APEC: An accesson-based method for single-cell chromatin accessibility analysis

Bin Li[1,3], Young Li[1,3], Kun Li[1], Lianbang Zhu[1], Qiaoni Yu[1], Jingwen Fang[1], Pengfei Cai[1], Chen Jiang[1], Kun Qu[1,2,*]

Affiliation

[1]Division of Molecular Medicine, Hefei National Laboratory for Physical Sciences at Microscale, The CAS Key Laboratory of Innate Immunity and Chronic Disease, School of Life Sciences, University of Science and Technology of China

[2]CAS Center for Excellence in Molecular Cell Sciences, University of Science and Technology of China

[3]Cofirst authors

*Correspondence: Kun Qu (qukun@ustc.edu.cn)

Contact Information

Kun Qu, Ph.D.

Division of Molecular Medicine, Hefei National Laboratory for Physical Sciences at Microscale, The CAS Key Laboratory of Innate Immunity and Chronic Disease, School of Life Sciences, University of Science and Technology of China

Hefei, Anhui, China, 230027

Email: qukun@ustc.edu.cn

Phone: +86-551-63606257

Abstract:

The development of sequencing technologies has promoted the survey of genome-wide chromatin accessibility at single-cell resolution; however, comprehensive analysis of single-cell epigenomic profiles remains a challenge. Here, we introduce an accessibility pattern-based epigenomic clustering (APEC) method, which classifies each individual cell by groups of accessible regions with synergistic signal patterns termed “accessons”. By integrating with other analytical tools, this python-based APEC package greatly improves the accuracy of unsupervised single-cell clustering for many different public data sets. APEC also identifies significant differentially accessible sites, predicts enriched motifs, and projects pseudotime trajectories. Furthermore, we developed a fluorescent tagmentation- and FACS-sorting-based single-cell ATAC-seq technique named ftATAC-seq and investigated the per cell regulome dynamics of mouse thymocytes. Associated with ftATAC-seq, APEC revealed a detailed epigenomic heterogeneity of thymocytes, characterized the developmental trajectory and predicted the regulators that control the stages of maturation process. Overall, this work illustrates a powerful approach to study single-cell epigenomic heterogeneity and regulome dynamics.

INTRODUCTION

As a technique for probing genome-wide chromatin accessibility in a small number of cells *in vivo*, the assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) has been widely applied to investigate the cellular regulomes of many important biological processes[1], such as hematopoietic stem cell (HSC) differentiation[2-4], embryonic development[5, 6], neuronal activity and regeneration[7-10], tumor cell metastasis[11], and patient responses to anticancer drug treatment[12]. Recently, several experimental schemes have been developed to capture chromatin accessibility at single-cell/nucleus resolution, i.e., single-cell ATAC-seq (scATAC-seq)[13], single-nucleus ATAC-seq (snATAC-seq)[14, 15], and single-cell combinatorial indexing ATAC-seq (sci-ATAC-seq)[16, 17], which significantly extended researchers' ability to uncover cell-to-cell epigenetic variation and other fundamental mechanisms that generate heterogeneity from identical DNA sequences. By contrast, the in-depth analysis of single-cell chromatin accessibility profiles for this purpose remains a challenge. Numerous efficient algorithms have been developed to accurately normalize, cluster and visualize cells from single-cell transcriptome sequencing profiles, including but not limited to SCnorm[18], Seurat[19], SC3[20], SIMLR[21], bigSCale[22], and SCANPY[23]. However, most of these algorithms are not directly compatible with a single-cell ATAC-seq dataset, for which the signal matrix is much sparser. To characterize scATAC-seq data, the Greenleaf lab developed an algorithm named chromVAR[24], which aggregates mapped reads at accessible sites based on annotated motifs of known transcription factors (TFs) and thus projects the sparse per accessible peak per cell matrix to a bias-corrected deviation per motif per cell matrix and significantly stabilizes the data matrix for downstream clustering analysis. Other mathematical tools, such as the latent semantic indexing (LSI) and density-based clustering methods, have also been applied to process single-cell/nucleus ATAC-seq data[15, 17]. However, none of these methods can precisely distinguish cells from low sequencing depth without prior knowledge of essential principle components or TF motifs. Therefore, a refined algorithm is urgently needed to better categorize cell subgroups with minor differences under low coverage, thereby providing a deeper mechanistic understanding of single-cell epigenetic heterogeneity and regulation.

RESULTS

Accesson-based algorithm improves single-cell clustering

Here, we introduce a new single-cell chromatin accessibility analysis toolkit named APEC (accessibility pattern-based epigenomic clustering), which combines peaks with the same signal fluctuation among all single cells into peak groups, termed "accessons", and converts the original sparse cell-peak matrix to a much denser cell-accesson matrix for cell type categorization (Figure 1a). In contrast to previous motif-based methods (e.g., chromVAR), this accesson-based reduction scheme naturally groups synergistic accessible regions together without a priori knowledge of genetic information (such as TF motifs) and provides more efficient, accurate and rapid cell clustering from single-cell ATAC-seq profiles. More conveniently, APEC integrates all necessary procedures, from raw sequence trimming, alignment, and quality control (Supplementary Figure 1a-1c) to cell clustering, motif enrichment, and pseudotime trajectory prediction into a head-to-toe program package that has been made available on GitHub (https://github.com/QuKunLab/APEC).

To test the performance of APEC, we first obtained data from previous publications, which performed scATAC-seq on lymphoid-primed multipotent progenitors (LMPPs), monocytes, HL-60 lymphoblastoid cells (HL60), and blast cells and leukemic stem cells (LSCs) from two acute myeloid leukemia (AML) patients[24]. Compared to the motif-based method chromVAR[24, 25], this new accesson-based algorithm more precisely and clearly clustered cells into their corresponding identities (Figure 1b-1e). For instance, distinct cell types, such as LMPPs, monocytes and HL60 cells, were more vividly separated from each other (Adjusted Rand Index (ARI)=0.95, compared to ARI=0.59 for chromVAR); similar cells, such as the blast cells and LSCs from two AML patients, were ambiguous in chromVAR (ARI=0.36) but were more clearly categorized in both the hierarchical clustering heatmap and the tSNE scattering plot in APEC (ARI=0.69). The contribution of the minor differences between similar cells is aggregated in accessons but diluted in motifs. For example, APEC identified prominent superenhancers around the E3 ligase inhibitor gene *N4BP1*[26] and the MLL fusion gene *GPHN*[27] in the LSC cells from AML patient 1 (P1-LSC) but not in the other cell types (Figure 1f, Supplementary Figure 1d). We noticed that all peaks in these superenhancers were classified into one accesson that was critical for distinguishing P1-LSCs from P2-LSCs, P1-blast cells and P2-blast cells. However, these peaks were distributed in multiple TF motifs, which significantly diluted the contributions of the minor differences (Supplementary Figure 1e-f). To test the robustness of APEC at low sequencing depth, we

107 randomly selected reads from the raw data and calculated the ARI values for each sampled data.
108 Compared with chromVAR, the APEC algorithm exhibits better robustness at sequencing depth
109 as low as 10% of the original data (Supplementary Figure 1g).
110
111 **APEC is applicable to multiple single-cell chromatin detection techniques**
112
113 To evaluate the compatibility and performance of APEC with other single-cell chromatin
114 accessibility detection techniques, such as snATAC-seq[15], transcript-indexed scATAC-seq[28] and
115 sciATAC-seq[16], APEC was also tested with the data sets generated by those experiments. For
116 example, APEC discovered 10 cell subpopulations in adult mouse forebrain snATAC-seq data[15],
117 including three clusters of excitatory neurons (EX1-3), five groups of inhibitory neurons (IN1-4),
118 astroglia cells (AC), oligodendrocyte cells (OC), and microglial cells (MG; Figure 2a & 2b), as
119 defined by the chromatin accessibilities at the loci of cell type-specific genes (Figure 2c).
120 Compared to published results[15], APEC identified 4 rather than 2 distinct inhibitory subpopulations,
121 among which IN1 and IN4 were more similar and IN2 and IN3 were more distinct (Figure 2d). The
122 motif enrichment analysis module in APEC identified cell type-specific regulators that are
123 consistent with previous publications[15]. For example, the NEUROD1 and OLIG2 motifs were
124 generally enriched on excitatory clusters (EX1~3); the MEF2C motif was more enriched on EX3
125 than on EX1/2 neurons; the motifs of MEIS2 and DLX2 were differentially enriched on two
126 subtypes of inhibitory neurons (IN2 and IN3, respectively); and the NOTO, SOX2, and ETS1
127 motifs were enriched on the AC, OC, and MG clusters, respectively (Figure 2e). These results
128 suggest that APEC is capable of identifying cell subtype-specific regulators.

129 Since the divergence of the gene expression levels in a single cell is much greater than
130 that of the chromatin accessibilities, single-cell transcriptome analysis usually identifies more cell
131 subpopulations. Therefore, it is critical to anchor the cell types identified from scATAC-seq to
132 those from scRNA-seq. Lake et al. identified dozens of excitatory and inhibitory neuronal subtypes
133 in the adult human brain using snDrop-seq and scTHS-seq experiments[14] and provided tens of
134 signature genes that distinguished these cell types. Interestingly, the accessons that represent
135 these signature genes were also distinctly enriched at corresponding clusters of neurons. For
136 example, the upright part of the EX1 cell cluster in snATAC-seq enriched accessons represents
137 the genes *Cbln2* and *Col5a2*, which are specific genes in clusters Ex1/2/3a that were defined in
138 the scDrop-seq data (Figure 2f). The left part of the EX1 cell cluster in snATAC-seq matched the
139 Ex3b/3c/3e clusters in scDrop-seq (marked by *Nefm*), EX2 matched Ex4/5/6 (marked by *Foxp2*
140 and *Pcp4*), and EX3 matched Ex3d (marked by *Phactr2*). The same method also works to anchor

141 inhibitory neurons, as the IN2 cells in the snATAC-seq data corresponded to the In1/2 clusters in
142 the scDrop-seq data (marked by *Cck* and *Cnr1*), the IN3 cells corresponded to the In6b/8 clusters
143 (marked by *Stxbp6* and *Tac1*), the IN4 cells corresponded to the In1c/3 clusters (marked by *Vip*
144 and *Tshz2*), and the low right branch of IN1 corresponded to the In7 cluster (marked by *Npy*)
145 (Figure 2g). These results highlight the potential advantages of the accesson-based method for
146 integrative analysis of scRNA-seq and scATAC-seq data.

147 In addition, due to the sparser per-cell-per-peak fragment count matrix, more than 29.7%
148 (946 out of 3034) of high-quality cells were previously unable to be correctly assigned into any
149 subpopulation of interest[15], but APEC successfully categorized all cells into their corresponding
150 subtypes, confirming its high sensitivity. In contrast, chromVAR misclustered AC and EX4 with
151 inhibitory neurons, although the same parameters were applied (Supplementary Figure 2a-2c).
152 These results confirm that this accesson-based APEC method can better distinguish and
153 categorize single cells with great sensitivity and reliability.

154

155 **APEC constructs a pseudotime trajectory that predicts cell differentiation lineage**

156

157 Cells are not static but dynamic entities, and they have a history, particularly a
158 developmental history. Although single-cell experiments often profile a momentary snapshot, a
159 number of remarkable computational algorithms have been developed to pseudo-order cells
160 based on the different points they were assumed to occupy in a trajectory, thereby leveraging
161 biological asynchrony[29, 30]. For instance, Monocle[30, 31] constructs the minimum spanning tree, and
162 Wishbone[32] and Spring[33] construct the nearest neighbor graph from single-cell transcriptome
163 profiles. These tools have been widely used to depict neurogenesis[34], hematopoiesis[35, 36] and
164 reprogramming[37]. APEC integrates the Monocle algorithm into the accesson-based method and
165 enables pseudotime prediction from scATAC-seq data[38] and was applied to investigate HSC
166 differentiation linages (Figure 3a). Principal component analysis (PCA) of the accesson matrix
167 revealed multiple stages of the lineage during HSC differentiation (Figure 3b) and was consistent
168 with previous publications[3, 38]. After utilizing the Monocle package, APEC provided more precise
169 pathways from HSCs to the differentiated cell types (Figure 3c). In addition to the differentiation
170 pathways to MEP cells through the CMP state and to CLP cells through the LMPP state, MPP
171 cells may differentiate into GMP cells through two distinct trajectories: Path A through the CMP
172 state and Path B through the LMPP state, which is consistent with the composite model of HSC
173 and blood lineage commitment[39]. Notably, APEC suggested that $CD34^+$ plasmacytoid dendritic

174 cells (pDCs) from the bone marrow (Supplementary Figure 3) were derived from CLP cells on the
175 psuedotime trajectory (Figure 3c), which also agrees with a previous report[40]. Furthermore, APEC
176 is capable of evaluating the deviation of each TF along the single-cell trajectory to determine the
177 regulatory mechanisms during HSC differentiation. As expected, the HOX motif is highly enriched
178 in the accessible sites of HSCs/MPP cells, as are the GATA1, CEBPB and TCF4 motifs, which
179 exhibit gradients that increase along the erythroid, myeloid and lymphoid differentiation pathways,
180 respectively[38] (Figure 3d). In addition, we can see that the TF regulatory strategies of the two
181 paths from MMPs towards GMP cells were very different. Finally, we generated a hematopoiesis
182 tree based on the APEC analysis (Figure 3e).
183
184 **APEC reveals the single-cell regulatory heterogeneity of thymocytes**
185
186 T cells generated in the thymus play a critical role in the adaptive immune system, and
187 the development of thymocytes can be divided into 3 main stages based on the expression of the
188 surface markers CD4 and CD8, namely, CD4 CD8 double-negative (DN), CD4 CD8 double-
189 positive (DP) and CD4 or CD8 single-positive (CD4SP or CD8SP, respectively) stages[41]. However,
190 due to technical limitations, our genome-wide understanding of thymocyte development at single-
191 cell resolution remains unclear. Typically, more than 80% of thymocytes stay in the DP stage in
192 the thymus, whereas DN cells account for only approximately 3% of the thymocyte population. To
193 eliminate the impacts of great differences in proportion, we developed a fluorescent tagmentation-
194 and FACS-sorting-based scATAC-seq strategy (ftATAC-seq), which combined the advantages of
195 ATAC-see[42] and Pi-ATAC-seq[43] to manipulate the desired number of target cells by indexed
196 sorting (Figure 4a). Tn5 transposomes were fluorescently labeled in each cell to evaluate the
197 tagmentation efficiency so that cells with low ATAC signals could be gated out easily
198 (Supplementary Figure 4a, Figure 4b). With ftATAC-seq, we acquired high-quality chromatin
199 accessibility data for 352 index-sorted DN, DP, CD4SP, and CD8SP single cells and 352 mixed
200 thymocytes (Figure 4b). We applied APEC to mouse thymocyte ftATAC-seq data to investigate
201 the chromatin accessibility divergence during the developmental process and to reveal refined
202 regulome heterogeneity at single-cell resolution. Taking into account all 130685 peaks called from
203 the raw sequencing data, APEC aggregated 600 accessons and successfully assigned over 92%
204 of index-sorted DN, DP, CD4SP and CD8SP cells into the correct subpopulations (Figure 4c, 4d),
205 providing a much better classification than chromVAR (Supplementary Figure 4b, 4c), for which
206 this rate was only 56%. As expected, the majority of randomly sorted and mixed thymocytes were
207 classified into DP subtypes based on similarity hierarchical clustering, which was consistent with

208 the cellular subtype proportions in the thymus. APEC further classified all thymocytes into 14
209 subpopulations, including 2 DN, 7 DP, 1 CD4SP, 2 CD8SP, 1 coherence (Coh.A) and 1 transition
210 (Tran.A) state, suggesting that extensive epigenetic heterogeneity exists among cells with the
211 same CD4 and CD8 surface markers (Figure 4e). For instance, there are four main subtypes of
212 DN cells, according to the expression of the surface markers CD44 and CD25[44], while two clusters
213 were identified in ftATAC-seq. The accessibility signals around the *Il2ra* (Cd25) and *Cd44* gene
214 loci demonstrated that DN.A1 comprised $CD44^{+}CD25^{-}$ and $CD44^{+}CD25^{+}$ DN subtypes (DN1 and
215 DN2), and DN.A2 cells comprised $CD44^{-}CD25^{+}$ and $CD44^{-}CD25^{-}$ subtypes (DN3 and DN4),
216 suggesting significant chromatin changes between DN2 and DN3 cell development (Figure 4f).
217 Many TFs have been reported to be essential in regulating thymocyte development, and
218 we found that their motifs were remarkably enriched at different stages during the process (Figure
219 4g). For instance, Runx3 is well known for regulating CD8SP cells[45], and we observed significant
220 enrichment of the RUNX motif on DN cells and a group of CD8SP cells. Similarly, the TCF[46, 47],
221 RORC[48] and NFkB[49] family in regulating the corresponding stages during this process. More
222 enriched TF motifs in each cell subpopulation were also observed, suggesting significant
223 regulatory divergence in thymocytes (Supplementary Figure 4d). Interestingly, two clusters of
224 CD8SP cells appear to be differentially regulated based on motif analysis, in which CD8.A1 cells
225 are closer to DP cells, while CD8.A2 cells are more distant at the chromatin level, suggesting that
226 CD8.A2 cells are more mature CD8SP cells. In addition to the well-defined subtypes, APEC also
227 found a mixed cell population without specific features that was termed the coherence state
228 (Coh.A) and a transitional population between DP and SP cells (Tran.A).
229 APEC is capable of integrating single-cell transcriptional and epigenetic information by
230 scoring gene sets of interest based on their nearby peaks from scATAC-seq, thereby converting
231 the chromatin accessibility signals to values that are comparable to gene expression profiles
232 (online **Methods**). To test the performance of this integrative analysis approach and to evaluate
233 the accuracy of thymocyte classification by APEC, we assayed the transcriptomes of single
234 thymocytes and obtained 357 high-quality scRNA-seq profiles using the SMART-seq2 protocol[50].
235 Unsupervised analysis of gene expression profiles clustered these thymocytes into 13 groups in
236 Seurat[19] (Supplementary Figure 5a, 5b), and each subpopulation was identified based on known
237 feature genes (Supplementary Figure 5c, 5d). We then compared the adjusted scores obtained
238 from APEC with the single-cell RNA expression profile and observed a strong correlation between
239 the subtypes identified from the transcriptome and the subtypes identified from chromatin
240 accessibility (Figure 4h), confirming the reliability and stability of cellular classification using APEC.
241

242 **APEC reconstructs the thymocyte developmental trajectory from ftATAC-seq profiles**

243

244 APEC is capable of constructing a pseudotime trajectory and then predicting the cell
245 differentiation lineage from a "snapshot" of single-cell epigenomes (Figure 3). We applied APEC
246 to recapitulate the developmental trajectory and thereby reveal the single-cell regulatory dynamics
247 during the maturation of thymocytes. Psuedotime analysis based on single-cell ftATAC-seq data
248 shaped thymocytes into 5 developing stages (Figure 5a, Supplementary Figure 6a-b), where most
249 of the cells in stages 1, 2, 4, and 5 were DN, DP, CD8SP and CD4SP cells, respectively. APEC
250 also identified a transitional stage 3, which consisted of DP, coherence and transitional cells.
251 Interestingly, the psuedotime trajectory suggests three developmental pathways for this process,
252 one of which started with stage 1 (DN) and ended in stage 2 (DP), and the other two of which
253 started with stage 1 (DN), went through a transitional stage 3 (a mixture of DP, Coh and Tran)
254 and then bifurcated into stage 4 (CD8SP) and 5 (CD4SP). The predicted developmental trajectory
255 could also be confirmed by the gene expression of surface markers, such as Cd4, Cd8, Runx3
256 and Ccr7 (Figure 5b). To evaluate the gene ontology (GO) enrichments over the entire process,
257 we implemented an accesson-based GO module in APEC, which highlights the significance of
258 the association between cells and biological function (Figure 5c). For instance, T cells selections,
259 including β-selection, positive selection and negative selection, start from the DN3 stage.
260 Consistent with this process, we observed a strong "T cell selection" GO term on the trajectory
261 path after DN.A1. Since TCR signals are essential for T cell selection, we also observed the "T
262 cell activation" GO term accompanied by "T cell selection". Meanwhile, the regulation of protein
263 binding signal was also decreased at SP stages, indicating the necessity of weak TCR signal for
264 the survive of SP T cells during negative selection.

265 To further uncover the regulatory mechanism underlying this developmental process,
266 APEC was implemented to identify stage-specific enriched TFs along the trajectory and pinpoint
267 the "pseudotime" at which the regulation occurs. In addition to the well-studied TFs mentioned
268 above (Figure 4g, Supplemental Figure 4c), APEC also identified Zeb1[51], Ctcf[52] and Id4 as
269 potential stage-specific regulators (Figure 5d). Interestingly, the Id4 motif enriched on DP cells
270 was also reported to regulate apoptosis in other cell types[53, 54]. Associated with the fact that the
271 vast majority of DP thymocytes die because of a failure of positive selection[55], we hypothesize
272 that stage 2 may be the path towards DP cell apoptosis. We then checked the distribution of DP
273 cells along the stage 2 trajectory and found that most DP.A1 cells were scattered in "early" stage
274 2, and they were enriched with GO terms such as "T cell selection", "cell activation" and
275 "differentiation" (Figure 5e, Supplementary Figure 6c). However, most DP.A5-A6 cells were

distributed at the end of stage 2, and their principle accessons were enriched with GO terms such as “apoptosis” and “chromatin modification”. These results suggest that a majority of DP thymocytes undergo T cell selection and enter an apoptosis state. Although it is believed that more than 95% of DP thymocytes are subjected to death in positive selection, only a small proportion of apoptotic cells could be detected in a snapshot of the thymus. By comparing the number of cells near stage 3 with all the cells in stage 2, we estimated that ~3-5% of cells would survive positive selection, which is consistent with previous publications[56, 57]. Our data suggest that before entering an apoptotic stage, DP thymocytes that fail selection could have already committed to apoptosis at the chromatin level.

DISCUSSION

Here, we introduced an accesson-based algorithm for single-cell chromatin accessibility analysis. Without any prior information (such as motifs), this approach generated more refined cell groups with reliable biological functions and properties. Integrating the new algorithm with all necessary chromatin sequencing data processing tools, APEC provides a comprehensive solution for transforming raw experimental single-cell data into final visualized results. In addition to better clustering of subtle cell subtypes, APEC is also capable of locating potential specific superenhancers, searching enriched motifs, estimating gene opening scores, and building time-dependent cell developmental trajectories, and it is compatible with many existing single-cell accessibility datasets. Despite these advantages, the biological implications of accessons are still obscure, especially for those that involve only a small number of peaks; therefore, further investigations may require uncovering the biology that underlies accessons.

To evaluate the performance of this approach in the context of the immune system, we also adopted APEC with scATAC-seq technology to investigate the regulome dynamics of the thymic development process. We developed a novel method of ftATAC-seq that captures Tn5-tagged single cells of interest and outlines the chromatin accessibility heterogeneity and dynamics during this process. Coordinated with essential cell surface markers, APEC provided a much more in-depth classification of thymocytes than the conventional DN, DP, CD4SP and CD8SP stages based on single-cell chromatin status. By reconstructing the developmental pseudotime trajectory, APEC discovered a transitional stage before thymocytes bifurcate into CD4SP and CD8SP cells and inferred that one of the stages leads to cell apoptosis. APEC analysis suggested that DP cells

were gradually programmed to undergo apoptosis at the chromatin level; however, further studies are needed to fully understand the regulatory mechanism of this process.

Acknowledgments

This work was supported by the National Key R&D Program of China (2017YFA0102903 to K.Q.) and by National Natural Science Foundation of China grants 91640113 (to K.Q.) and 31771428 (to K.Q.). It was also supported by Anhui Provincial Natural Science Foundation grant BJ2070000097 (to B.L.). We thank the Howard Chang lab at Stanford University for helpful discussion. We thank the USTC supercomputing center and the School of Life Science Bioinformatics Center for providing supercomputing resources for this project.

Authors' contributions

KQ, BL and YL conceived the project, BL developed the APEC software and performed all data analysis with helps from QY, JF, PC, and CJ. YL developed ftATAC-seq technique and performed all scATAC-seq and scRNA-seq experiments with helps from LZ. KL analyzed scRNA-seq data. BL, YL and KQ wrote the manuscript with inputs from all other authors.

Data and code availability

Mouse thymocytes ftATAC-seq data can be obtained from the Genome Sequence Archive at BIG Data Center with the accession number CRA001267 and is available via http://bigd.big.ac.cn/gsa/s/yp1164Et. Other published data sets used in this study are available from NIH GEO with accession numbers GSE74310[3], GSE65360[13], GSE96772[38], and GSE100033[15]. APEC pipeline can be downloaded from the GitHub website (https://github.com/QuKunLab/APEC).

334 ## REFERENCES

335 1. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native
336 chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins
337 and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).
338 2. Ye, M. et al. Hematopoietic Differentiation Is Required for Initiation of Acute Myeloid Leukemia.
339 *Cell Stem Cell* **17**, 611-623 (2015).
340 3. Corces, M.R. et al. Lineage-specific and single-cell chromatin accessibility charts human
341 hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193-1203 (2016).
342 4. Yu, V.W.C. et al. Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of
343 Hematopoietic Stem Cells. *Cell* **167**, 1310-1322 e1317 (2016).
344 5. Wu, J. et al. The landscape of accessible chromatin in mammalian preimplantation embryos.
345 *Nature* **534**, 652-657 (2016).
346 6. Wu, J. et al. Chromatin analysis in human early development reveals epigenetic transition during
347 ZGA. *Nature* **557**, 256-260 (2018).
348 7. Jorstad, N.L. et al. Stimulation of functional neuronal regeneration from Muller glia in adult mice.
349 *Nature* **548**, 103-107 (2017).
350 8. Su, Y. et al. Neuronal activity modifies the chromatin accessibility landscape in the adult brain.
351 *Nat Neurosci* **20**, 476-483 (2017).
352 9. Xiang, Y. et al. Fusion of Regionally Specified hPSC-Derived Organoids Models Human Brain
353 Development and Interneuron Migration. *Cell Stem Cell* **21**, 383-398 e387 (2017).
354 10. de la Torre-Ubieta, L. et al. The Dynamic Landscape of Open Chromatin during Human Cortical
355 Neurogenesis. *Cell* **172**, 289-304 e218 (2018).
356 11. Denny, S.K. et al. Nfib Promotes Metastasis through a Widespread Increase in Chromatin
357 Accessibility. *Cell* **166**, 328-342 (2016).
358 12. Qu, K. et al. Chromatin Accessibility Landscape of Cutaneous T Cell Lymphoma and Dynamic
359 Response to HDAC Inhibitors. *Cancer Cell* **32**, 27-41 e24 (2017).
360 13. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation.
361 *Nature* **523**, 486-490 (2015).
362 14. Lake, B.B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the
363 human adult brain. *Nat Biotechnol* **36**, 70-80 (2018).
364 15. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain
365 reveals cell-type-specific transcriptional regulation. *Nat Neurosci* **21**, 432-439 (2018).
366 16. Cusanovich, D.A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial
367 cellular indexing. *Science* **348**, 910-914 (2015).
368 17. Cusanovich, D.A. et al. The cis-regulatory dynamics of embryonic development at single-cell
369 resolution. *Nature* **555**, 538-542 (2018).
370 18. Bacher, R. et al. SCnorm: robust normalization of single-cell RNA-seq data. *Nat Methods* **14**, 584-
371 586 (2017).
372 19. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic
373 data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
374 20. Kiselev, V.Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486
375 (2017).
376 21. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-
377 cell RNA-seq data by kernel-based similarity learning. *Nat Methods* **14**, 414-416 (2017).
378 22. Iacono, G. et al. bigSCale: an analytical framework for big-scale single-cell data. *Genome Res* **28**,
379 878-890 (2018).

380 23. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis.
381 *Genome Biol* **19**, 15 (2018).
382 24. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-
383 associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975-978 (2017).
384 25. Zamanighomi, M. et al. Unsupervised clustering and epigenetic classification of single cells. *Nat*
385 *Commun* **9**, 2410 (2018).
386 26. Oberst, A. et al. The Nedd4-binding partner 1 (N4BP1) protein is an inhibitor of the E3 ligase Itch.
387 *Proc Natl Acad Sci U S A* **104**, 11280-11285 (2007).
388 27. Eguchi, M. et al. GPHN, a novel partner gene fused to MLL in a leukemia with t(11;14)(q23;q24).
389 *Genes Chromosomes Cancer* **32**, 212-221 (2001).
390 28. Satpathy, A.T. et al. Transcript-indexed ATAC-seq for precision immune profiling. *Nat Med* **24**,
391 580-590 (2018).
392 29. Bendall, S.C. et al. Single-cell trajectory detection uncovers progression and regulatory
393 coordination in human B cell development. *Cell* **157**, 714-725 (2014).
394 30. Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by
395 pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).
396 31. Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods*
397 **14**, 979-982 (2017).
398 32. Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data.
399 *Nat Biotechnol* **34**, 637-645 (2016).
400 33. Weinreb, C., Wolock, S. & Klein, A.M. SPRING: a kinetic interface for visualizing high dimensional
401 single-cell expression data. *Bioinformatics* **34**, 1246-1248 (2018).
402 34. Habib, N. et al. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons.
403 *Science* **353**, 925-928 (2016).
404 35. Olsson, A. et al. Single-cell analysis of mixed-lineage states leading to a binary cell fate choice.
405 *Nature* **537**, 698-+ (2016).
406 36. Zhou, F. et al. Tracing haematopoietic stem cell formation at single-cell resolution. *Nature* **533**,
407 487-+ (2016).
408 37. Treutlein, B. et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell
409 RNA-seq. *Nature* **534**, 391-+ (2016).
410 38. Buenrostro, J.D. et al. Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape
411 of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548 e1516 (2018).
412 39. Adolfsson, J. et al. Identification of Flt3+ lympho-myeloid stem cells lacking erythro-
413 megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**, 295-
414 306 (2005).
415 40. Chistiakov, D.A., Orekhov, A.N., Sobenin, I.A. & Bobryshev, Y.V. Plasmacytoid dendritic cells:
416 development, functions, and role in atherosclerotic inflammation. *Front Physiol* **5**, 279 (2014).
417 41. Germain, R.N. T-cell development and the CD4-CD8 lineage decision. *Nat Rev Immunol* **2**, 309-322
418 (2002).
419 42. Chen, X. et al. ATAC-see reveals the accessible genome by transposase-mediated imaging and
420 sequencing. *Nat Methods* **13**, 1013-1020 (2016).
421 43. Chen, X. et al. Joint single-cell DNA accessibility and protein epitope profiling reveals
422 environmental regulation of epigenomic heterogeneity. *Nat Commun* **9**, 4590 (2018).
423 44. Godfrey, D.I., Kennedy, J., Suda, T. & Zlotnik, A. A developmental pathway involving four
424 phenotypically and functionally distinct subsets of CD3-CD4-CD8- triple-negative adult mouse
425 thymocytes defined by CD44 and CD25 expression. *J Immunol* **150**, 4244-4252 (1993).
426 45. Taniuchi, I. et al. Differential requirements for Runx proteins in CD4 repression and epigenetic
427 silencing during T lymphocyte development. *Cell* **111**, 621-633 (2002).

428 46. Ioannidis, V., Beermann, F., Clevers, H. & Held, W. The beta-catenin--TCF-1 pathway ensures
429 CD4(+)CD8(+) thymocyte survival. *Nat Immunol* **2**, 691-697 (2001).
430 47. Yu, S. et al. The TCF-1 and LEF-1 transcription factors have cooperative and opposing roles in T cell
431 development and malignancy. *Immunity* **37**, 813-826 (2012).
432 48. Sun, Z. et al. Requirement for RORgamma in thymocyte survival and lymphoid organ development.
433 *Science* **288**, 2369-2373 (2000).
434 49. Gerondakis, S., Fulford, T.S., Messina, N.L. & Grumont, R.J. NF-kappaB control of T cell
435 development. *Nat Immunol* **15**, 15-25 (2014).
436 50. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat*
437 *Methods* **10**, 1096-1098 (2013).
438 51. Higashi, Y. et al. Impairment of T cell development in deltaEF1 mutant mice. *J Exp Med* **185**, 1467-
439 1479 (1997).
440 52. Heath, H. et al. CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *EMBO J*
441 **27**, 2839-2850 (2008).
442 53. Andres-Barquin, P.J., Hernandez, M.C. & Israel, M.A. Id4 expression induces apoptosis in astrocytic
443 cultures and is down-regulated by activation of the cAMP-dependent signal transduction pathway.
444 *Exp Cell Res* **247**, 347-355 (1999).
445 54. Carey, J.P., Knowell, A.E., Chinaranagari, S. & Chaudhary, J. Id4 promotes senescence and
446 sensitivity to doxorubicin-induced apoptosis in DU145 prostate cancer cells. *Anticancer Res* **33**,
447 4271-4278 (2013).
448 55. Surh, C.D. & Sprent, J. T-cell apoptosis detected in situ during positive and negative selection in
449 the thymus. *Nature* **372**, 100-103 (1994).
450 56. Huesmann, M., Scott, B., Kisielow, P. & von Boehmer, H. Kinetics and efficacy of positive selection
451 in the thymus of normal and T cell receptor transgenic mice. *Cell* **66**, 533-540 (1991).
452 57. Shortman, K., Vremec, D. & Egerton, M. The kinetics of T cell antigen receptor expression by
453 subgroups of CD4+8+ thymocytes: delineation of CD4+8+3(2+) thymocytes as post-selection
454 intermediates leading to mature T cells. *J Exp Med* **173**, 323-332 (1991).
455

FIGURES

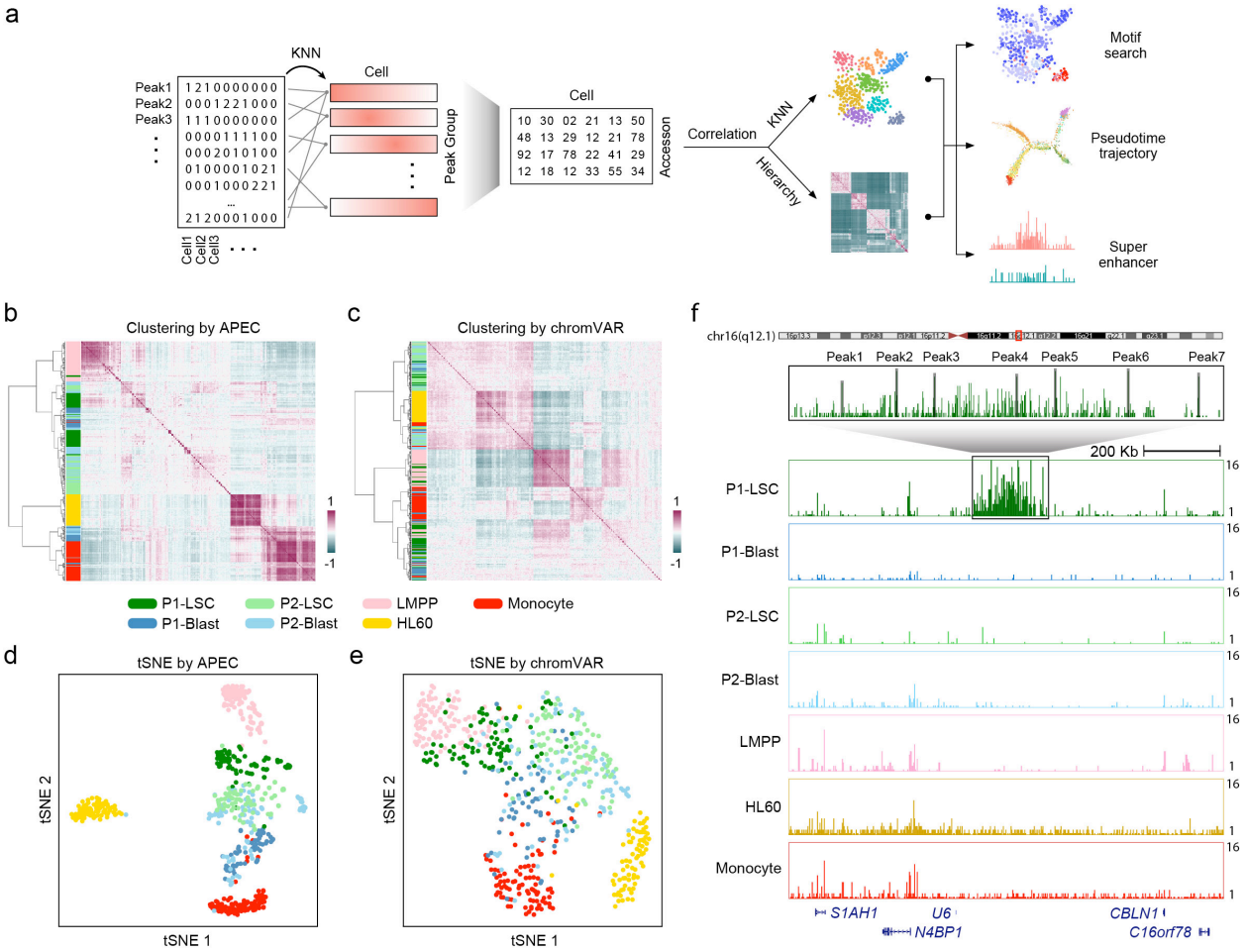

Figure 1. The accesson matrix constructed from the sparse fragment count matrix improved the clustering of scATAC-seq data. (**a**) Step-by-step workflow of APEC. Peaks were grouped into accessons by their accessibility pattern among cells with the K nearest neighbors (KNN) method. (**b, c**) Hierarchical clustering of cell-cell correlations based on the accesson matrix (from APEC) and the motif matrix (from chromVAR). The scATAC-seq data include leukemic stem cells (LSCs), leukemia blast cells, lymphoid-primed multipotential progenitors (LMPPs), HL60 cells, and monocytes. P1, acute myeloid leukemia (AML) patient 1 (SU070); P2, AML patient 2 (SU353). The cells are labeled by their fluorescence indices. (**d, e**) t-Distributed Stochastic Neighbor Embedding (tSNE) diagrams based on the accesson matrix and the motif matrix. (**f**) Fragment counts were specifically enriched in the superenhancer region upstream of N4BP1 in P1-LSCs.

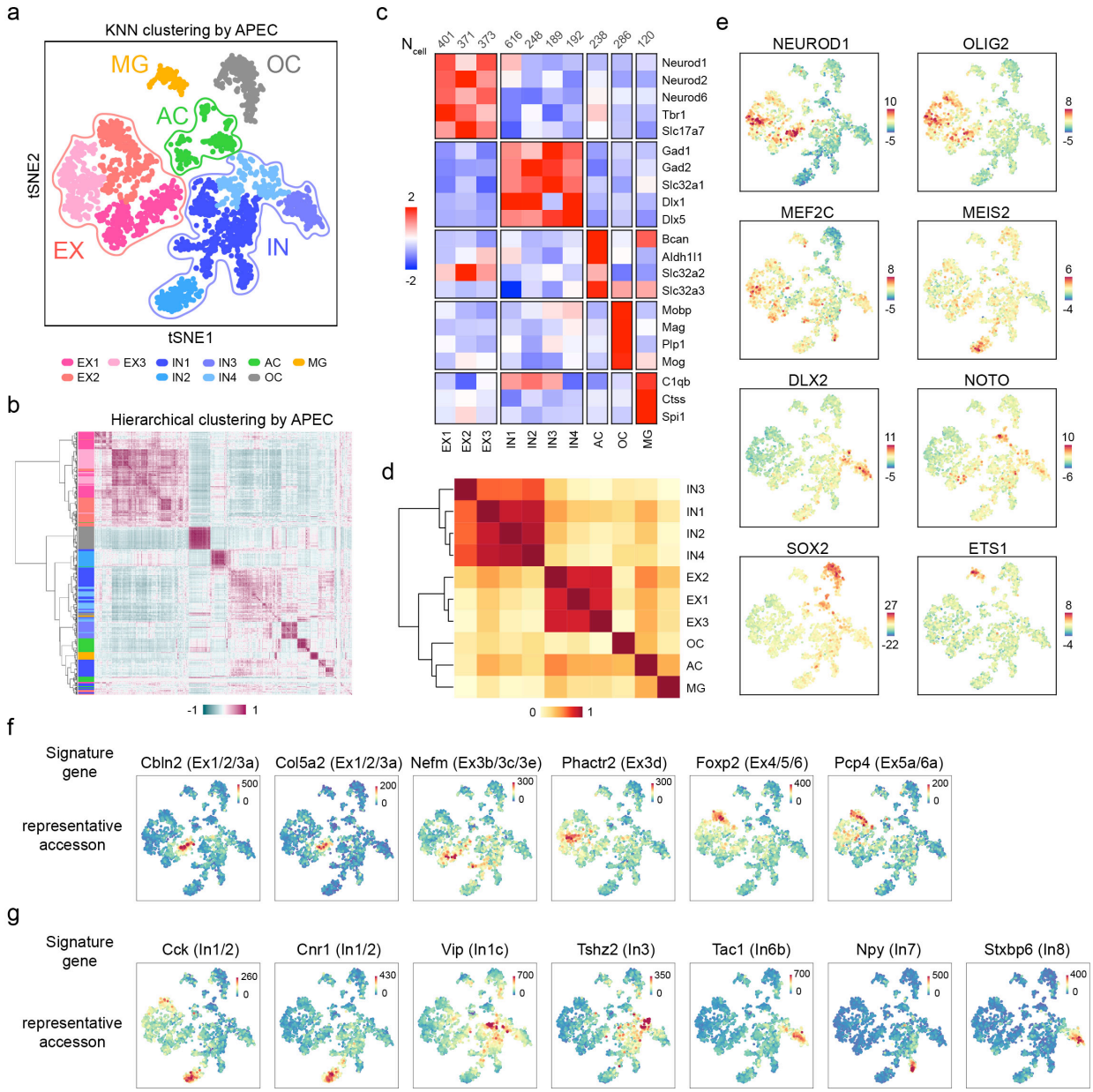

Figure 2. APEC improved the cell type classification of adult mouse forebrain snATAC-seq data. (**a**) A tSNE diagram demonstrates the KNN clustering of forebrain cells. (**b**) Hierarchical clustering of the cell-cell correlation matrix. The side bar denotes cell clusters from the KNN method. (**c**) Average of the marker gene scores for each cell cluster, normalized by the standard score (z-score). The top row lists the cell numbers for each cluster. (**d**) Hierarchical clustering of the cluster-cluster correlation matrix. (**e**) Differential enrichments of cell type-specific motifs in different clusters. (**f, g**) Intensity of representative accessons associated with signature genes of excitatory (Ex) and inhibitory (In) neuron subtypes. The subtypes listed in parentheses were defined by the signature genes in the results from scRNA-seq data[14].

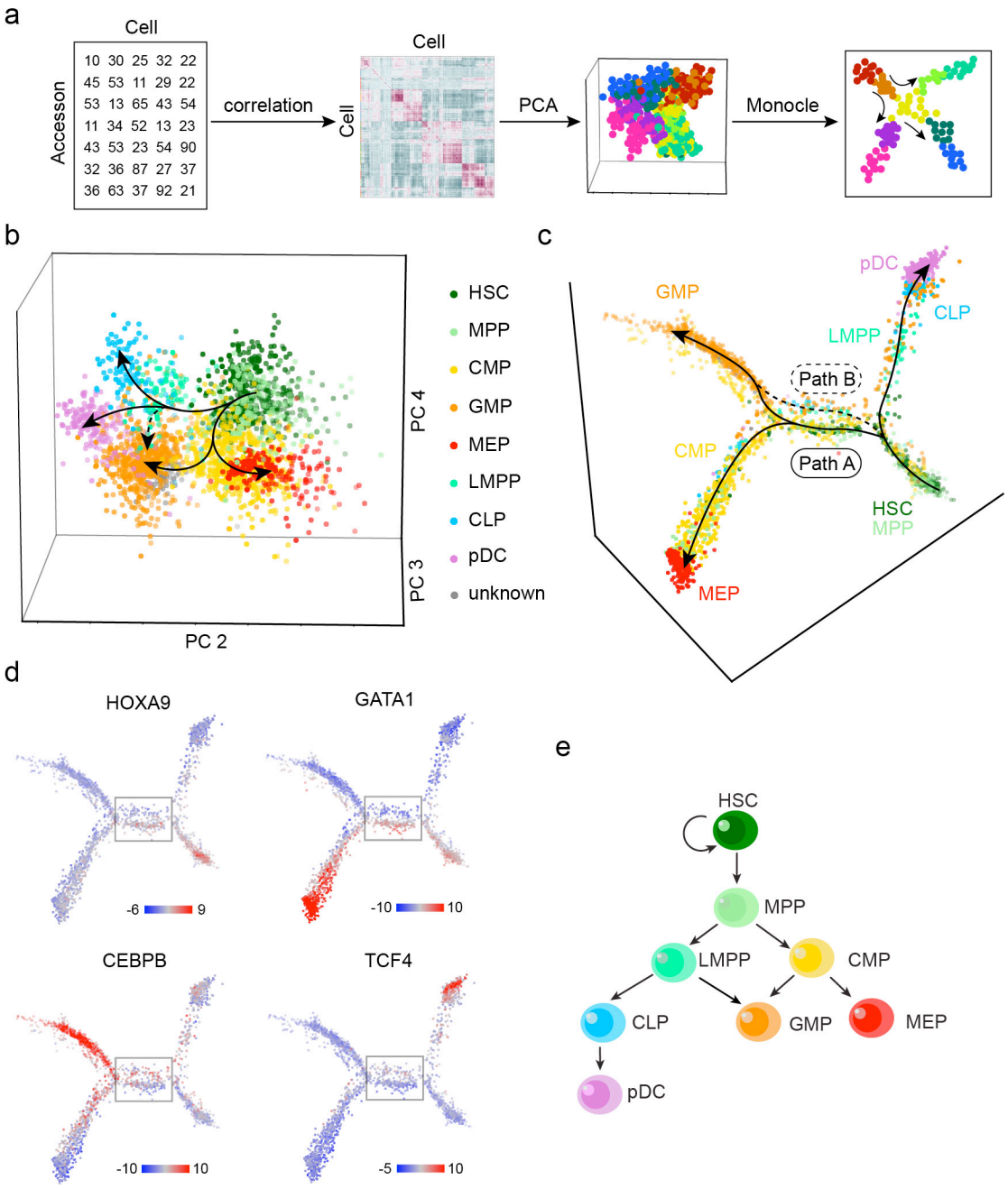

Figure 3. APEC constructed a differentiation pathway from scATAC-seq data from human hematopoietic cells. (**a**) The pseudotime trajectory construction scheme based on the accesson matrix and Monocle. (**b**) Principal component analysis (PCA) of the accesson matrix for human hematopoietic cells. The first principal component is not shown here because it was highly correlated with sequencing depth[38]. HSC, hematopoietic stem cell; MPP, multipotent progenitor; LMPP, lymphoid-primed multipotential progenitor; CMP, common myeloid progenitor; CLP, common lymphoid progenitor; pDC, plasmacytoid dendritic cell; GMP, granulocyte-macrophage

487 progenitor; MEP, megakaryocyte-erythroid progenitor; unknown, unlabeled cells. (**c**) Pseudotime
488 trajectory for the same data constructed by calling Monocle on the accesson matrix. Paths A and
489 B represent different pathways for GMP cell differentiation. (**d**) The deviations of significant
490 differential motifs (HOXA9, GATA1, CEBPB, and TCF4) plotted on the pseudotime trajectory. (**e**)
491 Modified schematic of human hematopoietic differentiation.

492

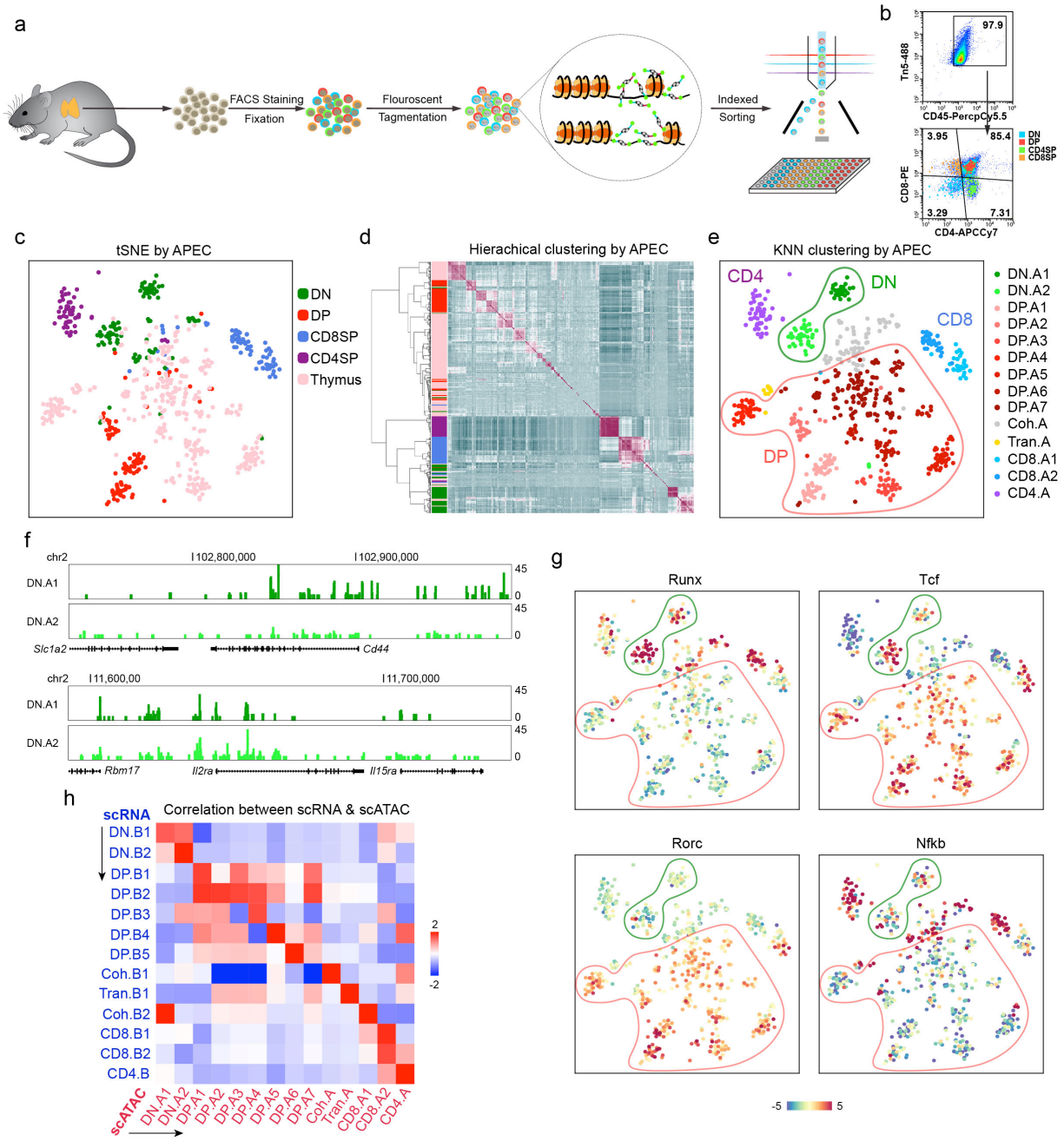

Figure 4. APEC accurately identified cell subtypes based on scATAC-seq data from *Mus musculus* thymocytes. (**a**) Experimental workflow of the fluorescent tagmentation- and FACS-sorting-based scATAC-seq strategy (ftATAC-seq). (**b**) Indexed sorting of double-negative (DN), double-positive (DP), $CD4^+$ single-positive (CD4SP), and $CD8^+$ single-positive (CD8SP) cells with strong tagmentation signals. (**c**) The tSNE of thymocyte single-cell ftATAC-seq data based on the accesson matrix, in which the cells are labeled by the sorting index. (**d**) Hierarchical clustering of the cell-cell correlation matrix. On the sidebar, each cell was colored by the sorting index. (**e**) The

501 accesson-based KNN method clustered thymocytes into 14 subtypes. DN.A1&A2, double-
502 negative clusters; DP.A1~A7, double-positive clusters; Coh.A, coherent state; Tran.A, transition
503 state; CD8.A1&A2, $CD8^+$ single-positive clusters; CD4.A, $CD4^+$ single-positive cluster. (**f**) Average
504 fragment counts of two DN clusters around the marker genes Cd44 and Il2ra. (**g**) Differential
505 enrichment of the motifs Runx, Tcf, Rorc, and Nfkb in the cell clusters. (**h**) Correlation of the cell
506 clusters identified by data from single-cell transcriptome (SMART-seq) and chromatin accessibility
507 (ftATAC-seq) analysis.

508

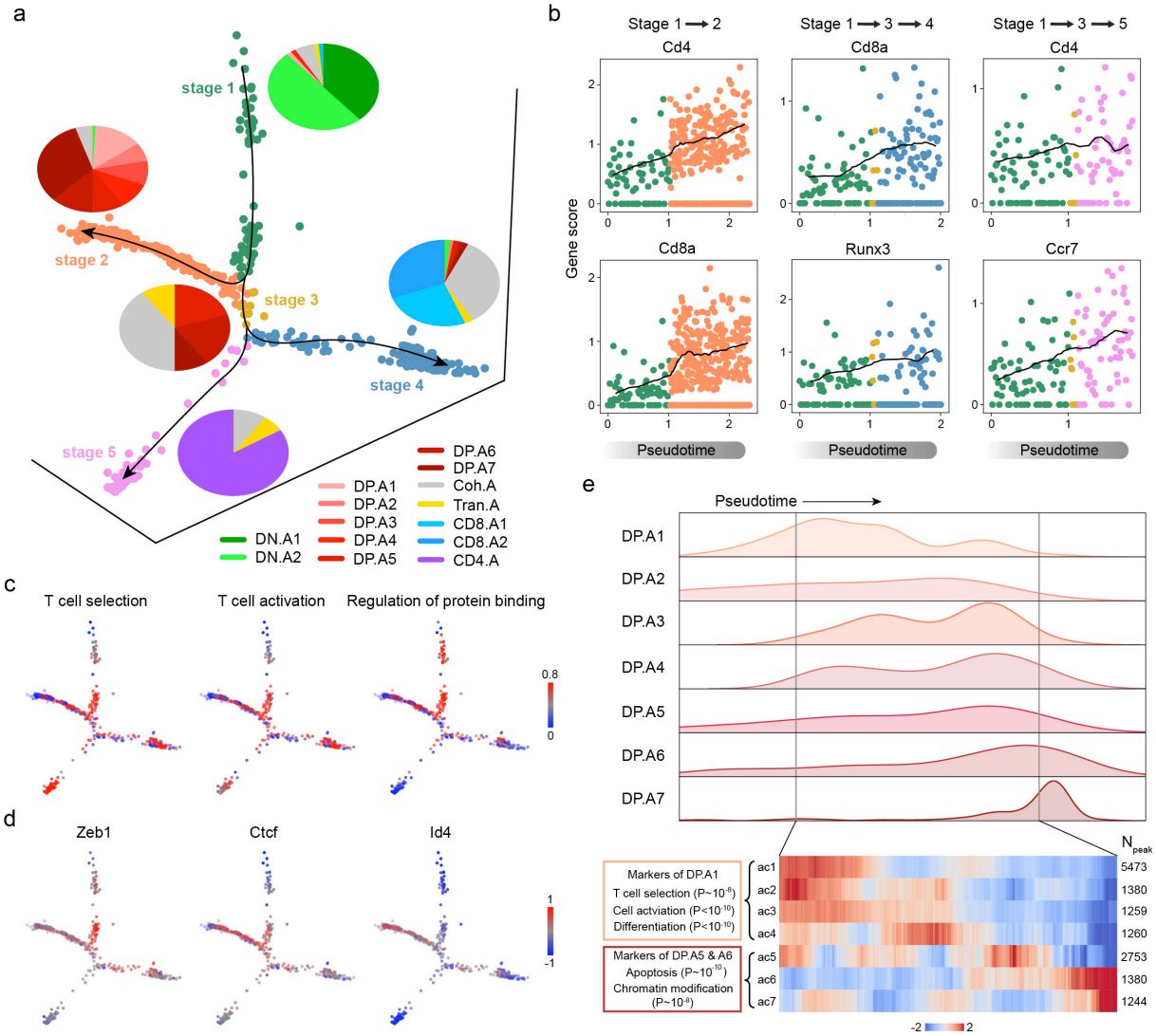

Figure 5. APEC depicted the developmental pathways of *Mus musculus* thymocytes by pseudotime analysis. (**a**) Pseudotime trajectory based on the accesson matrix of thymocyte ftATAC-seq data. Cell colors were defined by the developmental stages along pseudotime. Pie charts show the proportion of cell clusters at each stage. (**b**) Normalized scores of important marker genes (Cd8a, Cd4, Runx3, and Ccr7) along each branch of the pseudotime trajectory. (**c**) Accesson weight scores of important functional GO terms along each branch of the pseudotime trajectory. (**d**) Enrichment of specific motifs searched from the differential accessons of each cell subtype. (**e**) On the stage 2 branch, the cell number distribution of clusters DP.A1~A7 along pseudotime (upper panel) and the intensity of marker accessons of DP.A1 and DP.A5/A6 (lower panel).

521 ## METHODS

522 **Mice.** C57BL/6 mice were purchased from Beijing Vital River Laboratory Animal Technology and
523 maintained under specific pathogen-free conditions until the time of experiments. All mouse
524 experiments in this study were reviewed and approved by the Institutional Animal Care and Use
525 Committee of the University of Science and Technology of China.

526 **ftATAC-seq on mouse thymocytes.** Alexa fluor 488-labeled adaptor oligonucleotides were
527 synthesized at Sangon Biotech as follows: Tn5ME, 5'-[phos]CTGTCTCTTATACACATCT-3';
528 AF488-R1, 5'-AF488- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3'; and AF488-R2, 5'-
529 AF488-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3'. Then, 50 μM of AF488-
530 R1/Tn5ME and AF488-R2/Tn5ME were denatured separately in TE buffer (Qiagen) at 95 °C for
531 5 min and cooled down to 22 °C at 0.1 °C/s. AF488-labeled adaptors were assembled onto Robust
532 Tn5 transposase (Robustnique) according to the user manual to form fluorescent transposomes.

533 Thymus tissues isolated from 6- to 8-week-old male mice were gently ground in 1 mL of RPMI-
534 1640. Thymocytes in a single-cell suspension were counted after passing through a 40 μm nylon
535 mesh. A total of 1×10^6 thymocytes were stained with PerCP-Cy5.5-anti-CD45, PE-anti-CD8a
536 and APC-Cy7-anti-CD4 antibodies (Biolegend) and then fixed in 1× PBS containing 1% methanol
537 at room temperature for 5 min. After washing twice with 1× PBS, the cells were counted again. A
538 total of 1×10^5 fixed cells were resuspended in 40 μL of 1× TD buffer (5 mM Tris-HCl , pH 8.0, 5
539 mM $MgCl_2$, and 10% DMF) containing 0.1% NP-40. Then, 10 μL of fluorescent transposomes
540 were added and mixed gently. Fluorescent tagmentation was conducted at 55 °C for 30 min and
541 stopped by adding 200 μL of 100 mM EDTA directly to the reaction mixture. The cells were loaded
542 on a Sony SH800S sorter, and single cells of the $CD45^+$/AF488-$Tn5^{hi}$ population were index-
543 sorted into each well of 384-well plates. The 384-well plates used to acquire sorted cells were
544 loaded with 2 μL of release buffer (50 mM EDTA, 0.02% SDS) before use. After sorting, the cells
545 in the wells were incubated for 1 min. Plates that were not processed immediately were preserved
546 at -80 °C.

547 To prepare a single-cell ATAC-seq library, plates containing fluorescently tagmented cells were
548 incubated at 55 °C for 30 min. Then, 4.2 μL of PCR round 1 buffer (1 μL of 100 μM $MgCl_2$, 3 μL
549 of 2× I-5 PCR mix [MCLAB], and 0.1 μL each of 10 μM R1 and R2 primers) were added to each
550 well, followed by PCR: 72 °C for 10 min; 98 °C for 3 min; 10 cycles of 98 °C for 10 s, 63 °C for 30
551 s and 72 °C for 1 min; 72 °C for 3 min; and holding at 4 °C. Thereafter, each well received 4 μL
552 of PCR round 2 buffer (2 μL of I-5 PCR Mix, 0.5 μL each of Ad1 and barcoded Ad2 primers, and
553 1 μL of ddH_2O), and final PCR amplification was carried out: 98 °C for 3 min; 12 cycles of 98 °C

554 for 10 s, 63 °C for 30 s and 72 °C for 1 min; 72 °C for 3 min; and holding at 4 °C. Wells containing
555 different Ad2 barcodes were collected together and purified with a QIAquick PCR purification kit
556 (Qiagen). Libraries were sequenced on an Illumina HiSeq X Ten system.

557 **SMART-seq on thymocytes.** Thymocytes were stained and sorted directly into 384-well plates
558 without fixation. SMART-seq was performed as described with some modifications.[58] Reverse
559 transcription and the template-switch reaction were performed at 50 °C for 1 hr with Maxima H
560 Minus Reverse Transcriptase (Thermo Fisher); for library construction, 0.5-1 ng of cDNA was
561 fragmented with 0.05 μL of Robust Tn5 transposome in 20 μL of TD buffer at 55 °C for 10 min,
562 then purified with 0.8× VAHTS DNA Clean Beads (Vazyme Biotech), followed by PCR
563 amplification with Ad1 and barcoded Ad2 primers and purification with 0.6× VAHTS DNA Clean
564 Beads. Libraries were sequenced on an Illumina HiSeq X Ten system.

565 **Data source.** All experimental raw data used in this paper are available online. The single-cell
566 data for mouse thymocytes captured by the ftATAC-seq experiment can be obtained from the
567 Genome Sequence Archive at BIG Data Center with the accession number CRA001267 and is
568 available via http://bigd.big.ac.cn/gsa/s/yp1164Et. Other published data sets used in this study
569 are available from NIH GEO: (1) scATAC-seq data for LSCs and leukemic blast cells from patients
570 SU070 and SU353, LMPP cells, and monocytes from GSE74310[3]; (2) scATAC-seq data for HL-
571 60 cells from GSE65360[13]; and (3) scATAC-seq data for hematopoietic development (HSCs,
572 MPPs, CMPs, LMPPs, GMPs, EMPs, CLPs and pDCs) from GSE96772[38]. APEC is also
573 compatible with a preprocessed fragment count matrix from the snATAC-seq data for the
574 forebrain of adult mice (p56) from GSE100033[15].

575 **Preparing the fragment count matrix from the raw data.** APEC adopted the general mapping,
576 alignment, peak calling and motif searching procedures to process the scATAC-seq data. To trim
577 the adapters in the raw data (in paired-end fastq format files for each single-cell sample), we
578 implanted the python version trimming code from our previous published pipeline (ATAC-pipe)[59].
579 Then, APEC used BOWTIE2 to map the trimmed sequencing data to the corresponding genome
580 index and used PICARD for the sorting, duplicate removal, and fragment length counting of the
581 aligned data. The pipeline called peaks from the merged file of all cells by MACS2, ranked and
582 filtered out the low quality peaks based on the false discovery rate (Q-value). Genomic locations
583 of the peaks were annotated by HOMER, and motifs searched by FIMO. APEC calculates the
584 number of fragments and the percent of reads mapped to the TSS region (±2000 BP) for each
585 cell, and filters out high quality cells for downstream analysis. All required files for the hg19 and
586 mm10 assembly have been integrated into the pipeline. If users want to process data from other

587 species, they can also download corresponding reference files from the UCSC website. By
588 combining existing tools, APEC made it possible to finish all of the above data processing steps
589 by one command line, and generate a fragment count matrix for subsequent cell clustering and
590 differential analysis.

591 **Accesson-based clustering algorithm.** We define accesson as a set of peaks with similar
592 accessibility patterns across all single cells, similar to the definition of gene modules for RNA-seq
593 data. The peaks of a same accesson can be distant from each other on the genome, and
594 sometimes on multiple chromosomes. After preprocessing, a filtered fragment count matrix **M** is
595 obtained, and APEC groups peaks to construct accessons and then performs cell clustering
596 analysis as follows:

597 (1) Normalization of the fragment count matrix. Each matrix element M_{ij} represents the
598 number of raw reads in cell i and peak j, and element M_{ij} was then normalized by the
599 total number of reads in each cell i, as if there are 10,000 reads in each cell.

600 $$M'_{ij} = log_2\left(\frac{M_{ij}\times 10000}{\sum_{j'} M_{ij'}} + 1\right)$$

601 (2) Constructing accessons. The top 40 principal components of the normalized matrix **M'**
602 were used to construct the connectivity matrix ($\mathbf{C_{peak}}$) of peaks by the K-nearest-neighbor
603 (KNN) method. Based on the matrix $\mathbf{C_{peak}}$, all peaks were grouped by agglomerative
604 clustering with the Ward's method, and the sum of one peak group was an accesson. In
605 processing of all datasets in this study, the default number of accessons was set to 600.
606 We recommend using a flexible number of accessons so that you can accumulate enough
607 peaks in one accesson while avoiding incorrect grouping of differential peaks. However,
608 not all accessons were used for cell clustering in the next step. Sparse accessons with 4
609 or fewer peaks were discarded since they will interfere with the clustering accuracy. Only
610 accessons containing 5 or more peaks were retained in the accesson count matrix $\mathbf{M_a}$.
611 Each row of $\mathbf{M_a}$ is an accesson, each column is a cell and the elements of $\mathbf{M_a}$ represents
612 the cumulative read counts of each accesson in each cell. If less than 30% of the
613 accessons contain enough number of peaks, the users may consider to reduce the default
614 accesson number to avoid sparse accessons.

615 (3) Cell clustering. From the accesson matrix $\mathbf{M_a}$, APEC calculated the Pearson correlation
616 between each pair of cells, and then performed both hierarchical and KNN clustering on
617 the correlation matrix to categorize cells into different clusters. The number of cell clusters
618 can be predicted by the Louvain method, or inputted by the users. By default, cell

clustering was performed in the high-dimensional PCA transformed space, but also supports clustering in the tSNE space.

(4) Comparison with other clustering methods. To investigate the accuracy of clusters generated by different algorithms, APEC provides two ways to compare cell clusters: a) The contingency matrix, in which each element represents the number of common cells between two clusters from different methods (e.g., hierarchy and KNN clustering, or accesson-based and motif-based clustering); b) The ARI value, which evaluates the similarity of clustering results from two different algorithms[20]. Moreover, one clustering method can be compared with known cell types in the original single-cell data (such as the FACS index) to confirm the accuracy of the cell type classification algorithm.

Gene score evaluated by peaks around the TSS. To evaluate the accessibility score of one gene, we calculated the average count of all peaks around its TSS (±20000 BP) as its raw score (S_{ij} for cell i and gene j). Then, we obtained the gene accessibility score by normalizing the raw score ($S'_{ij} = S_{ij} * 10000/\sum_i S_{ij}$), which is in a range comparable to the gene expression from scRNA-seq data. The average score of all cells in one cluster represents the accessibility of a cell type ($\bar{S}_{kj}$ for cell cluster k and gene j). We normalized the gene score matrix $\bar{S}$ by caculating z-score for each row and column, and the final matrix $\bar{S}_z$ represents the relative strength of gene accessibility for each cell type.

Significant differential peaks, genes and motifs. APEC used the Student's t-test to estimate the significance of the fragment count differences between cell clusters, with P-value and fold changes, and one can determine the thresholds to identify significant differential peaks for each cluster. The significant differential genes of each cell cluster can also be acquired from the accessibility score ($\bar{S}_{kj}$) by the same method. To accurately quantify the enrichment of motifs on each cell, APEC applied the bias-corrected deviation algorithm from chromVAR[24]; thus, the chromVAR algorithm has been embedded into the pipeline to facilitate the calculation of the corrected deviation of the motifs. In this python version of chromVAR, permuted sampling and background deviation calculation can be run in parallel on multiple processors to reduce the computer time. The differentially enriched motifs were defined by a fold change >1 in the average motif deviation between one cluster and another.

Potential super-enhancers. Here, we defined a super-enhancer as a long continuous genomic area containing many accessible regions and have the same accessibility pattern in different cells. Many different motifs appear in one super-enhancer, therefore, the motif-based clustering method

651 cannot reflect the critical contributions from super-enhancers for cell clustering. However, the
652 accesson-based algorithm can group most peaks in one super-enhancer to one accesson since
653 they always present the same accessibility pattern between cells. APEC identified super-
654 enhancers by counting the number of peaks in a 1 million BP genomic area that belong to a same
655 accesson. It also requires that more than 3/4 of the putative peaks in one super-enhancer be
656 adjacent on the initial peak list. The pipeline can also aggregate bam files by cell types/clusters
657 and convert them to BigWig format for users to upload to the UCSC genome browser for
658 visulization.

659 **Pseudotime trajectory.** As a tool to simulate the time-dependent variation of gene expression
660 and the cell development pathway, Monocle has been widely used for the analysis of single-cell
661 RNA-seq experiments[30, 60]. APEC reduced the dimension of the accesson count matrix $\mathbf{M_a}$ by
662 PCA, and then performed pseudotime analysis using the Monocle program. For complex datasets,
663 it is necessary to limit the number of principal components, since too many features will cause
664 too many branches on the pseudo-time trajectory, and makes it difficult for a user to identify the
665 biological significance of each branch. For the hematopoietic single cell data and thymocyte data,
666 we used the top 5 principal components of the accesson matrix to construct the developmental
667 and differentiation trajectories.

668 **Parameter settings for each analysis.** In the quality control (QC) step, cells are filtered by two
669 constraints: the percentage of the fragments in peaks (P_f) and the total number of valid fragments
670 (N_f). However, there is no fixed cutoff for these two parameters since the quality of different cell
671 types and/or experiment batches are completely different. The total number of peaks is usually
672 limited to approximately 50000 to reduce computer time, but we recommend using all peaks if the
673 users want to obtain better cell clusters. (1) For the scATAC-seq data from leukemic cells (P1/P2
674 LSCs and blast cells, LMPPs, HL60 cells, and monocytes), the threshold of -log(Q-value) was set
675 to 8 to retain 42139 high-quality peaks for subsequent processing. In the QC step, we set the P_f
676 cutoff to 0.05 and the N_f cutoff to 800. (2) For the snATAC-seq data from the adult mouse forebrain,
677 all peaks and the raw count matrix obtained from the original data source were adopted in the
678 analysis. (3) For the data set from hematopoietic cells, the -log(Q-value) threshold of high-quality
679 peaks was set to 35 to retain 54212 peaks, and the cutoff values of P_f and N_f were 0.1 and 1000,
680 respectively. (4) For the ftATAC-seq data from thymocytes, all 130685 peaks called by MACS2
681 were reserved for the fragment count matrix (Q-value<0.05), and we retained cells with P_f >0.2
682 and N_f >2000.

683 **SMART-seq data analysis with Seurat.** For the analysis of SMART-seq data from mouse
684 thymocytes, we employed STAR (version 2.5.2a) with the ratio of mismatches to mapped length
685 (outFilterMismatchNoverLmax) less than or equal to 0.05, translated output alignments into
686 transcript coordinates (i.e., quantMode TranscriptomeSAM) for mapping[61] (Dobin et al., 2013) and
687 used RSEM[62] (Bo et al., 2011) to calculate the TPM of genes. For QC, we excluded cells in which
688 fewer than 2000 genes were detected and genes that were expressed in only 3 or fewer cells.
689 Seurat filtered cells with several specific parameters to limit the number of genes detected in each
690 cell to 2000~6000 and the proportion of mitochondrial genes in each cell was set to less than 0.4
691 (i.e., low.thresholds=c(2000,-Inf), high.thresholds=c(6000,0.4)). Additionally, the top 12 principal
692 components were used for dimension reduction with a resolution of 3.2 (dims.use =1:12,
693 resolution=3.2), followed by cell clustering and differential expressed gene analysis[63].

694 **Association of cell clusters from scATAC-seq and scRNA-seq data.** To determine the
695 association between cell clusters from epigenomics and transcriptome sequencing, we calculated
696 the P-values of Fisher's exact test of marker/nonmarker genes between each pair of cell clusters
697 from scATAC-seq and scRNA-seq data. For example, for cell cluster i from ftATAC-seq and cell
698 cluster j from SMART-seq, if the number of consensus marker genes in both cluster i and j is G_{11},
699 the number of genes that are not markers in either cluster i or j is G_{22}, and the number of markers
700 in either cluster i (or cluster j) is G_{12} (or G_{21}), then the 2 by 2 matrix **G** can be directly used for
701 Fisher's exact test to evaluate the P-valuie A_{ij} between cluster i and j. We calculated the logarithm
702 of matrix **A** to obtain matrix **A'**, then calculated the z-score for each row and column of **A'** to
703 determine the correlation of cell clusters from different experiments.

704 **Biological function of accesson.** We defined the functional characteristics of each accesson by
705 the GO terms and motifs enriched on its peaks. The GO terms of an accesson were obtained by
706 submitting all of its peaks to the GREAT website[64]. The logarithm of the P-value of each GO term
707 in each accesson was filled into a (GO terms) × (accessons) matrix **L**. The significance of each
708 GO term on each cell was evaluated by the product of the matrix **L** and the accesson reads count
709 matrix $\mathbf{M_a}$. Then we calculated the z-score for each row of this product matrix, and plotted the z-
710 score as the GO-term score on the trajectory diagram. To assess the motif enrichment of the
711 accessons, we used the Centrimo tool of MEME suite[65] to search for the enriched motifs for the
712 peaks of each accesson, and applied the same algorithm as the GO term score to obtain the motif
713 score.

714

715 ## REFERENCES

716 58. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181
717 (2014).
718 59. Zuo, Z. et al. ATAC-pipe: general analysis of genome-wide chromatin accessibility. *Briefings in*
719 *Bioinformatics*, bby056-bby056 (2018).
720 60. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat Methods*
721 **14**, 309-315 (2017).
722 61. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
723 62. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without
724 a reference genome. *Bmc Bioinformatics* **12**, 323 (2011).
725 63. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using
726 Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
727 64. McLean, C.Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat*
728 *Biotechnol* **28**, 495-501 (2010).
729 65. Bailey, T.L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-
730 208 (2009).

731

732