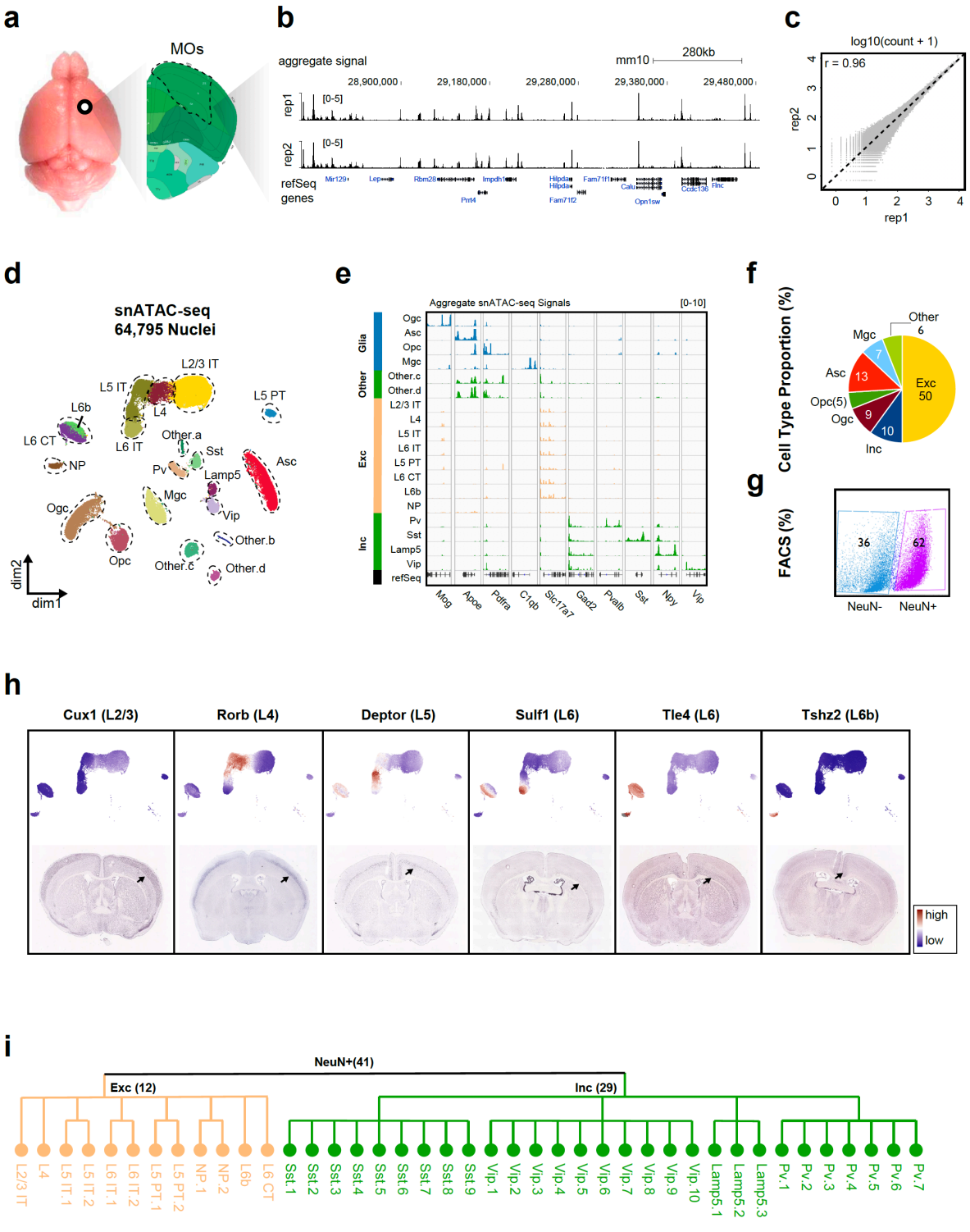
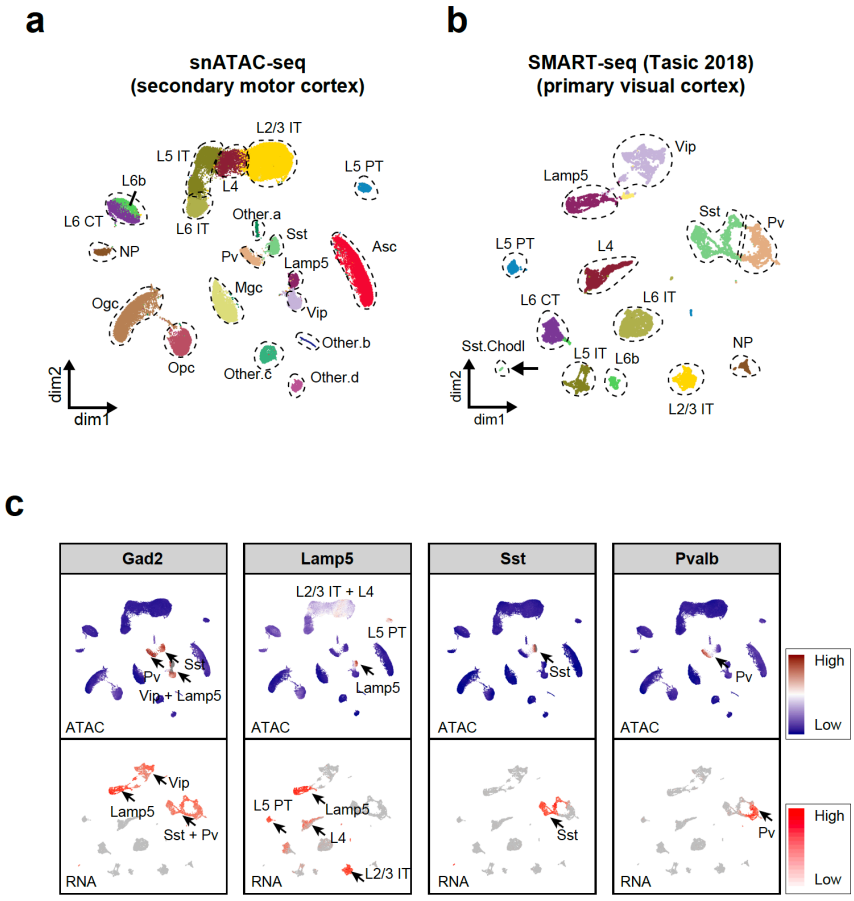


Figure 4. A high-resolution *cis*-regulatory atlas of the mouse motor cortex. (**a**) Illustration of secondary motor cortex (MOs) in the adult mouse brain. (**b**) Genome browser view of snATAC-seq aggregated signals for two biological replicates. (**c**) Reproducibility of aggregate signals for two biological replicates (rho=0.96, P < 1e-10).

(**d**) Two-dimensional visualization of SnapATAC clustering result. (**e**) Genome browser view of aggregate signal for each cluster at canonical marker genes. *Mog* is expressed in Oligodendrocyte cells; *Apoe* is expressed in Astrocyte cells; *Pdgfra* is expressed in Oligodendrocyte precursor cells; *C1qb* is expressed in Microglia cells; *Slc17a7* is expressed in excitatory cells; *Gad2* is expressed in inhibitory cells; *Pvalb* is strongly expressed in inhibitory Pvalb subtype; *Vip* is primarily expressed in inhibitory Vip subtype; *Sst* is expressed in inhibitory Sst subtype cells. *Npy* is expressed in inhibitory Lamp5 cells. (**f**) Cellular composition of cell types according to the SnapATAC clustering results. (**g**) Neuron versus non-neuron cell composition based on FACS sorting. (**h**) Imputed gene body accessibility level at marker genes for layer-specific excitatory neurons. (**i**) Dendrogram describing the taxonomy of neuronal subtypes.



340 **Figure 5. Comparison of neuron clusters with single cell RNA-seq.** (**a**) Two-
341 dimensional visualization of single-nucleus ATAC-seq clusters of mouse secondary motor
342 cortex. (**b**) Two-dimensional visualization of single neuron clusters from mouse frontal
343 visual cortex using SMART-seq V4. (**c**) Comparison of imputed gene body accessibility
344 and gene expression level at canonical marker genes. *Gad2* is highly expressed in
345 inhibitory neurons; *Lamp5* is expressed in inhibitory Lamp5 subtype, excitatory L2/3 IT,
346 L4 and L5 PT neurons; *Sst* is expressed in inhibitory Sst subtype; *Pvalb* is expressed in
347 inhibitory Pvalb subtype.
348
349 **SnapATAC uncovers candidate *cis*-elements active in rare cell populations**
350 A key utility of single cell chromatin accessibility analysis is to identify regulatory
351 sequences in the genome. By pooling reads from nuclei in each cluster, cell-type specific
352 chromatin landscapes can be obtained (**Fig. 6a**). Focusing on the major cell types
353 described in Figure 4d, peaks of chromatin accessibility signals were determined in each
354 cell cluster containing at least 500 cells, resulting in a combined total of 316,257 unique
355 candidate *cis*-elements (**Supplementary Table 5**). Most notably, 56%
356 (190,818/337,932) of these open chromatin regions are not detected in the analysis of
357 bulk ATAC-seq data of the same brain region (**Methods; Fig. 6b; Supplementary**
358 **Table 6**). We hypothesized that these open chromatin regions not detected in the bulk
359 ATAC-seq analysis may represent those that are only accessible in minor cell populations.
360 Supporting this hypothesis, nearly 80% of these elements were detected from only one
361 cell cluster (**Fig. 6c**).
362
363 Several lines of evidence support that these additional open chromatin regions are
364 functional elements, rather than technical noises. First, these sequences showed
365 significantly higher conservation than randomly selected genomic sequences with
366 comparable mappability scores (**Fig. 6d**). Second, these open chromatin regions display
367 enrichment for transcription factor binding motifs corresponding to transcription factors
368 (TFs) that play important regulatory roles in the corresponding cell types
369 (**Supplementary Table 7**). For example, the binding motif for Mef2c is highly enriched
370 in novel candidate cis-elements identified from Pvalb neuronal subtype (P-value = 1e-363;
371 **Fig. 7e-f**), consistent with previous report that Mef2c is upregulated in embryonic

372 precursors of Pv interneurons[29]. Similarly, the binding motif for ETS-factor PU.1, a known
373 transcription regulator of microglia[30], was highly enriched in the novel elements detected
374 from microglia (P-value = 1e-2250) (**Supplementary Table 7**). Finally, the new open
375 chromatin regions tend to test positive in transgenic reporter assays. Comparison to the
376 VISTA enhancer database[31] shows that enhancer activities of 256 of the newly identified
377 open chromatin regions have been previously tested using transgenic reporter assays in
378 e11.5 mouse embryos (**Supplementary Table 8**). 65% (167/256) of them drive
379 reproducible reporter expression in at least one embryonic tissue, substantially higher
380 than background rates (9.7%) estimated from regions in the VISTA database that lack
381 canonical enhancer mark (manuscript under review)[32]. Here, we displayed four examples
382 where elements were only present in rare population are tested positive in the brain
383 function associating regions (**Fig. 6g**). It is important to note that this comparison only
384 considers the *in vivo* enhancer function but does not validate the exact tissue-specificity
385 because the current data cannot exclude the possibility of a regulatory element being an
386 enhancer in other tissues.

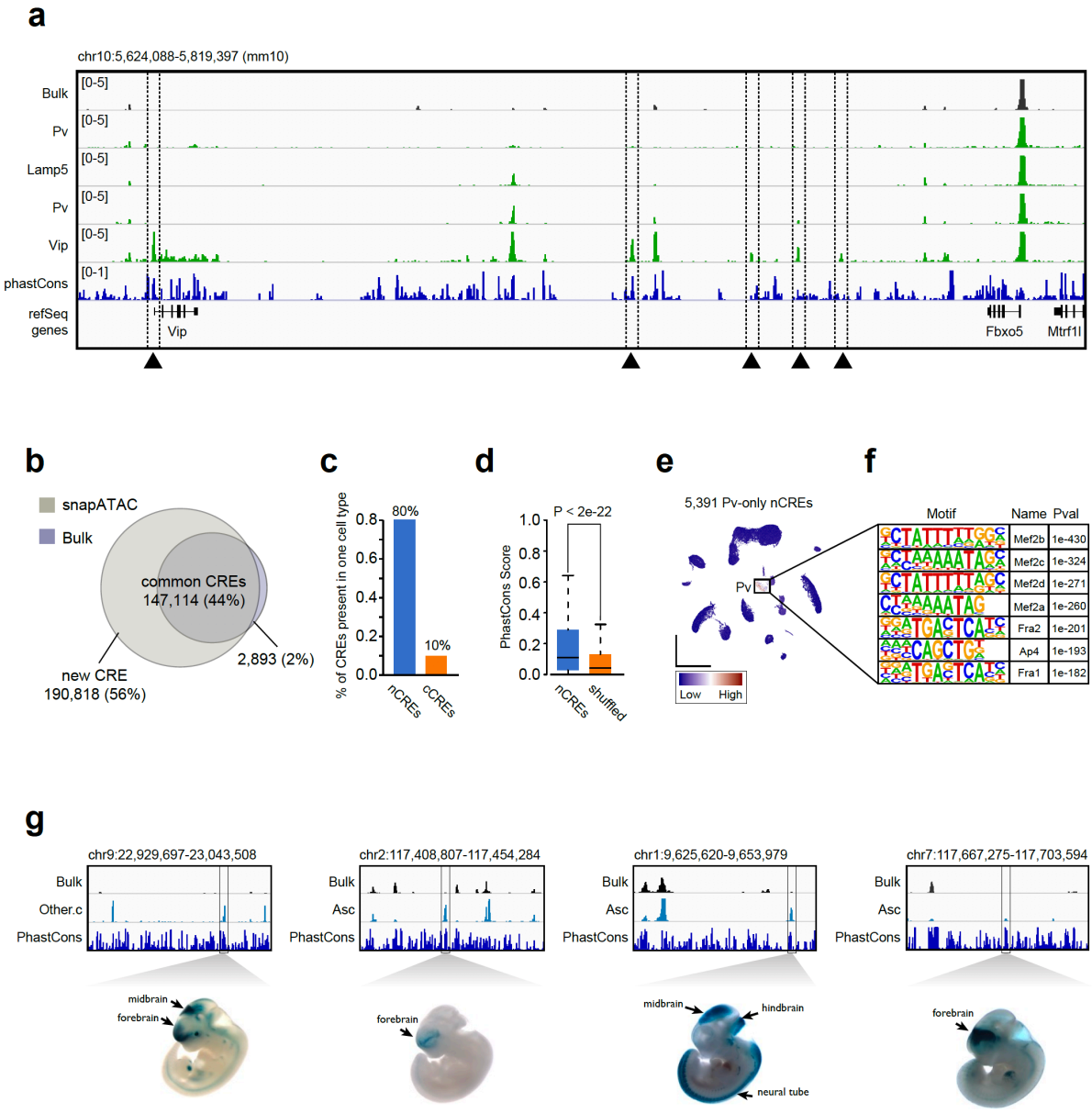


Figure 6. SnapATAC uncovers novel candidate *cis*-regulatory elements in rare cell types. (**a**) Genome browser view of 20Mb region flanking gene *Vip*. Dash line highlighting five regulatory elements specific to Vip subtypes that are under-represented in the conventional bulk ATAC-seq signal. (**b**) Over fifty percent of the regulatory elements identified from 20 major cell populations are new compared to that of bulk ATAC-seq. (**c**) Eighty percent of new elements present in only one cell type. (**d**) Sequence conservation comparison between new elements and randomly chosen genomic regions. (**e**) 5,391 Pv-specific new elements are highly enriched in Pv subtypes. (**f**) Top seven

396 motifs enriched in Pv-specific new elements. (**g**) Four new elements are tested positive
397 using transgenic mouse assay according to VISTA database.
398
399
400 **Extension of SnapATAC to other single cell epigenomics datasets**
401 Although SnapATAC was designed to analyze single cell ATAC-seq data, it is also
402 applicable to other sparse single cell datasets. To demonstrate this, SnapATAC was used
403 to reanalyze a variety of published single cell epigenomics and transcriptomics datasets.
404 When applied to a set of 14,963 sparse single nucleus RNA-seq (sNuc-seq) datasets from
405 adult post-mortem human brain tissue[33], SnapATAC uncovered distinct clusters
406 corresponding to all known major cell types (**Methods**; **Fig. 7a; Supplementary Fig.**
407 **21-22**). Sub-clustering of GABAergic neurons further identified 11 subtypes with
408 distinct gene expression patterns (**Supplementary Fig. 23**), including two Sst, three
409 Pv, three Vip subtypes and three clusters enriched for Lamp5 gene expression. Similarly,
410 analyzing a dataset of 2,784 methylomes form single neuronal nuclei in the human frontal
411 cortex[27], SnapATAC identified all the major and subtypes in excellent agreement with the
412 previous classification (**Fig. 7b; Supplementary Fig. 24**). When applied to single cell
413 H3k4me2 ChIP-seq data[34], SnapATAC correctly distinguished mouse embryonic stem
414 cells (ESCs) from mouse embryonic fibroblasts (MEFs) cells (**Methods; Fig. 7c**). Finally,
415 when used to analyze multiplexing single cell Hi-C[35], SnapATAC separates HeLa S3 from
416 HAP1 cells (**Methods; Fig. 7d**). Taken together, these results show that SnapATAC can
417 be used to process other single cell epigenomics datasets.
418
419

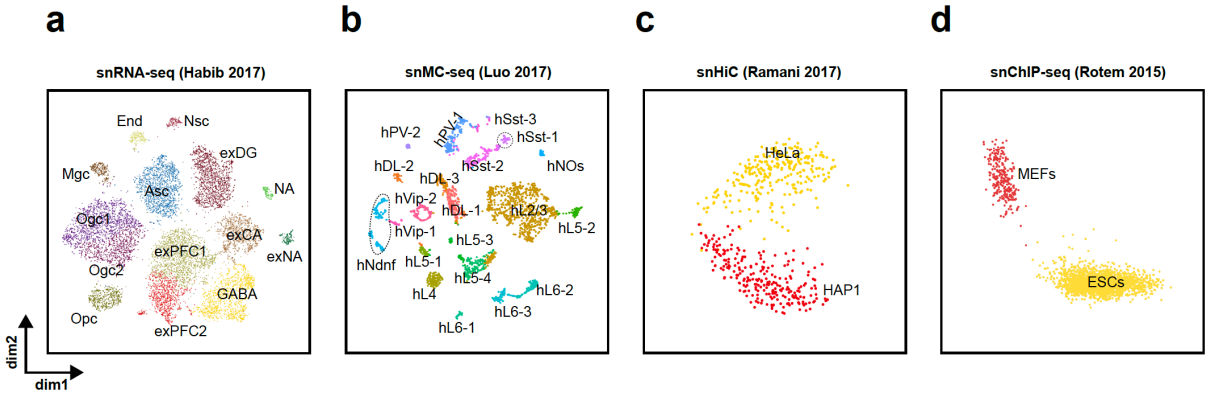


Figure 7. Applying SnapATAC to other single cell modalities. (**a**) A two-dimensional visualization plot of 14,963 DroNc-seq single-nucleus RNA-seq from adult frozen human hippocampus and prefrontal cortex. Nuclei are color-coded by cluster membership. exPFC: glutamatergic neurons from the PFC; GABA: GABAergic interneurons; exCA1/3: pyramidal neurons from the hip CA region; exDG: granule neurons from the hip dentate gyrus region; ASC: astrocytes; MGC: microglia; OGC, oligodendrocytes; OPC, oligodendrocyte precursor cells; NSC: neuronal stem cells; SMC: smooth muscle cells; END: endothelial cells. (**b**) A two-dimensional visualization plot of 2,784 methylomes form single neuronal nuclei in the human frontal cortex. (**c**) A two-dimensional visualization of single-cell Hi-C from HeLa S3 and HAP1 cells. (**d**) A two-dimensional visualization plot of H3K4me2 single cell ChIP-seq from ESCs and MEFs.

Conclusion & Discussion

In summary, SnapATAC is a comprehensive bioinformatic solution for single cell ATAC-seq analysis. The open-source software runs on commonly available and inexpensive hardware, making it accessible to any researcher using single-cell ATAC-seq data. Through extensive benchmarking using a variety of single cell chromatin datasets, SnapATAC outperforms existing methods substantially in both clustering accuracy and scalability. Although designed for analyzing single cell ATAC-seq datasets, SnapATAC can also be used to analyze a broad range of sparse single cell epigenomics data, such as single cell Hi-C, single cell ChIP-seq and single cell methylomes.

444 Applying SnapATAC to a new in-house dataset including >60,000 high quality single cell
445 ATAC-seq profiles from mouse secondary motor cortex, led to a single cell atlas of
446 candidate cis regulatory elements for this mouse brain region. The cellular diversity
447 identified by chromatin accessibility is at an unprecedented resolution and is consistent
448 with mouse neurogenesis and taxonomy revealed by single cell transcriptome data.
449 Besides characterizing cell types, SnapATAC identifies candidate cis-regulatory
450 sequences in each of the major cell types and infers transcription factors that control cell-
451 type specific gene expression programs. Although this study primarily focused on the
452 major cell types, additional neuronal subtypes can be identified through sub-clustering.
453 To obtain a robust signal for each of the subtypes, would require substantially more cells
454 and, more importantly, further anatomical, physiological, and functional experimental
455 validation.
456
457 One of the most exciting features of SnapATAC is its ability to identify candidate *cis*
458 regulatory elements active only in rare cell population. A large fraction (56%) of the
459 candidate *cis*-elements identified using SnapATAC analysis of the mouse secondary
460 motor cortex are not detected in bulk analysis. Most of these elements appear to be active
461 in individual cell types that account for 1% or less of the total cell population. While
462 further experiments to thoroughly validate the function of these additional open
463 chromatin regions is still needed, the ability for SnapATAC to uncover *cis*-elements from
464 minor cell types of a complex tissue will certainly greatly expand the catalog of cis
465 regulatory sequences in the genome.
466
467 **Data availability**
468 Raw and processed data to support the findings of this study have been deposited to
469 NCBI Gene Expression Omnibus with the accession number GSE126724.
470
471 **Code availability**
472 The scripts and pipeline for the analysis can be found at
473 https://github.com/r3fang/SnapATAC.

Acknowledgements

We thank D. Gorkin, R. Raviram, J. Hocker and X. Zhou for proofreading and suggestions for the manuscript. We thank S. Kuan for sequencing support. We thank O. Poirion, C. Zhang and B. Li for Bioinformatics support. We thank F. Xie for sharing processed single nucleus methylome data. 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.; 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.

488 **Methods**

489 **1. SnapATAC Pipeline**

490 *1.1 Barcode Demultiplexing*

491 Using a custom python script, we first de-multicomplexed FASTQ files by integrating the
492 cell barcode into the read name in the following format: "@" + "barcode" + ":" +
493 "original_read_name".

494

495 *1.2 Alignment & Sorting*

496 De-multicomplexed reads were aligned to the corresponding reference genome (i.e.
497 mm10 or hg19) using bwa[36] (0.7.13-r1126) in pair-end mode with default parameter
498 settings. Alignments were then sorted based on the read name using samtools[37] (v1.9).

499

500 *1.3 Fragmentation & Filtration*

501 Pair-end reads were converted into fragments and only those that are 1) properly paired
502 (according to SAM flag value); 2) uniquely mapped (MAPQ > 30); 3) with length less than
503 1000bp were kept.

504

505 *1.4 Duplicates Removal*

506 Sorted by barcode, fragments belonging to the same cell (or barcode) were automatically
507 grouped together which allowed for removing PCR duplicates for each cell separately.

508

509 *1.5 Snap File Generation*

510 Next, using filtered and sorted bam file, we generated a snap-format (Single-Nucleus
511 Accessibility Profiles) file which is hierarchically structured hdf5 file that contains the
512 following sessions: header (HD), cell-by-bin accessibility matrix (AM), cell-by-peak
513 matrix (PM), cell-by-gene matrix (GM), barcode (BD) and fragment (FM). HD session
514 contains snap-file version, date, alignment and reference genome information. BD
515 session contains all unique barcodes and corresponding meta data. AM session contains
516 multiple cell-by-bin matrices of different resolutions (or bin sizes). PM session contains
517 cell-by-peak count matrix. GM session contains cell-by-gene count matrix. FM session
518 contains all usable fragments for each cell. Fragments are indexed for fast search which

allows for generation of cell-type specific chromatin landscapes after clustering. Detailed information about snap file can be found in **Supplementary Note 1**.

1.6 Cell-by-Bin Count Matrix Generation

Using snap file, we next created cell-by-bin count matrices of different resolutions. The genome was segmented into uniform-sized bins and single cell ATAC-seq profiles were represented as cell-by-bin matrix with each element indicating number of open chromatin fragments overlapping with a given bin in a certain cell. A snap file allows for storing multiple cell-by-bin count matrices with different resolutions. For MOs snATAC-seq, we created snap file with 1kb, 5kb and 10kb resolution.

1.7 Barcode Selection

We next identified the high-quality barcodes based on the following criteria. 1) Total Sequencing Fragments (>1,000); 2) Mapping Ratio (>0.8); 3) Properly Paired Ratio (>0.9); 4) Duplicate Ratio (<0.5); 5) Mitochondrial Ratio (<0.1). We abandoned the use of reads in peak ratio as a metric for cell selection for two reasons. First, we found the reads-in-peak ratio is highly cell type specific. For instance, according to published single cell ATAC-seq, human fibroblast (BJ) cells have significantly higher reads in peak ratio (40-60%) versus 20-40% for GM12878 cells. Similarly, we found Glia cells overall have very different reads in peak ratio distribution compared to neuronal cells. We suspect this may reflect the nucleus size or global chromatin accessibility. Second, population-defined set of accessibility peaks are incomplete and are biased to the dominant populations. As shown in this study, for a complex tissue such as mammalian brain, we found over 50% of the peaks present in the rare populations are not identified from the aggregate signal of snATAC-seq. Therefore, we abandoned the use of reads in peak ratio for cell selection.

1.8 Bin Size Selection

Using the remaining cells, we sought to determine the optimal bin size based on the correlation between replicates. We recommend choosing the smallest bin size (or highest resolution) whose Pearson correlation between replicates is greater than 0.95. If there are no biological replicates available, we recommend splitting the cells into pseudo-replicates. In this study, we use 5kb unless noted.

551

552 *1.9 Matrix Binarization*

553 After choosing the optimal bin size, we found the vast majority of the items in the cell-by-
554 bin count matrix is "0", indicating either inaccessible (closed chromatin) or missing data.
555 Among the non-zero elements, some items have abnormally high coverage (often > 200)
556 perhaps due to alignment error. Therefore, we first removed the top 0.1% items of the
557 highest coverage in the matrix before converting it into a binary matrix.

558

559 *1.10 Feature Selection*

560 We next filtered any bins overlapping with the ENCODE blacklist
561 (http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/) to prevent from any
562 potential artifacts. Bins of exceedingly high coverage which likely represent the genomic
563 regions that are invariable between cells such as housekeeping gene promoters were
564 removed. We noticed that filtering bins of extremely low coverage perhaps due to random
565 noise can also improve the robustness of the downstream clustering analysis. In detail,
566 we calculated the coverage of each bin using the binary matrix and normalized the
567 coverage by log10(count + 1). We found the log-scaled coverage obey approximately a
568 gaussian distribution (**Supplementary Fig. 25**) which is then converted into zscore.
569 Bins with zscore beyond ± 2 were filtered before further analysis.

570

571 *1.11 Jaccard Index Matrix*

572 Next, we converted the genome-wide cell-by-bin matrix into a cell-by-cell similarity
573 matrix by calculating the Jaccard index between every two cells in the basis of genome-
574 wide profile overlaps. Usually, the number of cells is far smaller than number of bins,
575 therefore, it immediately reduces the dimensionality and increase the scalability of the
576 pipeline. However, the time for computing Jaccard matrix increases exponentially with
577 cell number growth. To solve the problem of big data, 1) we first divided the cells into
578 groups and calculated a sub Jaccard index matrix separately in parallel. For instance,
579 given that there are 50,000 cells in total, we first split the cells into 10 chunks with each
580 chunk containing 5,000 cells. Then we calculated the pairwise sub jaccard index matrix
581 between every two chunks. Finally, we created the entire Jaccard index matrix by

582 combining all sub Jaccard matrices. This allows for in-parallel computing. 2) To further
583 speed up this process, instead of calculating a full Jaccard matrix by comparing every two
584 cells, we calculated a partial Jaccard matrix by estimating the similarity between N cells
585 with a subset of randomly chosen K cells (K << N) (k=2000 used in this study unless
586 noted). We found that, without sacrificing the performance (**supplementary Fig. 26**),
587 this can substantially improve the scalability of the pipeline, making it possible for
588 processing millions of cells in the future.
589
590 *6.12 Normalization*
591 Theoretically, the entries of the Jaccard matrix M_{ij}, would reflect the true similarity
592 between cell i and j. However, due to the differing coverage between cells, this becomes
593 not the case. If there is a high sequencing depth of cell i, then M_{ij} will tend to have higher
594 Jaccard index, regardless whether i and j is actually similar or not (**Supplementary Fig.**
595 **1-2**).
596
597 This can also be proved as below. Given 2 cells i and j and let X_i and X_j be the binary vector.
598 The coverage of i and j is $C_i = sum(X_i)$ and $C_j = sum(X_j)$ (C_i >0, C_j >0), then let $P_i =$
599 $C_i/|X_i|$ and $P_j = C_j/|X_j|$ where $|X_i|$ and $|X_j|$ is the number of bins. Then the expected
600 Jaccard index between cell i and j is:
601
602 $$E_{ij} = (P_i * P_j)/(P_i + P_j - P_i P_j)$$
603
604 Because $P_i * P_j > 0$, then
605
606 $$E_{ij} = 1/(1/P_i + 1/P_j - 1).$$
607
608 Now it is obvious to see that the increase of either P_i or P_j will result in an increase of E_{ij}.
609
610 Here, we propose three different approaches to normalize Jaccard matrix, namely
611 observed over expected (OVE), observed over neighbor (OVN) and iterative matrix
612 balancing (ICE).

613

614 1.12.1 OVE: we first estimated the expected Jaccard index E_{ij} as described above, assuming
615 cells have random profiles. We noticed that E_{ij} usually underestimates similarity for high-
616 coverage cells, to adjust for this, we performed linear regression between expected E and
617 observed M and used residuals as normalized matrix N. Residuals matrix N was then
618 standardized for each cell.

619

620 1.12.2 OVN: the second approach estimated the expected cell-by-cell similarity using
621 neighboring cells. In detail, for every pair of cells i and j, according to the coverage, we
622 selected two groups of cells G_i^k and G_j^k representing the k nearest neighboring cells for i
623 and j with closest coverage. After removing common cells shared by G_i^k and G_j^k, we next
624 calculated the Jaccard matrix Jaccad(G_i^k, G_j^k) between these two groups of cells, the
625 average value of which was used as expected value to correct the bias in M_{ij}.

626

627 1.12.3 ICE: we also borrowed the idea of matrix balancing which is a technique commonly
628 used in Hi-C matrix normalization. We adapted "normICE" function in HiTC R package
629 which normalizes Hi-C matrix using matrix balancing algorithm that consists of
630 iteratively estimating the matrix bias using the l1 norm.

631

632 To compare the performance of different normalization methods, we performed principle
633 component analysis (PCA) against the normalized matrix and examined the degree of
634 association between the first principle component and sequencing depth. Overall, a
635 higher correlation indicates the dominate variance between cells is the read depth rather
636 than the meaningful biological variance. For all the data sets we have tested, OVN and
637 OVE overall shows a comparable performance (**Supplementary Fig. 2**), however, as
638 OVE is substantially faster at least according to our implementation, therefore, we choose
639 it as our final normalization method as used in this study. All the analysis is using OVE
640 unless noted. But all three methods are implemented in SnapATAC package.

641

642 To further demonstrate the performance of the normalization, we applied it to previously
643 published human scATAC-seq data from 10 cell lines[13]. The effect of normalization is

644 clearly evident from inspecting the heatmap. Cell types that are difficult to distinguished
645 in the original matrix become visibly distinct in the normalized matrix (**Supplementary**
646 **Fig. 2a-b**). Further applying linear dimensionality reduction against both matrices, we
647 found the first principal component of the raw matrix is strongly correlated with the
648 coverage (rho=-0.90, P < 1e-10; **Supplementary Fig. 2c**), whereas the first dimension
649 of the normalized matrix successfully distinguished BJ, TF from other cell types (rho=-
650 0.04; **Supplementary Fig. 2d**).
651
652 We next tested it against other published datasets. When applied to human Occipital Lobe
653 scTHS-seq[9], the first principal component of normalized matrix separates neuronal from
654 non-neuronal cells (**Supplementary Fig. 2e**). Similarly, when applied to the drosophila
655 embryo sci-ATAC-seq data[6], the first dimension now distinguished 4 major cell clades
656 (**Supplementary Fig. 2f**). Together, all suggest that SnapATAC is able to adjust for the
657 coverage bias.
658
659 *6.13 Dimensionality Reduction*
660 Like any other type of single-cell analysis, scATAC-seq contains extensive technical noise.
661 To overcome this challenge, we performed Principle Component Analysis (PCA) to
662 combine information across a correlated feature set hereby creating a mega-feature and
663 exclude the variance potential resulting from technical noise.
664
665 *1.14 Determining Significant Principle Components*
666 It is both critical and challenging to decide how many principle components (PCs) to
667 include for the downstream analysis. A variety of methods have been developed to identify
668 optimal number of PCs. For instance, JackStraw[38] can specify significant components for
669 PCA through permutation-based statistical test, however, this gets extensively time-
670 consuming when cell number is large. Instead, we recommend using an *ad hoc* approach
671 for choosing the optimal number of components. One approach as proposed by Sauret[39]
672 to simplify look at the variance plot and find the "elbow" point. The other heuristic
673 approach, we found also useful, is to plot every two pairs of PCs and simply look at the
674 plot and choose number of PCs that stop separating cells.
675

676 *1.15 Clustering.*

677 Using the selected significant PCs, we next calculated pairwise Euclidean distance
678 between every two cells, using this distance, we created a k-nearest neighbor graph in
679 which every cell is represented as a node and edges are drawn between cells within k
680 nearest neighbors. Edge weight between any two cells are refined by shared overlap in
681 their local neighborhoods using Jaccard similarity. Finally, we applied community finding
682 algorithm Louvain to identify the 'communities' in the resulting graph which represents
683 groups of cells sharing similar profiles, potentially originating from the same cell type.
684 This method is also known as 'Louvain-Jaccard'[40].

685

686 *1.16 Visualization.*

687 We next project the high-dimension data into a 2D space using BH t-SNE[41] implemented
688 by Rtnse package or FI-tsne[42] or UMAP[43] to visualize and explore the data. All three
689 methods are integrated into SnapATAC package.

690

691 *1.17 Cluster Annotation.*

692 To annotate the identified clusters, we next calculated the gene-body accessibility level
693 for every cell and annotated the cluster based on the marker genes identified from
694 previous single cell RNA sequencing. Note that clustering is unsupervised while
695 annotation is a supervised procedure that requires expert knowledge. Recently, a method
696 called Garnett[44] is developed to automatically annotate ATAC-seq clusters using single
697 cell RNA-seq. To further enhance the structure of data and remove the noise, we adopted
698 MAGIC[45] to smooth the gene accessibility signal.

699

700 *1.18 Identification of Cis-Elements*.

701 Cells belonging to the same cluster are pooled to create ensemble signal for each of the
702 cell type. This allows for identifying *Cis*-elements *de novo* from each of the clusters.
703 MACS2[46] (version 2.1.2) was used for generating signal tracks and peak calling with the
704 following parameters: --nomodel --qval 1e-2 -B --SPMR --call-summits.

705

706 *1.19 Cell-by-Peak Accessible Matrix.*

Merging peaks identified from each cluster, we create a reference list of regulatory elements. Using this reference map, we next create a cell-by-peak count matrix.

1.20 Motif Analysis.

We performed motif discovery using homer to infer the potential master regulator that control for gene expression in each of the cell types.

1.21 GREAT Analysis.

We next performed Genomic Region Enrichment Analysis (GREAT) to predict the function of each cluster.

2. Analysis of simulated single cell ATAC-seq

First, we downloaded the alignment files (bam files) for ten bulk ATAC-seq experiment from ENCODE (**Supplementary Table 1**). From each bam file, we simulated 200 single cell ATAC-seq datasets by randomly down sampling to a variety of coverages ranging from 1,000 to 10,000 reads per cells. Using simulated single cell ATAC-seq datasets, we created a cell-by-bin matrix with 5kb bin size for SnapATAC clustering. Merging peaks downloaded from ENCODE for each experiment, we created a cell-by-peak matrix for LSA, Cis-Topic, Cicero and chromVAR clustering. Code used in this study can be found in **Supplementary Note 2**.

3. Analysis of published single cell ATAC-seq datasets

3.1 Analysis of scATAC-seq datasets from human cell lines.

We obtained scATAC-seq count matrix from GEO (GSE99172). Analysis code used in this study is available in **Supplementary Note 3**.

3.2 Analysis of scTHS-seq datasets from human Occipital Lobe.

The cell-by-peak matrix was generated and shared by Aerts Lab (http://scenic.aertslab.org/cisTopic/counts_Lake.Rds). Analysis code used in this study is available in **Supplementary Note 4**.

3.3 Analysis of sci-ATAC-seq datasets from mouse atlas.

739 We downloaded processed data for each tissue from GEO (GSE111586) and generated the
740 snap file with cell-by-bin matrix at 5kb bin resolution. Analysis code used in this study is
741 available in **Supplementary Note 5**.
742
743 **4. Tissue collection & Nuclei Perspiration**
744 Adult C57BL/6J male mice were purchased from Jackson Laboratories. Brains were
745 extracted from P56-63 old mice and immediately sectioned into 0.6 mm coronal
746 sections, starting at the frontal pole, in ice-cold dissection media. The secondary motor
747 cortex (MOs) region was dissected from the first three slices along the anterior-posterior
748 axis according to the Allen Brain reference Atlas (http://mouse.brain-map.org/, see
749 Supplementary S27 for depiction of posterior view of each coronal slice; dashed line
750 highlights the MOs regions on each slice). Slices were kept in ice-cold dissection media
751 during dissection and immediately frozen in dry ice for posterior pooling and nuclei
752 production. For nuclei isolation, the MOs dissected regions from 15-23 animals were
753 pooled, and two biological replicas were processed for each slice. Nuclei were isolated as
754 described in previous studies[27,47], except no sucrose gradient purification was
755 performed. Flow cytometry analysis of brain nuclei was performed as described in Luo
756 et al[27].
757
758 **5. Tn5 transposase purification & loading**
759 Tn5 transposase was expressed as an intein chitin-binding domain fusion and purified
760 using an improved version of the method first described by Picelli et al[48]. T7 Express
761 lysY/I (C3013I, NEB) cells were transformed with the plasmid pTXB1-ecTn5 E54K L372P
762 (#60240, Addgene)[48]. An LB Ampicillin culture was inoculated with three colonies and
763 grown overnight at 37°C. The starter culture was diluted to an OD of 0.02 with fresh
764 media and shaken at 37°C until it reached an OD of 0.9. The culture was then immediately
765 chilled on ice to 10°C and expression was induced by adding 250 μM IPTG (Dioxane Free,
766 CI8280-13, Denville Scientific). The culture was shaken for 4 hours at 23°C after which
767 cells were harvested in 2 L batches by centrifugation, flash frozen in liquid nitrogen and
768 stored at -80°C. Cell pellets were resuspended in 20 ml of ice cold lysis buffer (20 mM
769 HEPES 7.2-KOH, 0.8 M NaCl, 1 mM EDTA, 10% Glycerol, 0.2% Triton X-100) with
770 protease inhibitors (cOmplete, EDTA-free Protease Inhibitor Cocktail Tablets,

771 11873580001, Roche Diagnostics) and passed three times through a Microfluidizer (lining
772 covered with ice water, Model 110L, Microfluidics) with a 5 minute cool down interval in
773 between each pass. Any remaining sample was purged from the Microfluidizer with an
774 additional 25 ml of ice-cold lysis buffer with protease inhibitors (total lysate volume
775 ~50ml). Samples were spun down for 20 min in an ultracentrifuge at 40K rpm (L-80XP,
776 45 Ti Rotor, Beckman Coulter) at 4°C. ~45 ml of supernatant was combined with 115 ml
777 ice cold lysis buffer with protease inhibitors in a cold beaker (total volume = 160 ml) and
778 stirred at 4°C. 4.2ml of 10% neutralized polyethyleneimine-HCl (pH 7.0) was then added
779 dropwise. Samples were spun down again for 20 min in an ultracentrifuge at 40K rpm (L-
780 80XP, 45 Ti Rotor, Beckman Coulter) at 4°C. The pooled supernatant was loaded onto
781 ~10ml of fresh Chitin resin (S6651L, NEB) in a chromatography column (Econo-Column
782 (1.5 × 15 cm), Flow Adapter: 7380015, Bio-Rad). The column was then washed with 50-
783 100 ml lysis buffer. Cleavage of the fusion protein was initiated by flowing ~20ml of
784 freshly made elution buffer (20 mM HEPES 7.2-KOH, 0.5 M NaCl, 1 mM EDTA, 10%
785 glycerol, 0.02% Triton X-100, 100mM DTT) onto the column at a speed of 0.8ml/min for
786 25 min. After the column was incubated for 63 hrs at 4°C, the protein was recovered from
787 the initial elution volume and a subsequent 30 ml wash with elution buffer. Protein-
788 containing fractions were pooled and diluted 1:1 with buffer [20 mM HEPES 7.2-KOH,1
789 mM EDTA, 10% glycerol, 0.5mM TCEP) to reduce the NaCl concentration to 250mM. For
790 cation exchange, the sample was loaded onto a 1ml column HiTrap S HP (17115101, GE),
791 washed with Buffer A (10mM Tris 7.5, 280 mM NaCl, 10% glycerol, 0.5mM TCEP) and
792 then eluted using a gradient formed using Buffer A and Buffer B (10mM Tris 7.5, 1M NaCl,
793 10% glycerol, 0.5mM TCEP) (0% Buffer B over 5 column volumes, 0-100% Buffer B over
794 50 column volumes, 100% Buffer B over 10 column volumes). Next, the protein-
795 containing fractions were combined, concentrated via ultrafiltration to ~1.5 mg/mL and
796 further purified via gel filtration (HiLoad 16/600 Superdex 75 pg column (28989333,
797 GE)) in Buffer GF (100mM HEPES-KOH at pH 7.2, 0.5 M NaCl, 0.2 mM EDTA, 2mM
798 DTT, 20% glycerol). The purest Tn5 transposase-containing fractions were pooled and 1
799 volume 100% glycerol was added to the preparation. Tn5 transposase was stored at -20°C.
800
801 To generate Tn5 transposomes for combinatorial barcoding assisted single nuclei
802 ATAC-seq, barcoded oligos were first annealed to pMENTs oligos (95 °C for 5 min,

803 cooled to 14 °C at a cooling rate of 0.1 °C/s) separately. Next, 1 µl barcoded transposon
804 (50 µM) was mixed with 7 ul Tn5 (~7 µM). The mixture was incubated on the lab bench
805 at room temperature for 30 min. Finally, T5 and T7 transposomes were mixed in a 1:1
806 ratio and diluted 1:10 with dilution buffer (50 % Glycerol, 50 mM Tris-HCl (pH=7.5),
807 100 mM NaCl, 0.1 mM EDTA, 0.1 % Triton X-100, 1 mM DTT). For combinatorial
808 barcoding, we used eight different T5 transposomes and 12 distinct T7 transposomes,
809 which eventually resulted in 96 Tn5 barcode combinations per sample[7]
810 (**Supplementary Table 9**).
811

812 **6. Single-nucleus ATAC-seq data generation**

813 Combinatorial ATAC-seq was performed as described previously with modifications[5,7].
814 For each sample two biological replicates were processed. Nuclei were pelleted with a
815 swinging bucket centrifuge (500 x g, 5 min, 4°C; 5920R, Eppendorf). Nuclei pellets were
816 resuspended in 1 ml nuclei permeabilization buffer (5 % BSA, 0.2 % IGEPAL-CA630, 1mM
817 DTT and cOmpleteTM, EDTA-free protease inhibitor cocktail (Roche) in PBS) and
818 pelleted again (500 x g, 5 min, 4°C; 5920R, Eppendorf). Nuclei were resuspended in
819 500 µL high salt tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM
820 potassium-acetate, 11 mM Mg-acetate, 17.6% DMF) and counted using a hemocytometer.
821 Concentration was adjusted to 4500 nuclei/9 µl, and 4,500 nuclei were dispensed into
822 each well of a 96-well plate. Glycerol was added to the leftover nuclei suspension for a
823 final concentration of 25 % and nuclei were stored at -80°C. For tagmentation, 1 µL
824 barcoded Tn5 transposomes[7,48] (**Supplementary Table 9**) were added using a
825 BenchSmart™ 96 (Mettler Toledo), mixed five times and incubated for 60 min at 37 °C
826 with shaking (500 rpm). To inhibit the Tn5 reaction, 10 µL of 40 mM EDTA were added
827 to each well with a BenchSmart™ 96 (Mettler Toledo) and the plate was incubated at 37
828 °C for 15 min with shaking (500 rpm). Next, 20 µL 2 x sort buffer (2 % BSA, 2 mM EDTA
829 in PBS) were added using a BenchSmart™ 96 (Mettler Toledo). All wells were combined
830 into a FACS tube and stained with 3 µM Draq7 (Cell Signaling). Using a SH800 (Sony),
831 20 nuclei were sorted per well into eight 96-well plates (total of 768 wells) containing
832 10.5 µL EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma)[7]. Preparation of
833 sort plates and all downstream pipetting steps were performed on a Biomek i7 Automated
834 Workstation (Beckman Coulter). After addition of 1 µL 0.2% SDS, samples were

835 incubated at 55 °C for 7 min with shaking (500 rpm). We added 1 µL 12.5% Triton-X to
836 each well to quench the SDS and 12.5 µL NEBNext High-Fidelity 2× PCR Master Mix
837 (NEB). Samples were PCR-amplified (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72
838 °C 60 s) × 12 cycles, held at 12 °C). After PCR, all wells were combined. Libraries were
839 purified according to the MinElute PCR Purification Kit manual (Qiagen) using a vacuum
840 manifold (QIAvac 24 plus, Qiagen) and size selection was performed with SPRI Beads
841 (Beckmann Coulter, 0.55x and 1.5x). Libraries were purified one more time with SPRI
842 Beads (Beckmann Coulter, 1.5x). Libraries were quantified using a Qubit fluorimeter (Life
843 technologies) and the nucleosomal pattern was verified using a Tapestation (High
844 Sensitivity D1000, Agilent). The library was sequenced on a HiSeq2500 sequencer
845 (Illumina) using custom sequencing primers, 25% spike-in library and following read
846 lengths: 50 + 43 + 40 + 50 (Read1 + Index1 + Index2 + Read2)[7].
847
848 **7. Bulk ATAC-seq data generation**
849 ATAC-seq was performed on 30,000-50,000 nuclei as described previously with
850 modifications[3]. Nuclei were thawed on ice and pelleted for 5 min at 500 x g at 4 °C. Nuclei
851 pellets were resuspended in 30 µl tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8),
852 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF) and counted on a hemocytometer.
853 30,000-50,000 nuclei were used for tagmentation and the reaction volume was adjusted
854 to 19 µl using tagmentation buffer. After addition of 1 µl TDE1 (Illumina FC-121-1030),
855 tagmentation was performed at 37°C for 60 min with shaking (500 rpm). Tagmented
856 DNA was purified using MinElute columns (Qiagen), PCR-amplified for 8 cycles with
857 NEBNext® High-Fidelity 2X PCR Master Mix (NEB, 72°C 5 min, 98°C 30 s, [98°C 10 s,
858 63°C 30 s, 72°C 60 s] x 8 cycles, 12°C held). Amplified libraries were purified using
859 MinElute columns (Qiagen) and SPRI Beads (Beckmann Coulter). Sequencing was
860 carried out on a NextSeq500 using a 150-cycle kit (75 bp PE, Illumina).
861
862 **8. Bulk ATAC-seq data analysis**
863 ATAC-seq reads were mapped to reference genome mm10 using bowtie version 2.2.6 and
864 *samtools* version 1.2 to eliminate PCR duplicates and mitochondrial reads. The paired-
865 end read ends were converted to fragments. Using fragments, MACS2[46] version 2.1.2was
866 used for generating signal tracks and peak calling with the following parameters: -q 0.01

867 --nomodel -B --SPMR --keep-dup all. Library quality control for bulk ATAC-seq can be
868 found in Supplementary Table 10.
869
870 **9. Analysis of other single cell datasets**
871 *9.1 Analysis of single nucleus RNA-seq*.
872 We downloaded gene table from dbGaP under accession code phs000424.v8.p1.
873 Analysis code can be found in **Supplementary Note 6**.
874
875 *9.2 Analysis of multiplexing single cell Hi-C*.
876 The processed data is obtained from GEO with accession code GSE84920. Analysis code
877 used in this study can be found in **Supplementary Note 7**.
878
879 *9.3 Analysis of single cell ChIP-seq*.
880 We downloaded single cell matrix from https://pubs.broadinstitute.org/drop-chip/.
881 Analysis code used in this study can be found in **Supplementary Note 8**.
882
883 *9.4 Analysis of single nucleus methylome-seq*.
884 100kb-bin single nucleus methylome datasets were shared by Mukamel lab. We binarized
885 the data by converting the methylation level into z-score and then set the bins with z-score
886 less than -1.5 to 1.

887 **REFERENCE**

888 1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in
889 the human genome. *Nature* **489**, 57–74 (2012).
890 2. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature*
891 **488**, 116–120 (2012).
892 3. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J.
893 Transposition of native chromatin for fast and sensitive epigenomic profiling of open
894 chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–
895 1218 (2013).
896 4. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open
897 Chromatin across the Genome. *Cell* **132**, 311–322 (2008).
898 5. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by
899 combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
900 6. Cusanovich, D. A. *et al.* The *cis*-regulatory dynamics of embryonic development at
901 single-cell resolution. *Nature* **555**, 538–542 (2018).
902 7. Preissl, S. *et al.* Single-nucleus analysis of accessible chromatin in developing mouse
903 forebrain reveals cell-type-specific transcriptional regulation. *Nat. Neurosci.* **21**, 432
904 (2018).
905 8. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin
906 Accessibility. *Cell* **174**, 1309-1324.e18 (2018).
907 9. Lake, B. B. *et al.* Integrative single-cell analysis of transcriptional and epigenetic
908 states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
909 10. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of
910 regulatory variation. *Nature* **523**, 486–490 (2015).
911 11. Mezger, A. *et al.* High-throughput chromatin accessibility profiling at single-cell
912 resolution. *Nat. Commun.* **9**, 3647 (2018).
913 12. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the
914 regulatory epigenome. *Nat. Rev. Genet.* 1 (2019). doi:10.1038/s41576-018-0089-8
915 13. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring
916 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat.*
917 *Methods* **14**, 975–978 (2017).
918 14. González-Blas, C. B. *et al.* Cis-topic modelling of single cell epigenomes. *bioRxiv*
919 370346 (2018). doi:10.1101/370346
920 15. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell
921 Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).
922 16. Stuart, T. *et al.* Comprehensive integration of single cell data. *bioRxiv* 460147
923 (2018). doi:10.1101/460147
924 17. Tasic, B. *et al.* Adult mouse cortical cell taxonomy revealed by single cell
925 transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
926 18. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of
927 communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).

928 19. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors
929 Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol.*
930 *Cell* **38**, 576–589 (2010).
931 20. McLean, C. Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory
932 regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
933 21. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal
934 update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
935 22. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions
936 Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).
937 23. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell
938 differentiation. *Nature* **518**, 331–336 (2015).
939 24. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999-
940 1014.e22 (2018).
941 25. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the
942 Adult Mouse Brain. *Cell* **174**, 1015-1030.e16 (2018).
943 26. Tasic, B. *et al.* Shared and distinct transcriptomic cell types across neocortical areas.
944 *Nature* **563**, 72 (2018).
945 27. Luo, C. *et al.* Single-cell methylomes identify neuronal subtypes and regulatory
946 elements in mammalian cortex. *Science* **357**, 600–604 (2017).
947 28. Ecker, J. R. *et al.* The BRAIN Initiative Cell Census Consortium: Lessons Learned
948 toward Generating a Comprehensive Brain Cell Atlas. *Neuron* **96**, 542–557 (2017).
949 29. Mayer, C. *et al.* Developmental diversification of cortical inhibitory interneurons.
950 *Nature* **555**, 457–462 (2018).
951 30. Gallant, S. & Gilkeson, G. ETS transcription factors and regulation of immunity.
952 *Arch. Immunol. Ther. Exp. (Warsz.)* **54**, 149–163 (2006).
953 31. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—
954 a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92
955 (2007).
956 32. Gorkin, D. *et al.* Systematic mapping of chromatin state landscapes during mouse
957 development. *bioRxiv* 166652 (2017). doi:10.1101/166652
958 33. Habib, N. *et al.* Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat.*
959 *Methods* **14**, 955–958 (2017).
960 34. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by
961 chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
962 35. Ramani, V. *et al.* Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266
963 (2017).
964 36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler
965 transform. *Bioinformatics* **25**, 1754–1760 (2009).
966 37. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,
967 2078–2079 (2009).

38. Chung, N. C. & Storey, J. D. Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics* **31**, 545–554 (2015).
39. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
40. Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323.e30 (2016).
41. Maaten, L. van der. Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
42. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Efficient Algorithms for t-distributed Stochastic Neighborhood Embedding. *ArXiv171209005 Cs Stat* (2017).
43. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2018).
44. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *bioRxiv* 538652 (2019). doi:10.1101/538652
45. Dijk, D. van *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729.e27 (2018).
46. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
47. Lister, R. *et al.* Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science* **341**, 1237905 (2013).
48. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

993 **Supplementary Figures**

994 **Figure S1**. Comparison of normalization methods.

995 **Figure S2**. SnapATAC adjusts for coverage bias.

996 **Figure S3**. Validation of SnapATAC performance relative to alternative methods on
997 simulated datasets.

998 **Figure S4**. SnapATAC identifies cell types using off-peak reads.

999 **Figure S5**. Off-peaks in single cell ATAC-seq library are enriched for euchromatin
1000 regions.

1001 **Figure S6**. SnapATAC identifies potential batch effects and GM subtypes.

1002 **Figure S7**. Gene accessibility level for selected markers of cell types expected in the
1003 pre-frontal cortex.

1004 **Figure S8**. Sub-clustering identifies inhibitory neuronal subtypes in pre-frontal cortex
1005 sciATAC-seq data.

1006 **Figure S9**. SnapATAC identifies additional cell types from published mouse cerebellum
1007 and kidney sci-ATAC-seq.

1008 **Figure S10**. SnapATAC identifies additional cell types from LSA-defined major
1009 clusters.

1010 **Figure S11**. SnapATAC outperforms existing methods in scalability.

1011 **Figure S12**. Reproducibility between two biological replicates for mouse cortex at the
1012 level of aggregate signal.

1013 **Figure S13**. Distribution of single cell quality control metrics for replicate 1.

1014 **Figure S14**. Distribution of single cell quality control metrics for replicate 2.

1015 **Figure S15**. Reproducibility of clustering result between two biological replicates.

1016 **Figure S16**. Gene accessibility level for selected markers of major cell types expected in
1017 the mouse motor cortex.

1018 **Figure S17**. Gene accessibility level for selected markers of inhibitory neuronal
1019 subtypes expected in the mouse motor cortex.

1020 **Figure S18**. Comparison with scRNA-seq

1021 **Figure S19**. Subdivision reveals rare Sst subtypes.

1022 **Figure S20**. Subdivision reveals 37 neuronal subtypes.

1023 **Figure S21**. Application of SnapATAC to single-nucleus RNA-seq.

1024 **Figure S22**. Gene expression level for selected markers of major cell types expected in
1025 the human brain.
1026 **Figure S23**. Sub-clustering of GABAergic neurons reveals subtypes in single nucleus
1027 RNA-seq.
1028 **Figure S24**. Application of SnapATAC to single-nucleus methylome.
1029 **Figure S25**. Bin coverage distribution.
1030 **Figure S26**. Partial Jaccard index matrix using different number of features.
1031 **Figure S27**. Illustration of secondary motor cortex dissection.
1032 **Figure S28**. Gating strategy for nuclei sorting.
1033
1034 **Supplementary Tables**
1035 **Table S1**. Ten bulk ATAC-seq used for simulating single cell ATAC-seq datasets
1036 **Table S2**. Comparison of clustering performance of five methods on simulated datasets
1037 **Table S3**. Alignment statistics for mouse motor cortex single nucleus ATAC-seq library
1038 **Table S4**. Metadata of mouse MOs single neuron ATAC-seq
1039 **Table S5**. A merged list of *c*is-Regulatory elements from all cell clusters
1040 **Table S6**. *C*is-Regulatory elements identified using bulk ATAC-seq
1041 **Table S7**. Motifs enriched for cell-type specific elements
1042 **Table S8**. VISTA enhancers overlapping with new cis-elements
1043 **Table S9**. Barcode indices used for single nucleus ATAC-seq experiment
1044 **Table S10**. Alignment statistics for mouse motor cortex bulk ATAC-seq library
1045