UNIVERSITY OF CALIFORNIA SAN DIEGO

Single Cell Analysis of Chromatin Accessibility

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Rongxin Fang

Committee in charge:

Professor Bing Ren, Chair
Professor Vineet Bafna, Co-Chair
Professor Joseph R. Ecker
Professor Christopher K. Glass
Professor Eran A. Mukamel
Professor Kun Zhang

2019

The Dissertation of Rongxin Fang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2019

DEDICATION

I would like to dedicate this thesis to my family, especially my mom and dad who supported my scientific endeavors unconditionally throughout my life. And to my girlfriend Xinzhu (Xinxin), who has always sparked me joy.

TABLE OF CONTENTS

SIGNATURE PAGE..... iii

DEDICATION.....iv

TABLE OF CONTENTS v

LIST OF FIGURES..... viii

ACKNOWLEDGMENTS..... xiii

VITA xvi

ABSTRACT OF THE DISSERTATION..... xviii

INTRODUCTION 1

References 7

CHAPTER 1: SINGLE-NUCLEUS ANALYSIS OF ACCESSBILE CHROMATIN IN DEVELOPING MOUSE FOREBRAIN 12

1.1 Abstract..... 12

1.2 Introduction..... 13

1.3 Results..... 14

1.4 Discussion 25

1.5 Acknowledgments 27

1.6 Author Contributions 27

1.7 Figures..... 29

1.8 Supplementary Methods 34

1.9 Supplementary Figures..... 48

1.11 References 66

CHAPTER 2: COMPREHENSIVE ANALYSIS OF SINGLE CELL ATAC-SEQ DATA .. 74

2.1 Abstract..... 74

2.2 Introduction..... 75

2.3 Results 79
2.4 Discussion 93
2.5 Acknowledgments 96
2.6 Author Contributions 96
2.7 Figures 97
2.8 Supplementary Methods 105
2.9 Supplementary Figures 134
2.10 References 174
CHAPTER 3: MAPPING OF LONG-RANGE CHROMATIN INTERACTIONS BY PROXIMITY LIGATION-BASED CHIP-SEQ 178
3.1 Abstract 178
3.2 Introduction 179
3.3 Results 180
3.4 Acknowledgments 184
3.5 Author Contributions 184
3.6 Figures 185
3.7 Supplementary Methods 187
3.8 Supplementary Figures 191
3.9 References 195
CHAPTER 4: A TILING-DELETION BASED GENETIC SCREEN FOR CIS-REGULATORY ELEMENT IDENTIFICATION IN MAMMALIAM CELLS 197
4.1 Abstract 197
4.2 Introduction 198
4.3 Results 200
4.4 Discussion 209
4.5 Acknowledgements 211
4.6 Author Contributions 211
4.7 Figures 212

4.8 Supplementary Methods 217
4.9 Supplementary Figures 229
4.10 References 246

LIST OF FIGURES

Figure 1.1. Overview of the experimental and computational procedures of snATAC-seq 29

Figure 1.2. Deconvolution of cell types in the p56 mouse forebrain and identification of potential master regulators of each cell type 30

Figure 1.3. SnATAC-seq analysis reveals the timing of neurogenesis and gliogenesis during embryonic forebrain development 32

Figure 1.4. SnATAC-seq analysis uncovers cis regulatory elements and transcriptional regulators of lineage specification in the developing forebrain 33

Figure S1.1. SnATAC-seq protocol optimization 48

Figure S1.2. Isolation of single nuclei after tagmentation 50

Figure S1.3. Overview of snATAC-seq sequencing data and quality filtering for single nuclei 51

Figure S1.4. SnATAC-seq data sets are robust and reproducible 53

Figure S1.5. Clustering strategies, quality control of clusters and clustering result for individual experiments in adult forebrain 55

Figure S1.6. Ranking of gene loci (TSS ± 10kb) compared to other clusters in adult forebrain 57

Figure S1.7: Flow cytometric analysis of adult mouse forebrain and comparison to single cell RNA-seq data from different brain regions 58

Figure S1.8. Sub-classification of excitatory neurons into hippocampal and cortical neuron types 59

Figure S1.9. Cell-type specificity and coverage of the cis elements 60

Figure S1.10. Distinct chromatin accessibility profiles of two GABAergic neuron clusters 61

Figure S1.11. Comparison of chromatin accessibility and differentially methylated regions in neuronal subtypes 62

Figure S1.12. Dynamics of chromatin accessibility within distinct cell groups 63

Figure S1.13. Distal genomic element clusters are associated with distinct anatomical locations in the developing forebrain 65

Figure 2.1. Schematic overview of SnapATAC analysis workflow..... 97

Figure 2.2. SnapATAC links distal regulatory elements to putative target genes 98

Figure 2.3. SnapATAC constructs cellular trajectories for the developing mouse brain 100

Figure 2.4. SnapATAC outperforms current methods in accuracy, sensitivity, scalability and stability of identifying cell types in complex tissues..... 101

Figure 2.5. A high-resolution cis-regulatory atlas of mouse secondary motor cortex (MOs) 102

Figure 2.6. SnapATAC enables supervised annotation of new scATAC-seq dataset using reference cell atlas 104

Figure S2.1. Overview of SnapTools workflow..... 134

Figure S2.2. SnapATAC removes putative doublets using Scrublet 135

Figure S2.3. Choosing the optimal bin size 136

Figure S2.4. SnapATAC is robust to sequencing depth..... 138

Figure S2.5. SnapATAC is robust to other biases..... 139

Figure S2.6. Nyström sampling improves the scalability without sacrificing the performance..... 140

Figure S2.7. SnapATAC delineates cellular heterogeneity in published large-scale scATAC-seq datasets..... 142

Figure S2.8. SnapATAC predicts gene and enhancer pairing by integrating scATAC-seq and scRNA-seq..... 144

Figure S2.9. SnapATAC constructs cellular trajectories for the developing mouse brain 145

Figure S2.10. Evaluation of clustering accuracy of SnapATAC relative to alternative methods on simulated datasets 146

Figure S2.11. Evaluation of clustering accuracy on published single cell ATAC-seq datasets 148

Figure S2.12. Gene accessibility score of canonical marker genes projected onto t-SNE embedding of mouse secondary motor cortex (MOs-M1) snATAC-seq dataset to guide the cluster annotation 149

Figure S2.13. Evaluation of clustering sensitivity on in-house mouse secondary motor cortex dataset 151

Figure S2.14. Gene accessibility score of canonical marker genes projected onto t-SNE embedding for a 10X scATAC-seq dataset of the mouse brain to guide the cluster annotation 152

Figure S2.15. Evaluation of clustering sensitivity on a 10X scATAC-seq dataset from the Mouse Brain..... 154

Figure S2.16. Gene accessibility score of canonical marker genes projected onto the t-SNE embedding from 5K PBMC 10X dataset to guide the annotation of the clusters. 155

Figure S2.17. Evaluation of clustering sensitivity on a 5K PBMC 10X dataset..... 156

Figure S2.18. Off-peak reads can be used to distinguish different cell types 157

Figure S2.19. Off-peak reads reflect higher-order chromatin structure..... 158

Figure S2.20. SnapATAC is robust to technical variation 159

Figure S2.21. SnapATAC eliminates batch effect using Harmony 160

Figure S2.22. Single nucleus ATAC-seq datasets are reproducible between biological replicates..... 161

Figure S2.23. Barcode selection of MOs 163

Figure S2.24. Consensus clustering of MOs 164

Figure S2.25. MOs clustering result is reproducible between biological replicates 165

Figure S2.26. Gene accessibility score of canonical marker genes projected onto MOs t-SNE embedding to guide the cluster annotation 166

Figure S2.27. Iterative clustering identifies 17 GABAergic neuronal subtypes 167

Figure S2.28. Gene accessibility score of marker genes projected onto t-SNE embedding from GABAergic neurons to guide the cluster annotation 168

Figure S2.29. SnapATAC uncovers novel candidate cis-regulatory elements in rare cell types 170

Figure S2.30. Joint diffusion maps embedding for query (Mouse Brain 10X) and reference dataset (MOs snATAC) 171

Figure S2.31. SnapATAC is robust for supervised annotation of datasets containing cell types missing in the reference atlas 172

Figure S2.32. Iterative clustering does not substantially improve the clustering sensitivity 173

Figure 3.1. PLAC-seq reveals chromatin interactions in mammalian cells at high sensitivity and accuracy 185

Figure S3.1. Development and validation of PLAC-seq 191

Figure S3.2. Comparison of chromatin interactions detected by 4C-seq, PLAC-seq, and ChIA-PET at three genomic loci 193

Figure 4.1. CREST-seq experimental design and application to the POU5F1 locus in hESC 212

Figure 4.2. CREs tend to be associated with canonical active chromatin markers of cis-regulatory elements and dense TF clusters 214

Figure 4.3. The core promoter regions of MSH5, NEU1, and PRRC2A are required for optimal POU5F1 expression in hESC 215

Figure 4.4. Analysis of chromatin interactions between the enhancer-like promoters and POU5F1 promoter in hESC 216

Figure S4.1. Design of sgRNA pairs 229

Figure S4.2 CREST-seq library construction and quality control 230

Figure S4.3. Quality control of CREST-seq data from replicates 231

Figure S4.4. CREST-seq identifies the promoter and known enhancers of POU5F1 . 232

Figure S4.5. Chromatin features enriched on CREs 233

Figure S4.6. Genotype information for the mutant clones with genomic deletion on selected CREs 234

Figure S4.7. Genotype information for core promoter mutant clones 236

Figure S4.8. Characterization and quantification of eGFP levels in multiple core promoter deletion mutant clones 238

Figure S4.9. Quantification of POU5F1, MSH5, NEU1 and PRRC2A expression in various samples 240

Figure S4.10. The reduced eGFP expression in bi-allelic or P1 allelic specific mutants is not due to DSB induced transcription repression 241

Figure S4.11. Promoter-CREs are associated with active gene expression 242

Figure S4.12. List of features that distinguish POU5F1 regulatory promoters from other non-POU5F1-regulatory promoters 243

Figure S4.13. Analysis of Cis- and Trans-regulatory elements with dual sgRNA tiling deletion screen 244

Figure S4.14. The eGFP levels correlate with P1 allele specific POU5F1 expression 245

ACKNOWLEDGMENTS

First, I am deeply thankful to my PhD advisor, Prof. Bing Ren, for his support and guidance during my Ph.D. Bing has not only welcomed me to the lab but also taught me how to be a rigorous, creative and ambitious scientist. Second, I would like to express my sincere gratitude to my committee members Prof. Joseph R. Ecker, Prof. Vineet Bafna, Prof. Kun Zhang, Prof. Christopher K. Glass and Prof. Eran A. Mukamel for their insightful advice for my projects and continuous support for my career development. Next, I would also like to thank all the members of the Ren lab, from whom I always have unlimited support. Last but not the least, I would like to thank my family for unconditional support throughout my life.

The Introduction is, in part, based on the material currently being prepared for submission as “Comprehensive Analysis of Single Cell ATAC-seq Data”. Rongxin Fang, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Kai Zhang, Fangming Xie, Eran A. Mukamel, Yanxiao Zhang, M. Margarita Behrens, Joseph Ecker, and Bing Ren. The Introduction is also, in part, based on material as it appears as "Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation" in *Nature Neurosciences*, 2018. Sebastian Preissl, Rongxin Fang, Hui Huang, Yuan Zhao, Ramya Raviram, David U Gorkin, Yanxiao Zhang, Brandon C Sos, Veena Afzal, Diane E Dickel, Samantha Kuan, Axel Visel, Len A Pennacchio, Kun Zhang, Bing Ren. The Introduction is also, in part, based on the material as it may appear in *Cell Research*, 2017. “Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq”. Rongxin

Fang, Miao Yu, Guoqiang Li, Sora Chee, Tristin Liu, Anthony D Schmitt and Bing Ren. The Introduction is also, in part, based on the material as it may appear in *Nature Methods*, 2016. “A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells”. Yarui Diao, Rongxin Fang, Bin Li, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C Lin, Hui Huang, Tristin Liu, Ryan J Marina, Inkyung Jung, Yin Shen, Kun-Liang Guan and Bing Ren. The dissertation author was the primary investigator and author of these papers.

Chapter 1, in full, is a reformatted reprint of the material as it appears “Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation” in *Nature Neuroscience,* 2018. Sebastian Preissl, Rongxin Fang, Hui Huang, Yuan Zhao, Ramya Raviram, David U Gorkin, Yanxiao Zhang, Brandon C Sos, Veena Afzal, Diane E Dickel, Samantha Kuan, Axel Visel, Len A Pennacchio, Kun Zhang, Bing Ren. The dissertation author was a co-primary investigator and author of this paper.

Chapter 2, in full, is currently being prepared for submission of the material as “Comprehensive Analysis of Single Cell ATAC-seq Data”. Rongxin Fang, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Kai Zhang, Fangming Xie, Eran A. Mukamel, Yanxiao Zhang, M. Margarita Behrens, Joseph Ecker, and Bing Ren. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reformatted reprint of the material as it appears as "Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq" in *Cell Research*, 2016. Rongxin Fang, Miao Yu, Guoqiang Li, Sora Chee, Tristin Liu, Anthony D Schmitt & Bing Ren. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reformatted reprint of the material as it appears as " A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells" in *Nature Methods*, 2017. Yarui Diao, Rongxin Fang, Bin Li, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C Lin, Hui Huang, Tristin Liu, Ryan J Marina, Inkyung Jung, Yin Shen, Kun-Liang Guan & Bing Ren. The dissertation author was the primary investigator and author of this paper.

VITA

2012 Bachelor of Engineering, Yantai University

2013 Research Assistant, Chinese Academy of Sciences

2014 Research Assistant, Duke University

2019 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Rongxin Fang, Sebastian Preissl, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Eran A. Mukamel, Yanxiao Zhang, M. Margarita Behrens, Joseph Ecker, Bing Ren. "Comprehensive Analysis of Single Cell ATAC-seq Data" (in preparation).

Sebastian Preissl*, **Rongxin Fang***, Hui Huang, Yuan Zhao, Ramya Raviram, David U. Gorkin, Yanxiao Zhang, Brandon C. Sos, Veena Afzal, Diane E. Dickel, Samantha Kuan, Axel Visel, Len A. Pennacchio, Kun Zhang & Bing Ren. "Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation". *Nature Neurosciences*. 21(3):432-439 (2018).

Yarui Diao*, **Rongxin Fang***, Bin LI*, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C Lin, Hui Huang, Tristin Liu, Ryan J Marina, Inkyung Jung, Yin Shen, Kun-Liang Guan & Bing Ren. "A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells". *Nature Methods*. 14(6):629-635 (2017).

Rongxin Fang*, Miao Yu*, Guoqiang Li, Sora Chee, Tristin Liu, Anthony D Schmitt & Bing Ren. "Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq". *Cell Research*. 26(12):1345-1348 (2016).

Rongxin Fang*, Chengqi Wang*, Geir Skogerbo and Zhihua Zhang. "Functional diversity of CTCF is encoded in binding motifs". *BMC Genomics*. 28;16:649 (2015).

Guoqiang Li*, Yaping Liu*, Yanxiao Zhang, Naoki Kubo, Miao Yu, **Rongxin Fang**, Manolis Kellis, Bing Ren. "Simultaneous profiling of DNA methylation and chromatin architecture in mixed populations and in single cells". *Nature Methods* (2019).

Yanxiao Zhang*, Ting Li*, Sebastian Preissl*, Jonathan Grinstein, Elie Farah, Eugin Destici, Ah Young Lee, Sora Chee, Yunjiang Qiu, Kaiyue Ma, Zhen Ye, Quan Zhu, Hui Huang, Rong Hu, **Rongxin Fang**, Leqian Yu, Juan Carlos Belmonte, Jun Wu, Sylvia Evans, Neil Chi, Bing Ren. "Transcriptionally active HERV-H retrotransposons demarcate

topologically associating domains in human pluripotent stem cells". *Nature Genetics*. 51(9):1380-1388 (2019).

Ivan Juric*, Miao Yu*, Armen Abnousi*, Ramya Raviram, **Rongxin Fang**, Yuan Zhao, Yanxiao Zhang, Yuchen Yang, Yun Li, Bing Ren, Ming Hu. "MAPS: model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments". *PLOS Computational Biology*. 15;15(4):e1006982 (2019).

Qingfei Jiang, Jane Isquith, Maria Anna Zipeto, Raymond H Diep, Jessica Pham, Nathan Delos Santos, Eduardo Reynoso, Julisia Chau, Heather Leu, Elisa Lazzari, Etienne Melese, Wenxue Ma, **Rongxin Fang**, Sheldon Morris, Bing Ren, Gabriel Pineda, Frida Holm, Catriona Jamieson. "Hyper-Editing of Cell-Cycle Regulatory and Tumor Suppressor RNA Promotes Malignant Progenitor Propagation". *Cancer Cell*. 35(1):81-94.e7 (2018).

Verena M Link*, Sascha H Duttke*, Hyun B Chun*, Inge R Holtman, Emma Westin, Marten A Hoeksema, Yohei Abe, Dylan Skola, Casey E Romanoski, Jenhan Tao, Gregory J Fonseca, Ty D Troutman, Nathanael J Spann, Tobias Strid, Mashito Sakai, Miao Yu, Rong Hu, **Rongxin Fang**, Dirk Metzler, Bing Ren, Christopher K Glass. "Analysis of Genetically Diverse Macrophages Reveals Local and Domain-wide Mechanisms that Control Transcription Factor Binding and Function". *Cell*. 173(7):1796-1809.e17 (2018).

Yupeng He, Manoj Hariharan, David U Gorkin, Diane E Dickel, Chongyuan Luo, Rosa G Castanon, Joseph R Nery, Ah Young Lee, Brian A Williams, Diane Trout, Henry Amrhein, **Rongxin Fang**, Huaming Chen, Bin Li, Axel Visel, Len A Pennacchio, Bing Ren, Joseph R Ecker. "Spatiotemporal DNA methylome dynamics of the developing mammalian fetus". *BioRxiv* (2019).

Joshua Chiou*, Chun Zeng*, Zhang Cheng, Jee Yun Han, Michael Schlichting, Serina Huang, Jinzhao Wang, Yinghui Sui, Allison Deogaygay, Mei-Lin Okino, Yunjiang Qiu, Ying Sun, Parul Kudtarkar, **Rongxin Fang**, Sebastian Preissl, Maike Sander, David Gorkin, Kyle J Gaulton. "Single cell chromatin accessibility reveals pancreatic islet cell type- and state-specific regulatory programs of diabetes risk". *BioRxiv* (2019).

*** co-first authors**

ABSTRACT OF THE DISSERTATION

Single Cell Analysis of Chromatin Accessibility

by

Rongxin Fang

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2019

Professor Bing Ren, Chair
Professor Vineet Bafna, Co-Chair

The identity of each cell in the human body is established and maintained through distinct gene expression program, which is regulated in part by the chromatin accessibility. Until recently, our understanding of chromatin accessibility has depended

largely upon bulk measurements in populations of cells. Recent advances in the sequencing techniques have allowed for the identification of open chromatin regions in single cells. During my Ph.D., I have developed and used single cell sequencing techniques to study the diverse gene regulatory programs underlie the different cell types in mammalian complex tissues. In chapter 1, colleague and I developed Single Nucleus Assay of Transpose Accessible Chromatin using Sequencing (snATAC-seq), a combinatorial barcoding-assisted single-cell assay for probing accessible chromatin in single cells. We then used snATAC-seq to generate an epigenomic atlas of early developing mouse brain. The high-level noise of each single cell chromatin accessibility profile and the large volume of the datasets pose unique computational challenges. In chapter 2, I developed a comprehensive bioinformatics software package called SnapATAC for analyzing large-scale single cell ATAC-seq dataset. SnapATAC resolves the heterogeneity in complex tissues and maps the trajectories of cellular states. As a demonstration of its utility, SnapATAC was applied to 55,592 single-nucleus ATAC-seq profiles from the mouse secondary motor cortex. To further determine the target genes of the distal regulatory elements identified using snATAC-seq in different cell types, in chapter 3, colleague and I developed PLAC-seq, a cost-efficient method that identifies the long-range chromatin interaction at kilobase resolution. PLAC-seq improves the efficiency of detecting chromatin conformation by over 10-fold and reduces the input requirement by nearly 100-fold compared to the prior techniques. Finally, to probe the *in vivo* function of the regulatory sequences, I present a high-throughput CRISPR screening method (CREST-seq) for the unbiased discovery and functional assessment of enhancer sequences in the human genome. We used it to interrogate the 2-Mb *POU5F1* locus in

human embryonic stem cells and discovered that sequences previously annotated as promoters of functionally unrelated genes can regulate the expression of *POU5F1* from a long distance. We anticipate that these studies will help us understand the gene regulatory programs across diverse biological systems ranging from human disease to the evolution of species.

INTRODUCTION

Nearly two decades have passed since the human genome was first completely sequenced[1,2], yet the function of its roughly 3 billion nucleotides is still largely unknown. Decoding the human genome, especially the non-protein coding portion that harbors most of the sequence variants underlying the common human diseases, requires the knowledge of the promoters, enhancers, insulators and other regulatory elements[3]. Therefore, comprehensive mapping of the *cis*-regulatory sequences across diverse tissues and cell types in the human body is critical to understand the role of gene regulation in cell function and in human disease.

Since the *cis*-regulatory sequences are often marked by hypersensitivity to nucleases or transposases when they are active or poised to act, approaches to detect DNA accessibility, such as ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing)[4] and DNase-seq (DNase I hypersensitive sites sequencing)[5] have been widely used to map the candidate *cis*-regulatory sequences. However, conventional assays that use bulk tissue samples as input cannot resolve cell type specific usage of *cis* elements and lacks the resolution to study the temporal dynamics. To overcome this challenge, several single cell sequencing techniques have been developed to profile the chromatin accessibility in single cells. For instance, one approach relies on isolation of cell using microfluidic devices (Fluidigm, C1)[6]. Another type of approach involves combinatorial indexing to simultaneously analyze tens of thousands of cells[7]. However, to make these single cell analyses more widely applicable, it is necessary to optimize them for primary tissues.

In Chapter 1, colleague and I show that it is possible to isolate single nuclei from frozen tissues and assay chromatin accessibility in these nuclei in a massively parallel manner. We further apply this technique on the mouse forebrain through eight developmental stages, creating the first single cell epigenomic atlas of developing mouse brain.

Despite the recent advances in single cell ATAC-seq techniques, the exceeding sparsity of signals in each individual profile due to low detection efficiency (5-15% of peaks detected per cell)[7] and the growing volumes of the datasets present a unique computational challenge. To address this challenge, a number of unsupervised algorithms have been developed. For instance, one approach, chromVAR[8], groups similar cells together by dissecting the variability of transcription factor (TF) motif occurrence in the open chromatin regions in each cell. Another type of approach employs the natural language processing techniques such as Latent Semantic Analysis (LSA)[9] and Latent Dirichlet Allocation (LDA)[10] to group cells together based on the similarity of chromatin accessibility. A third approach analyzes the variability of chromatin accessibility in cells based on the k-mer composition of the sequencing reads from each cell[11,12]. A fourth approach, Cicero[13], infers cell-to-cell similarities based on the gene activity scores predicted from their putative regulatory elements in each cell.

However, several limitations still apply to these methods. First, the current analysis methods often require performing dimensionality reduction such as principle component

analysis (PCA) or singular value decomposition (SVD) on a cell matrix of hundreds of thousands of dimensions, scaling the analysis to millions of cells remains very challenging or nearly impossible. Second, the unsupervised identification of cell types or states in complex tissues using scATAC-seq dataset does not match the power of scRNA-seq[14]. One possibility is that the current methods rely on the use of pre-defined accessibility peaks based on the aggregate signals that potentially introduces bias to the cell type identification.

In Chapter 2, I will introduce a software package called Single Nucleus Analysis Pipeline for ATAC-seq (SnapATAC). Unlike previous methods, SnapATAC does not require population-level peak annotation prior to clustering. Instead, it resolves cellular heterogeneity by directly comparing the genome-wide accessibility profiles between cells with the use of the diffusion maps algorithm[15,16], which is highly robust to noise and perturbation. Furthermore, with the use of a sampling technique, Nyström method[17,17,18], SnapATAC improves the computational efficiency and enables the analysis of scATAC-seq from a million cells on regular hardware. Additionally, SnapATAC provides a collection of frequently used features, including integration of scATAC-seq and scRNA-seq dataset, prediction of enhancer-promoter interaction, discovery of key transcription factors, identification of differentially accessible elements, construction of trajectories during cellular differentiation, correction of batch effect and classification of new dataset based on existing cell atlas. Through extensive benchmarking using both simulated and empirical datasets from diverse tissues and species, we show that SnapATAC substantially outperforms its counterparts in accuracy, sensitivity, scalability and

reproducibility for cell type identification from complex tissues. Furthermore, we demonstrate the utility of SnapATAC by building a high-resolution single cell atlas of the mouse secondary motor cortex. This atlas comprises of ~370,000 candidate *cis*-regulatory elements in 31 distinct cell types, including rare neuronal cell types that account for less than 0.1% of the total population analyzed. Through motif enrichment analysis, we further infer potential key transcriptional regulators that control cell type specific gene expression programs in the mouse brain.

Formation of long-range chromatin loops is a crucial step in transcriptional activation of target genes by distal enhancers[19]. Mapping such structural features can help define target genes for enhancers and annotate non-coding sequence variants linked to human diseases[19–21]. Study of the higher-order chromatin organization has been facilitated by the development of chromosome conformation capture (3C)-based technologies[22,23]. Among the commonly used high-throughput 3C approaches are Hi-C[24] and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)[25]. Global analysis of long-range chromatin interactions using Hi-C has been achieved at kilobase resolution but requires billions of sequencing reads[26]. High-resolution analysis of long-range chromatin interactions at selected genomic regions can be attained cost-effectively through ChIA-PET[25,27]. However, ChIA-PET requires hundreds of million cells as starting materials, limiting its application to biological problems with limited materials.

In chapter 3, college and I developed Proximity Ligation-Assisted ChIP-seq (PLAC-seq) to reduce the amount of input materials and to improve the sensitivity and robustness

of the assay. Unlike ChIA-PET, PLAC-seq conducts proximity ligation in nuclei prior to chromatin shearing and immunoprecipitation. As a result, we demonstrated that, compared to ChIA-PET, PLAC-seq greatly improves the efficiency of detecting the long-range chromatin conformation reads and significantly lowers the input materials.

Despite that millions of candidate *cis*-regulatory sequences have been annotated in the human genome on the basis of biochemical signatures such as histone modification, transcription factor (TF) binding, and chromatin accessibility[3,28–32], only a handful of these candidate elements have been functionally validated in the native genomic context. High-throughput CRISPR–Cas9-mediated mutagenesis by single guide RNAs (sgRNAs) has been used to functionally characterize *cis*-regulatory elements in mammalian cells[33–37]. However, current approaches are limited because (1) not all sequences are suitable for CRISPR–Cas9-mediated genome editing, owing to the lack of protospacer-adjacent motifs (PAMs), which are required for targeting and DNA cutting by CRISPR–Cas9[38–40]; (2) CRISPR–Cas9-mediated genome editing with individual sgRNAs tends to cause point mutations or short insertions or deletions, thus necessitating the use of an unrealistically large number of sgRNAs to interrogate the human genome; and (3) it has been challenging to distinguish *cis*- and *trans*-regulatory elements.

In chapter 4, colleagues and I developed CREST-seq that allows the efficient discovery and functional characterization of the regulatory elements through the introduction of massively parallel kilobase-long deletions in the genome. We provide evidence in support of the utility of CREST-seq for the large-scale identification of *cis*-

regulatory elements in human embryonic stem cells (hESCs). We report the discovery of 45 regulatory sequences of *POU5F1*, and a surprisingly large number of enhancer-like promoters. Our results highlight a commonality that promoter of one gene can behave like an enhancer to regulate the expression of another gene.

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001).

2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. & Zhu, X. The Sequence of the Human Genome. Science 291, 1304–1351 (2001).

3. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

4. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods 10, 1213–1218 (2013).

5. Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S. & Crawford, G. E. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell 132, 311–322 (2008).

6. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y. & Greenleaf, W. J. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015).

7. Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C. & Shendure, J. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348, 910–914 (2015).

8. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature Methods 14, 975–978 (2017).

9. Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C. & Shendure, J. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174, 1309-1324.e18 (2018).

10. Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J. & Aerts, S. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nature Methods 16, 397–400 (2019).

11. de Boer, C. G. & Regev, A. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. BMC Bioinformatics 19, (2018).

12. Lareau, C. A., Duarte, F. M., Chew, J. G., Kartha, V. K., Burkett, Z. D., Kohlway, A. S., Pokholok, D., Aryee, M. J., Steemers, F. J., Lebofsky, R. & Buenrostro, J. D. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology (2019). doi:10.1038/s41587-019-0147-6

13. Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J. & Trapnell, C. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Molecular Cell 71, 858-871.e8 (2018).

14. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902.e21 (2019).

15. Coifman, R. R. & Lafon, S. Diffusion maps. Applied and Computational Harmonic Analysis 21, 5–30 (2006).

16. Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. & Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proceedings of the National Academy of Sciences 102, 7426–7431 (2005).

17. Kumar, S., Mohri, M. & Talwalkar, A. Ensemble Nystrom Method. 9

18. Li, M., Kwok, J. T. & Lu, B.-L. Making Large-Scale Nyström Approximation Possible. 12

19. Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. Cell Stem Cell 14, 762–775 (2014).

20. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature 502, 499–506 (2013).

21. Sexton, T. & Cavalli, G. The Role of Chromosome Domains in Shaping the Functional Genome. Cell 160, 1049–1059 (2015).

22. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics 14, 390–403 (2013).

23. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. Genes & Development 30, 1357–1382 (2016).

24. Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293 (2009).

25. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E. & Ruan, Y. An oestrogen-receptor-α-bound human chromatin interactome. Nature 462, 58–64 (2009).

26. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D.,

Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014).

27. Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-L., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G. & Ruan, Y. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell 163, 1611–1627 (2015).

28. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M. & Snyder, M. Architecture of the human regulatory network derived from ENCODE data. Nature 489, 91–100 (2012).

29. Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120 (2012).

30. Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., Yang, H., Wang, T., Lee, A. Y., Swanson, S. A., Zhang, J., Zhu, Y., Kim, A., Nery, J. R., Urich, M. A., Kuan, S., Yen, C., Klugman, S., Yu, P., Suknuntha, K., Propson, N. E., Chen, H., Edsall, L. E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.-Y., Chi, N. C., Antosiewicz-Bourget, J. E., Slukvin, I., Stewart, R., Zhang, M. Q., Wang, W., Thomson, J. A., Ecker, J. R. & Ren, B. Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell 153, 1134–1148 (2013).

31. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai,

L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).

32. Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B. E. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011).

33. Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.-C., Zhang, F., Orkin, S. H. & Bauer, D. E. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, 192–197 (2015).

34. Korkmaz, G., Lopes, R., Ugalde, A. P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R. & Agami, R. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nature Biotechnology 34, 192–198 (2016).

35. Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J. M., Gifford, D. K. & Sherwood, R. I. High-throughput mapping of regulatory DNA. Nature Biotechnology 34, 167–174 (2016).

36. Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y., Dixon, J., Maliskova, L., Guan, K., Shen, Y. & Ren, B. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Research 26, 397–405 (2016).

37. Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A. & Zhang, F. High-resolution interrogation of functional elements in the noncoding genome. Science 353, 1545–1549 (2016).

38. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science 337, 816–821 (2012).

39. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155, 733–740 (2009).

40. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 507, 62–67 (2014).

CHAPTER 1: SINGLE-NUCLEUS ANALYSIS OF ACCESSBILE CHROMATIN IN DEVELOPING MOUSE FOREBRAIN

1.1 Abstract

Analysis of chromatin accessibility can reveal the transcriptional regulatory sequences, but heterogeneity of primary tissues poses a significant challenge in mapping the precise chromatin landscape in specific cell types. Here, we report single nucleus ATAC-seq (snATAC-seq), a combinatorial barcoding-assisted single cell assay for transposase-accessible chromatin that is optimized for use on flash-frozen primary tissue samples. We apply this technique on the mouse forebrain through eight developmental stages. Through analysis of more than 15,000 nuclei, we identify 20 distinct cell populations corresponding to major neuronal and non-neuronal cell-types. We further define cell-type specific transcriptional regulatory sequences, infer potential master transcriptional regulators, and delineate developmental changes in forebrain cellular composition. Our results provide insight into the molecular and cellular dynamics that underlie forebrain development in the mouse and establish technical and analytical frameworks that are broadly applicable to other heterogeneous tissues.

1.2 Introduction

Transcriptional regulatory elements in the genome (*cis* regulatory elements) play fundamental roles in development and disease[1,2]. Analysis of chromatin accessibility in primary tissues using assays such as DNase-seq[3,4] and ATAC-seq[5,6] has identified millions of candidate *cis* elements in the human and mouse genomes[2,7]. However, we still lack precise information about the *cis* regulatory elements in specific cell types, because previous experiments performed on heterogeneous tissue samples yield an ensemble average signal from multiple constituent cell types. In some cases, specific cell types can be isolated from heterogeneous tissues using protein markers[6,8–10], but a more general strategy is needed to enable the study of cell type specific gene regulation on a larger scale.

In theory, single cell-based chromatin accessibility studies can be used for unbiased identification of subpopulations in a heterogeneous biological sample, and for identification of the regulatory elements active in each subpopulation. Indeed, proof of principle has been reported using cultured mammalian cells and cryopreserved blood cell-types[11–13]. However, to make these approaches more widely applicable, it is necessary to optimize them for primary tissues. One major difficulty in working with primary tissues is that they are typically preserved by flash freezing, which is not amenable to the isolation of intact single cells. Here, we show that it is possible to isolate single nuclei from frozen tissues and assay chromatin accessibility in these nuclei in a massively parallel manner.

1.3 Results

Method optimization and computational analysis framework. We adopted a combinatorial barcoding assisted single cell ATAC-seq strategy[12] and optimized it for frozen tissue samples (**Supplementary Methods**). Compared to previous reports[12], key modifications were made to maximally preserve nuclei integrity during sample processing and optimize transposase-mediated fragmentation of chromatin in individual nuclei (**Figure S1.1-S1.2**). We applied this modified protocol, hereafter referred to as snATAC-seq (single nucleus ATAC-seq), to mouse forebrain tissue from 8-week-old adult mice (P56) and from mouse embryos at seven developmental stages from embryonic day 11.5 (E11.5) to birth (P0) (**Figure 1.1a, b**). DNA libraries were sequenced to near saturation as indicated by a read duplication rate of 36-73% per sample. The barcode collision rate which assesses the probability of two nuclei sharing the same barcode combination was ~16% and slightly higher than expected and reported before (**Figure S1.3c**)[12]. We filtered out low-quality datasets using three stringent quality control criteria including read depth (**Figure S1.3d**), recovery rate of constitutively accessible promoters in each nucleus (**Figure S1.3e**), and signal-over-noise ratio estimated by fraction of reads in peak regions (**Figure S1.3f, Supplementary Methods**). In total, 15,767 high-quality snATAC-seq datasets were obtained. The median read depth per nucleus ranged from 9,275 to 18,397, with the median promoter coverage at 11.6% and the median fraction of reads in peak regions at 22%. Our protocol maintains the extraordinary scalability of combinatorial indexing, while featuring a ~6-fold increase in read depth per nucleus compared to previous reports. The high quality of the single nucleus chromatin accessibility maps was supported by strong concordance between the aggregate snATAC-seq data and bulk

ATAC-seq data (R > 0.9), and excellent reproducibility between independent snATAC-seq experiments (R > 0.91, **Figure 1.1c**, **Figure S1.4**).

The snATAC-seq profiles from each forebrain tissue arise from a mixture of distinct cell types. Enhancer regions are well known to display cell type-dependent chromatin accessibility[14], and are more effective at classifying cell types than promoters or transcriptomic data[11] (**Figure S1.5a, b**). Thus, we focused on Transcriptional Start Sites (TSS)-distal accessible chromatin regions (defined as all genomic elements outside a 2 kb window upstream the TSS), corresponding to putative enhancers, to group individual nucleus profiles into distinct cell types. We developed a novel computational framework to uncover distinct cell types from the snATAC-seq datasets without requiring prior knowledge (**Supplementary Methods**). First, we determined the open chromatin regions from the bulk ATAC-seq profiles of mouse forebrain tissue in seven fetal development time points and in adults, resulting in a total of 140,103 TSS-distal elements (**Figure 1.1d** and **Supplementary Methods**). Next, we constructed a binary accessibility matrix of open chromatin regions, using 0 or 1 to indicate absence or presence of a read at each open chromatin region in each nucleus (**Figure 1.1d**). We then calculated the pairwise similarity between cells using a Jaccard index, and applied a non-linear dimensionality reduction method, t-SNE[15], to project the Jaccard index matrix to a low-dimension space (**Figure 1.1d**)[16]. The final t-SNE plot depicts cell types as distinct clusters in a three-dimensional space (**Figure 1.1d**).

Identification of forebrain cell types from snATAC-seq profiles. We applied this computational framework first to 3,033 high-quality snATAC-seq profiles obtained from the adult forebrain (**Figure 1.2a**). As a negative control, we included 200 "shuffled" nuclear profiles (**Figure S1.5c, d** and **Supplementary Methods**). This analysis revealed 10 total clusters. As expected, the shuffled nuclei formed a distinct cluster with low intra-cluster similarity. In addition, one other cluster showed low intra-cluster similarity likely represents low quality nuclei or accessibility profiles resulting from barcode collision events (**Figure S1.3c**). After eliminating these nuclei, we determined 8 distinct cell type clusters from the adult forebrain (**Figure 1.2a** and **Figure S1.5c, d**). Notably, the clustering results were highly reproducible for two independent experiments (**Figure S1.5e, f**).

To categorize each cluster, we generated aggregate chromatin accessibility maps for each cluster and examined the patterns of chromatin accessibility at known cell type marker genes. We found three clusters with chromatin accessibility at *Neurod6* and other excitatory neuron-specific genes[17] (clusters EX1-3, **Figure 1.2b**, **Figure S1.6a**); two clusters with accessibility at the gene locus of *Gad1* likely representing inhibitory neurons (clusters IN1-2, **Figure 1.2b**, **Figure S1.6a**)[18]; one cluster with accessibility at the *Apoe* locus and other known astroglia markers[19] (cluster AC, **Figure 1.2b**); one cluster with accessibility at the *Mog gene locus* and other oligodendrocyte marker genes[20] (cluster OC, **Figure 1.2b**); and one microglia cluster with accessibility at genes encoding complement factors including the gene *C1qb* (cluster MG, **Figure 1.2b**, **Figure S1.6c-e**)[21]. We also compared the aggregate chromatin accessibility maps for each cluster to

previously published maps from sorted excitatory neurons[6], GABAergic neurons[8], microglia[21] and NeuN negative nuclei (which mostly comprise non-neuronal cells including astrocytes and oligodendrocytes[22]; **Figure 1.2b** and **Figure S1.7a-c**). Consistent with the accessibility patterns at marker gene loci, we observed that clusters EX1-3 were highly similar to sorted excitatory neurons. To further characterize the distinct excitatory neuron clusters, we compared EX1-3 with published bulk ATAC-seq data from different cortical layers and from dentate gyrus. Interestingly, we found that EX1 and EX3 were more similar to upper and lower cortical layers, respectively, whereas EX2 showed properties of dentate gyrus neurons (**Figure S1.8a**). Clusters IN1 was highly similar to sorted cortical GABAergic neurons[23]. Surprisingly, IN2 was more similar to sorted excitatory neurons than cortical GABAergic neurons. Distinctions between the inhibitory neuron clusters (IN1 and IN2) were not clear at this stage but came into focus later when we analyzed transcription factor (TF) motifs enriched in the accessible chromatin regions (described below). Clusters OC and AC resembled sorted NeuN negative cells, and cluster MG is similar to sorted microglia (**Figure 1.2b, c**)

According to our snATAC-seq data, the adult mouse forebrain consists of 52% excitatory neurons, 24% inhibitory neurons, 12% oligodendrocytes and 6% astrocytes and microglia, respectively (**Figure 1.2d**). Since the cell type proportion varies between different forebrain regions, for example cortex and hippocampus, the percentages derived from snATAC-seq represent an average of all forebrain regions (**Figure S1.7d, e**; **Figure 1.2e**). The predominance of neuronal nuclei derived from adult forebrain tissue was

confirmed by flow cytometry analysis using staining against the post-mitotic neuron marker NeuN[22] (**Figure S1.6b**; **Figure S1.7b, e**; **Figure 1.2e**).

Delineation of the *cis* regulatory landscape of specific cell types in the adult forebrain. The power of the snATAC-seq is not simply to delineate cell types, but further, to reveal the *cis*-regulatory landscape within each cell type. To this end, we calculated the cell type specificity of each putative *cis* regulatory element (i.e. chromatin accessibility region) using a Shannon entropy index (**Figure S1.9**). As expected, proximal promoter elements were accessible in more cell types, while the distal enhancer elements showed significantly higher cell type-specificity (Median value of 4.2% for proximal elements vs. 0.4% for distal elements) (**Figure S1.9a-d**). We next developed a feature selection method (**Supplementary Methods**) to identify the subset of elements that could best distinguish the 8 cell type clusters from each other. This approach identified 4,980 elements showing clear cell type dependent accessibility (**Figure 1.2e**). To gain insight into the key transcriptional regulators and pathways active in each cell type, we performed k-means clustering followed by motif enrichment analysis for these genomic elements (**Figure 1.2e, f** and **Figure S1.9d**). For each cell type, we observed an enrichment of binding motifs corresponding to key TFs (**Figure 1.2f**). For example, the binding motif for ETS-factor PU.1 was enriched in MG elements[24], motifs for SOX proteins were enriched in OC elements[25], bHLH motifs were enriched in EX1-3 elements, and DLX homeodomain factor motifs were enriched in IN elements (**Figure 1.2f**)[26]. Moreover, this analysis revealed an important difference between the inhibitory neuron clusters IN1 and IN2. We found that a binding motif for MEIS factors was enriched in a subset of elements specific

to IN2. Previous reports showed that MEIS2 plays a major role in generation of medium spiny neurons, the main GABAergic neurons in the striatum[27]. Accordingly, we identified gene loci of *Ppp1r1b* and *Drd1,* which encode markers of medium spiny neurons, to be highly accessible in IN2 but not IN1 (**Figure S1.10**)[27]. These data suggest that IN2 may represent medium spiny neurons, while IN1 could represent a distinct class of GABAergic neurons. We also identified motifs that were differentially enriched between EX1, EX2 and EX3. Notably, regions specific for EX1 and 3 were enriched for motifs from the Forkhead family and EX2 was enriched for motifs recognized by MEF2C (**Figure S1.8c**), which has been shown to play an important role in hippocampus mediated memory[28]. A comparison with data from cell-type specific differentially methylated regions identified by single cell DNA-methylation analysis of neurons showed that both methods were able to identify inhibitory and excitatory neuron specific elements (**Figure S1.11**)[29].

Profiling embryonic forebrain development using snATAC-seq. We next extended our framework by analyzing the snATAC-seq profiles derived from fetal mouse forebrains at seven developmental stages (**Figure 1.1b**), seeking to reveal developmental dynamics of transcriptional regulation at the cellular level. The developmental stages examined cover key events from the onset of neurogenesis to gliogenesis[30]. From 12,733 high-quality snATAC-seq profiles we identified 12 distinct sub-populations (**Figure 1.3a**) that exhibit changes in abundance through development (**Figure 1.3a-c**). This broad cell-type classification allowed us to profile the dynamic cis-regulatory landscape of forebrain development. Based on accessibility profiles at gene loci of known marker genes, we assigned these cell populations to radial glia, excitatory neurons, inhibitory neurons,

astrocytes and erythromyeloid progenitors (EMP) (**Figure 1.3b**)[24,31]. Interestingly, the EMP cluster was restricted to E11.5, whereas the astrocyte cluster was present after E16.5 and expanded dramatically around birth (**Figure 1.3b, c**)[30], highlighting two developmental processes: invasion of myeloid cells into the brain prior to neurogenesis, and gliogenesis succeeding neurogenesis after E16.5[30]. Mature excitatory neurons (eEX2) were indicated by increased accessibility at *Neurod6* which encodes a post-mitotic neuron marker, and absence of signal at the *Hes5 gene,* which encodes a Notch effector and a marker gene for neuronal progenitors (**Figure 1.3b, c**)[31]. This cell type expanded in abundance between E12.5 and E13.5 and followed the emergence of early differentiating neurons (eEX1, **Figure 1.3b, c**). Remarkably, inhibitory-neuron-like cells were already present at E11.5 (**Figure 1.3b**).

Identification of lineage specific transcriptional regulators during embryonic forebrain development. To identify the transcriptional regulatory sequences in each sub-population, we identified 16,364 genomic elements that show cell-population-specific chromatin accessibility and best separate the sub-cell populations (**Figure 1.4a**). To further characterize these elements, we performed motif enrichment analysis and gene ontology analysis of each cluster using GREAT[32]. Our analysis showed that genomic elements that were mostly associated with radial glia like cell groups (**Figure 1.4a**, RG1-4) fell into regulatory regions of genes involved in early forebrain developmental processes including “Forebrain regionalization” (**Figure 1.4b**, K1), “Central nervous system development” (**Figure 1.4b**, K3) or “Forebrain development” (**Figure 1.4b**, K5). These elements were enriched for homeobox motifs corresponding to LHX-transcription

factors including LHX2 (**Figure 1.4c**, K1,3,5), which is critical for generating the correct neuron numbers by regulating proliferation of neural progenitors[33] and for temporally promoting neurogenesis over astrogliagenesis[34]. Remarkably, one of these clusters was also enriched for both the proneural bHLH transcription factor ASCL1 (*Mash1*) and its co-regulator POU3F3 (*Brn1*) (**Figure 1.4c**, K5)[35]. ASCL1 is required for normal proliferation of neural progenitor cells and implicated in a DLX1/2 associated network that promotes GABAergic neurogenesis[36,37]. In line with this, associated genomic elements were also accessible in one inhibitory neuron cluster (eIN2, **Figure 1.4c**, K5).

We also identified transcriptional regulators that were specifically associated either with neurogenesis or gliogenesis during forebrain development. For example, the early astrocyte (eAC)-specific elements were located in open chromatin regions near genes involved in “glia cell fate commitment” and the top enriched transcription factor motif was NF1-halfsite (**Figure 1.4a-c**, K2). Previous studies showed that NF1 transcription factor NF1A alone is capable of specifying glia cells to the astrocyte lineage[25]. NFIX is another NF1 family member with proneural function[38]. This motif is enriched together with the bHLH transcription factor NEUROD1 binding sites mainly in open chromatin regions found in the excitatory neuron cell population (**Figure 1.4c**, K4,12,13)[31]. Based on chromatin accessibility profiles at marker gene loci, we have previously assigned two cell clusters to the excitatory neuron lineage (eEX1, eEX2, **Figure 1.3b**). Compared to cluster eEX2, eEX1 showed increased accessibility at both radial glia associated open chromatin (**Figure 1.4a**, K4; **Figure 1.3b**) and chromatin regions associated with “CNS neuron differentiation” (**Figure 1.4a**, K12). In addition, eEX1 nuclei preceded the emergence of

eEX2 nuclei during development (**Figure 1.3c**). These findings indicate that eEX1 might represent a transitional state during excitatory neuron differentiation.

The bHLH transcription factor family consists of several subfamilies that recognize different DNA motifs[39]. NEUROD1 belongs to a sub-family of transcription factors that bind to a central CAT motif whereas other transcription factors such as TCF12 preferentially bind to a CAG motif[39]. Our snATAC-seq profiles revealed an enrichment of the TCF12-binding motif in regions associated with “Cortex GABAergic interneuron differentiation” in contrast to the excitatory neuron associated enrichment for NEUROD1 (**Figure 1.4a-c**, K4, 11-13)[40]. Analysis of the inhibitory neuron cluster eIN3 specific genomic elements showed a remarkable bias in proximity to genes associated with “Skeletal muscle organ development” (**Figure 1.4a, b**, K8). More detailed analysis revealed that the underlying genes *Mef2c/d* and *Foxp1/2* as well as Drd2/3 encode transcription factors and dopamine receptors indicating differentiating striatal medium spiny neurons[41,42]. This finding was consistent with the enrichment for MEIS-homeodomain factors in these regions (**Figure 1.4c**, K8) comparable to the medium spiny neuron cluster in adult forebrain (**Figure 1.2e, f**, K8; **Figure S1.10**). Further, genomic elements specific to the EMP cluster were associated with genes involved in “Myeloid cell development” (**Figure 1.4a-c**, K14) and enriched for motifs of the ubiquitous AP-1 transcription factor complexes that have been described to play a role in shaping the enhancer landscape of macrophages[43].

Finally, we attempted to identify developmental dynamics of elements within each cell cluster (**Figure S1.11**). Our analysis revealed between 41 and 2,114 dynamic genomic elements for each cell type (**Figure S1.12c-g**). Regions that are more accessible after birth (P0) compared to early time points were enriched for the RFX1 motif in the GABAergic neuron including the cluster eIN1 as well as in the excitatory neuron cluster eEX2 (**Figure S12d, e**) indicating a general role of the evolutionary conserved RFX factors in perinatal adaptation of brain cells. Several family members including RFX1 are expressed in the brain and have been implicated to regulate cilia e.g. in sensory neurons[44].

Functional and anatomical annotation of identified candidate *cis*-regulatory elements. While assessment of open chromatin plays an important role in predicting regulatory elements in the genome. it does not provide direct information of functional activity. To address this point, we asked if cluster-specific trannsposase accessible chromatin in the embryonic forebrain overlaps with genomic elements tested in reporter assays to validate enhancer activity in mouse embryonic forebrain *in vivo*[45].First, we focused our analysis on all genomic elements with validated functional activity in the forebrain and a subset shown to be active only in the subpallium[46,47]. The subpallium is a brain region that gives rise to GABAergic and cholinergic neurons[46]. In total, 63.1 % (275/436) of all forebrain enhancer and 64.8% (59/91) of subpallial enhancer were represented in our subset of genomic elements, respectively, indicating a high degree of sensitivity. Next, we calculated the relative enrichment of subpallial enhancers over total forebrain enhancers for each cluster. Remarkably, subpallial enhancers were only

enriched in clusters K9-11, which were assigned to the GABAergic neuron lineage (**Figure 1.4d, e**; **Figure S1.13**). Next, we found that elements mainly accessible in radial glia cells were active in pallial regions (**Figure 1.4a**, K1, 3, 4; **Figure S1.13**). Surprisingly, elements of cluster K5 were active in dorsal and lateral pallial regions as well as in the lateral ganglionic eminence indicating conserved roles for these genomic elements in a wide variety of regions in the developing forebrain (**Figure 1.4a**; **Figure S1.13**). Integration of genomic elements identified by snATAC-seq in specific cell clusters with transgenic enhancer assays confirms the high specificity and sensitivity of snATAC-seq in identifying cell populations and their underlying regulatory elements.

1.4 Discussion

Tissue heterogeneity has been a significant hurdle in the dissection of gene regulatory programs driving mammalian development. While single cell-based analysis of chromatin accessibility has been reported, a major challenge lies in the requirement for fresh cell populations by the published methods, whereas most biological biopsy samples tissue banks are either frozen or in Formalin Fixed Paraffin Embedded blocks. We report here a general approach (snATAC-seq) and a computational framework that can be used to dissect cellular heterogeneity and delineate cell-type-specific gene regulatory sequences in snap frozen primary tissues. We applied snATAC-seq to heterogeneous forebrain samples from adult and embryonic mice and resolved specific cell types in these samples. Similar to other approaches such as single cell RNA-seq[48] and single cell DNA methylation analysis[29], snATAC-seq can be used to identify cell types de novo in a heterogeneous tissue, facilitating generation of cell atlases in the brain and other tissues. In addition, snATAC-seq catalogues the candidate enhancers for each cell type, enabling the dissection of gene regulatory programs without the need to purify specific cell types. As such, this method is particularly suitable for studying cell populations in complex tissues where cellular surface markers are not available. The current framework allows analysis of major cell-types with a relative abundance of at least 5% as shown for microglia in the adult forebrain. It is expected that increasing the number of cells profiled per experiment will linearly increase the sensitivity of cell type detection. Indeed, the presented combinatorial barcoding protocol can be scaled up to > 5,000 high quality nuclei per experiment simply by working in 384-well plate format rather than 96

well plates. Increasing the number of barcodes during tagmentation will also help to lower the final barcode collision rate without limiting the throughput.

Through integrative analysis of single nuclei chromatin accessibility profiles, we tracked changes in the relative proportions of these cell types during development, identified putative regulatory elements active within each cell type, and used those regulatory elements to reveal key TFs in specific forebrain cell types. Therefore, our results provide a unique view of the cell type specific *cis* regulatory landscape in the forebrain. We expect that with larger cell numbers in the future it will be possible to uncover previously unknown regulatory elements in rare cell types. Moreover, applying snATAC to human tissues samples and integration with genomic variants variant calls may reveal relative contributions of distinct cell-types to diseases like schizophrenia or Alzheimer's. We anticipate that our snATAC-seq approach will be a valuable tool for analysis of other brain regions and non-neuronal tissues and will help to pave the way to a better understanding of mammalian developmental programs.

1.5 Acknowledgments

This study was funded in part by the National Human Genome Research Institute (U54HG006997 to B.R.), National Institute Mental Health (1U19MH114831 to B.R., U01MH098977 to K.Z.), NIH (2P50 GM085764 to B.R.), and the Ludwig Institute for Cancer Research (to B.R.). S.P. was supported by a postdoctoral fellowship from the Deutsche Forschungsgemeinschaft (DFG, PR 1668/1-1). R.R. was supported by a Ruth L. Kirschstein National Research Service Award NIH/NCI T32 CA009523. We thank B. Li for bioinformatic support. We thank M. He and T. Osothprarop for providing the Tn5 enzyme. We thank D. Gao for sequencing on the MiSeq. Research conducted at the E.O. Lawrence Berkeley National Laboratory was performed under U.S. Department of Energy Contract DE-AC02-05CH11231, University of California.

Chapter 1, in full, is a reprint of the material as it appears in Nature Neuroscience 2018 “Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation”. Sebastian Preissl, Rongxin Fang, Hui Huang, Yuan Zhao, Ramya Raviram, David U. Gorkin, Yanxiao Zhang, Brandon C. Sos, Veena Afzal, Diane E. Dickel, Samantha Kuan, Axel Visel, Len A. Pennacchio, Kun Zhang & Bing Ren. The dissertation author was the primary investigator and author of this paper.

1.6 Author Contributions

Study was conceived and designed by B.R., S.P., R.F. and K.Z.; Study was overseen by B.R. and K.Z. ;Experiments performed by S.P., B.C.S, H.H.; Tissue collection by V.A., D.E.D., A.V., L.A.P.,S.P.; Sequencing performed by S.K.;

Computational strategy developed by R.F., S.P., Data analysis performed by S.P., R.F., Yu.Z., R.R., Ya.Z., D.U.G; Manuscript written by S.P., R.F. and B.R. A.V., L.A.P., and K.Z. provided input and edited the manuscript. All authors discussed results and commented on the manuscript.

1.7 Figures

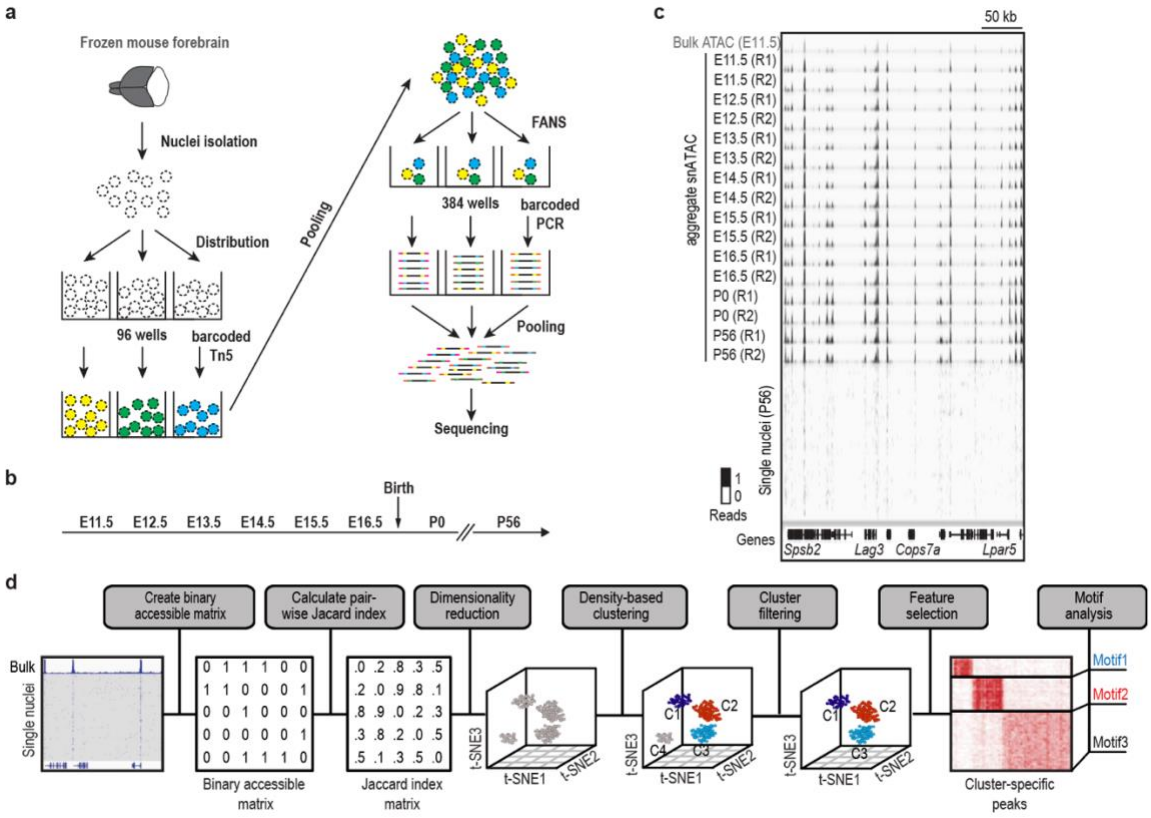
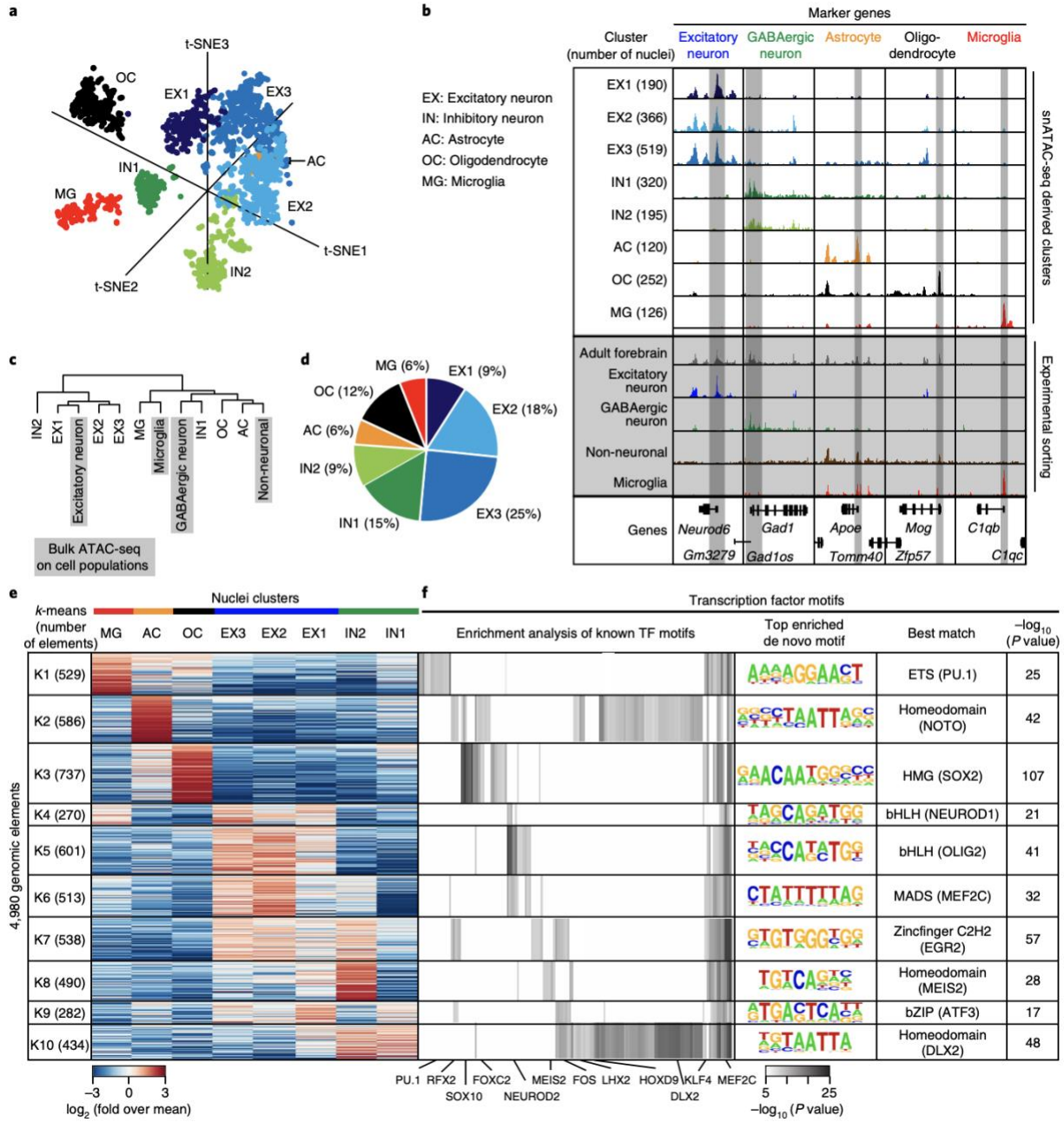


Figure 1.1. Overview of the experimental and computational procedures of snATAC-seq. (**a**) Following nuclei isolation from frozen forebrain tissue biopsies, tagmentation of 4,500 permeabilized nuclei was carried out using barcoded Tn5 in 96-well plates. After pooling, 25 nuclei were sorted into each well of a 384-well plate and PCR was carried out to introduce the second set of barcodes. FANS: Fluorescence assisted nuclei sorting. (**b**) Overview of the developmental time points examined in the current study. E: embryonic day; P: postnatal day; (**c**) Chromatin accessibility profiles of aggregate snATAC-seq (black tracks) agree with bulk ATAC-seq (grey, top track) and are consistent between independent experiments. (**d**) Framework of computational analysis of snATAC-seq data.

Figure 1.2. Deconvolution of cell types in the p56 mouse forebrain and identification of potential master regulators of each cell type. (**a**) Clustering of single nuclei from both experiments revealed 8 different cell groups in adult forebrain. (**b**) Aggregate chromatin accessibility profiles for each cell cluster and the bulk ATAC-seq for the sorted cell populations or the whole forebrain at several marker gene loci (Bulk data are shaded in grey). (**c**) Hierarchical clustering of aggregate single nuclei ATAC-seq data and the bulk ATAC-seq data sets. (**d**) Cellular composition of adult forebrain derived from snATAC-seq data. (**e**) K-means clustering of 4,980 genomic elements based on chromatin accessibility. (**f**) enrichment analysis for transcription factor motifs in each cell group. For enrichment of known motifs, one-tailed Fisher's Exact test was used to calculate significance[49]. Displayed p-values are Bonferroni corrected for multiple testing. For *de novo* motif enrichment testing a hypergeometric test was used[50]. Displayed p-values are not corrected for multiple testing.


a
t-SNE3
OC
EX1
EX3
AC
IN1
MG
EX2
t-SNE1
IN2
t-SNE2
b
Marker genes
Cluster (number of nuclei)
Excitatory neuron
GABAergic neuron
Astrocyte
Oligo-dendrocyte
Microglia
EX: Excitatory neuron
IN: Inhibitory neuron
AC: Astrocyte
OC: Oligodendrocyte
MG: Microglia
EX1 (190)
EX2 (366)
EX3 (519)
IN1 (320)
IN2 (195)
AC (120)
OC (252)
MG (126)
snATAC-seq derived clusters
Adult forebrain
Excitatory neuron
GABAergic neuron
Non-neuronal
Microglia
Experimental sorting
Genes
Neurod6
Gm3279
Gad1
Gad1os
Apoe
Tomm40
Mog
Zfp57
C1qb
C1qc
c
IN2
EX1
EX2
EX3
MG
IN1
OC
AC
Excitatory neuron
Microglia
GABAergic neuron
Non-neuronal
Bulk ATAC-seq on cell populations
d
MG (6%)
EX1 (9%)
OC (12%)
EX2 (18%)
AC (6%)
IN2 (9%)
IN1 (15%)
EX3 (25%)
e
Nuclei clusters
k-means (number of elements)
MG
AC
OC
EX3
EX2
EX1
IN2
IN1
4,980 genomic elements
K1 (529)
K2 (586)
K3 (737)
K4 (270)
K5 (601)
K6 (513)
K7 (538)
K8 (490)
K9 (282)
K10 (434)
−3
0
3
log2 (fold over mean)
f
Transcription factor motifs
Enrichment analysis of known TF motifs
Top enriched de novo motif
Best match
−log10 (P value)
ETS (PU.1)
25
Homeodomain (NOTO)
42
HMG (SOX2)
107
bHLH (NEUROD1)
21
bHLH (OLIG2)
41
MADS (MEF2C)
32
Zincfinger C2H2 (EGR2)
57
Homeodomain (MEIS2)
28
bZIP (ATF3)
17
Homeodomain (DLX2)
48
PU.1
RFX2
SOX10
FOXC2
NEUROD2
MEIS2
FOS
LHX2
HOXD9
KLF4
DLX2
MEF2C
5
25
−log10 (P value)

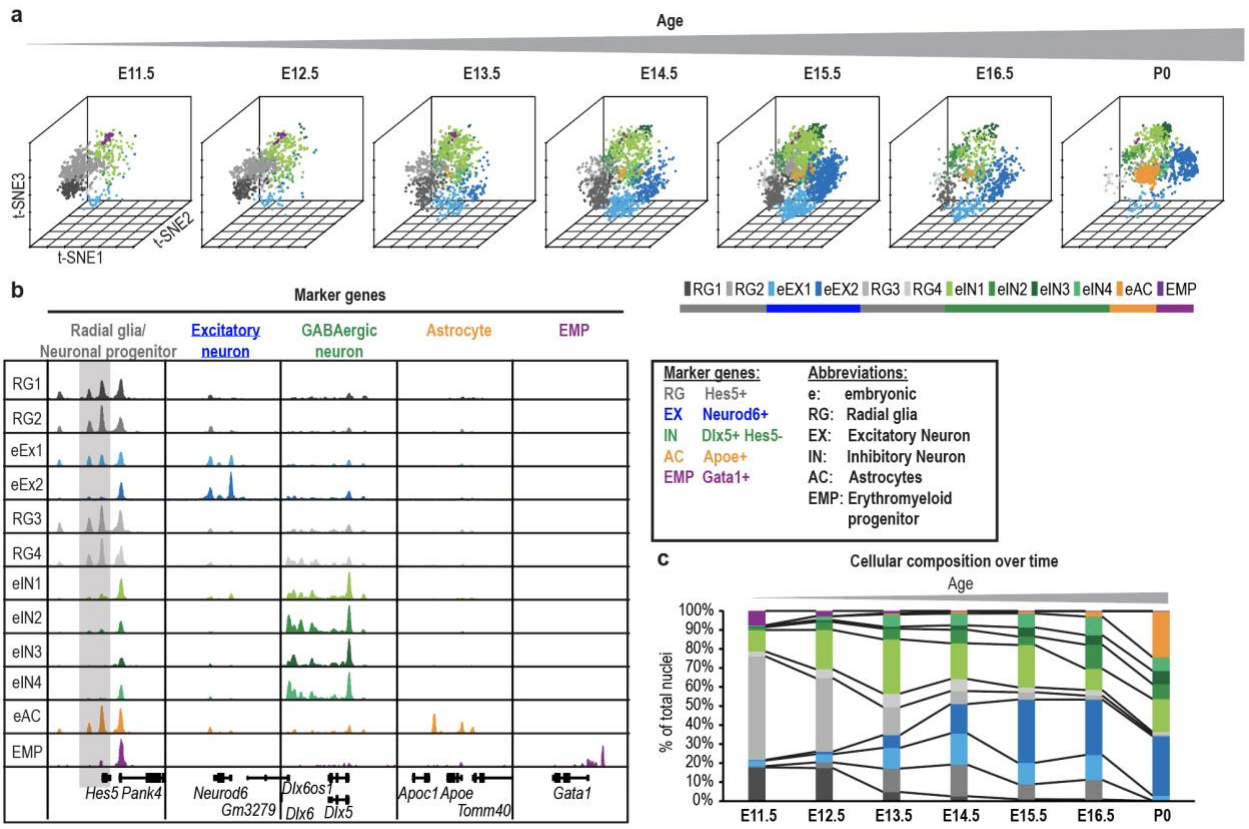


Figure 1.3. SnATAC-seq analysis reveals the timing of neurogenesis and gliogenesis during embryonic forebrain development. (**a**) Clustering of single nuclei from both independent experiments revealed 12 different cell groups with changing relative abundance. (**b**) Aggregate chromatin accessibility profiles for cell clusters and at marker gene loci used to assign cell types. For better visualization, *Hes5* gene locus is grey shaded. (**c**) Quantification of cellular composition during forebrain development.

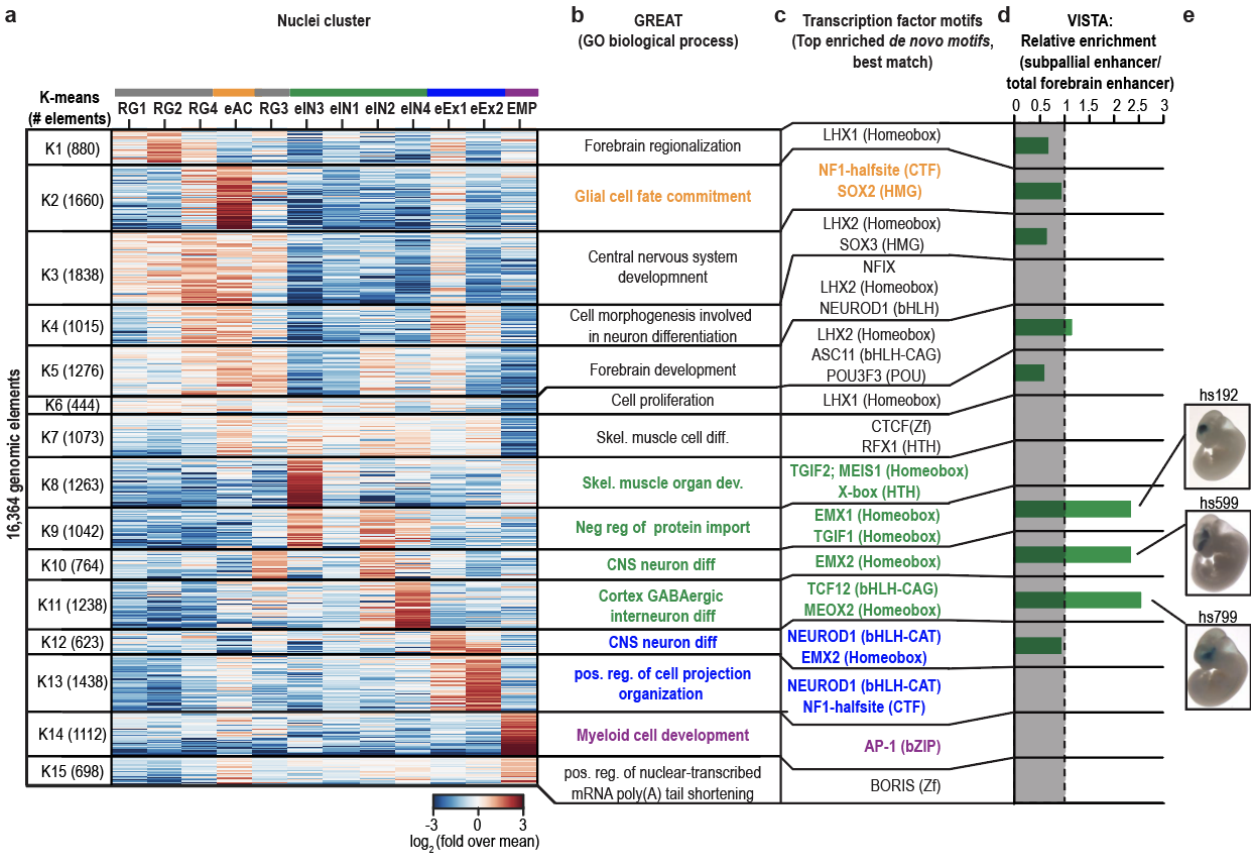


Figure 1.4. SnATAC-seq analysis uncovers cis regulatory elements and transcriptional regulators of lineage specification in the developing forebrain. (**a**) A heat map shows the results of K-means clustering of 16,364 candidate *cis* regulatory elements based on chromatin accessibility in different cell types. (**b**) Gene ontology analysis of each cell type using GREAT[32]. (**c**) Transcription factor motifs enriched in each group[50]. (**d**) Enrichment of enhancers that were functionally validated as part of the VISTA database[45]. (**e**) Representative images of transgenic mouse embryos showing LacZ reporter gene expression under control of the indicated subpallial enhancers. Pictures were downloaded from the VISTA database[45].

1.8 Supplementary Methods

Mouse tissues. All animal experiments were approved by the Lawrence Berkeley National Laboratory Animal Welfare and Research Committee or the University of California, San Diego, Institutional Animal Care and Use Committee. Forebrains from embryonic mice (E11.5-E16.5) and early postnatal mice (P0) were dissected from one pregnant female or one litter at a time and combined. For breeding, animals were purchased from Charles River Laboratories (C57BL/6NCrl strain) or Taconic Biosciences (C57BL/6NTac strain) for E14.5 and P0. Breeding animals for other time points were received from Charles River Laboratories (C57BL/6NCrl). Dissected tissues were flash frozen in a dry ice ethanol bath. For the adult time point (P56), the forebrain from 8-week old male C57BL/6NCrl mice (Charles River Laboratories) were dissected and flash frozen in liquid nitrogen separately. Tissues were pulverized in liquid nitrogen using pestle and mortar. For each time point two replicates were processed (n = 2 per time point).

Transposome generation. To generate A/B transposomes, A and B oligos were annealed to common pMENTs oligos (95°C 2 min, 14°C ∞ (cooling rate: 0.1°C/s)) separately. Next, barcoded transposons were mixed in a 1:1 molar ratio with unloaded transposase Tn5 which was generated at Illumina. Mixture was incubated for 30 min at room temperature. Finally, A and B transposomes were mixed. For combinatorial barcoding we used 8 different A transposons and 12 distinct B transposons which eventually resulted in 96 barcode combinations[51].

Combinatorial barcoding assisted single nuclei ATAC-seq. Combinatorial ATAC-seq was performed as described previously with modifications[12]. 5-10 mg frozen tissue was transferred to a 1.5 ml Lobind tube (Eppendorf) in 1 ml NPB (5 % BSA (Sigma), 0.2 % IGEPAL-CA630 (Sigma), cOmplete (Roche), 1 mM DTT in PBS) and incubated for 15 min at 4 °C. Nuclei suspension was filtered over a 30 µm Cell-Tric (Sysmex) and centrifuged for 5 min with 500 x g. Nuclei pellet was resuspended in 500 µl of 1.1x DMF buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF) and nuclei were counted using a hemocytometer. Concentration was adjusted to 500 /µl and 4500 nuclei were dispensed into each well of a 96 well plate. For tagmentation, 1 µl barcoded Tn5 transposome (0.25 µM)[51] was added to each well, mixed 5 times and incubated for 60 min at 37°C with shaking (500 rpm). To quench the reaction 10 µl 40 mM EDTA were added to each well and plate was incubated at 37°C for 15 min with shaking (500 rpm). 20 µl sort buffer (2 % BSA, 2 mM EDTA in PBS) were added to each well and all wells combined afterwards. Nuclei suspension was filtered using a 30 µm CellTric (Sysmex) into a FACS tube and 3 µM Draq7 (Cell Signalling) was added. Using a SH800 sorter (Sony) 25 nuclei were sorted per well into 4 96-well plates (total of 384 wells) containing 18.5 µl EB (50 pM Primer i7, 200 ng BSA (Sigma)). Sort plates were shortly spun down. After addition of 2 µl 0.2 % SDS samples were incubated at 55°C for 7 min with shaking (500 rpm). 2.5 µl 10% Triton-X was added to each well to quench SDS. Finally, 2 µl 25 µM Primer i5 and 25 µl NEBNext® High-Fidelity 2X PCR Master Mix (NEB) and samples were PCR amplified for 11 cycles (72°C 5 min, 98°C 30 s,[98°C 10 s, 63°C 30 s, 72°C 60 s] x 11, 72°C ∞). Following PCR, all wells were combined (around 15.5 mL) and mixed with 80 ml PB including pH-indicator (1:2500, Qiagen) and 4 ml Na-

Acetate (3 M, pH = 5.2). Purification was carried out on 4 columns following the MinElute® PCR Purification Kit manual (Qiagen). DNA was eluted with 15 µl EB and eluate from all four columns was combined in a LoBind Tube (Eppendorf). For Ampure XP Bead (Beckmann Coulter) cleanup 170 µl EB buffer and 110 µl Ampure XP Beads (0.55x) were added to 30 µl eluate. After incubation at room temperature for 5 min and magnetic separation supernatant was transferred to a new tube and another 190 µl Ampure XP Beads (1.5x) were added. After incubation beads were washed twice on the magnet using 500 µl 80 % EtOH. After drying the beads for 7 min at room temperature library was eluted with 20 µl EB (Qiagen). Libraries were quantified using Qubit fluoromoeter (Life technologies) and nucleosomal pattern was verified using Tapestation (High Sensitivity D1000, Agilent). 25 pM library was loaded per lane of a HiSeq2500 sequencer (Illumina) using custom sequencing primers[51] and following read lengths: 50 + 43 + 37 + 50 (Read1 + Index1 + Index2 + Read2). The first 8 bp of Index1 correspond to the p7 barcode and the last 8 bp to the i7 barcode. The first 8 bp of Index2 correspond to the i5 barcode and the last 8 bp to the p5 barcode. Since Index1 and 2 each contain 2 barcodes separated by a common linker sequence, we generated a spike-in library using different transposon and PCR primer sequences to balance the bases within each detection cycle. For the human-mouse mixture experiment, E15.5 forebrain and GM12878 nuclei were mixed in a 1:1 ratio prior to tagmentation. Samples were processed as above with the exceptions that just 96 wells were used after nuclei sorting and PCR amplification was performed for 13 cycles. The final library was loaded at 15 pM and sequenced using a MiSeq (Illumina) with following read lengths: PE 44 + 43 + 37 +44 (Read1 + Index1 + Index2 + Read2).

Cell culture. GM12878 (Coriell Institute for Medical Research) cells were cultured in RPMI1640 medium (Thermo Fisher Scientific) containing 2 mM L-glutamine (Thermo Fisher Scientific), 15% foetal bovine serum (Gemini Bioproducts) and 1 % Penicillin-Streptomicin (Thermo Fisher Scientific) in T25 Flasks (Corning) at 37°C under 5% carbon dioxide. For the snATAC-seq mixture experiment, cells were harvested by centrifugation, washed with PBS (Thermo Fisher Scientific) and resuspended in NPB (5 % BSA (Sigma), 0.2 % IGEPAL-CA630 (Sigma), cOmplete (Roche), 1 mM DTT in PBS). Samples were incubated 5 min at 4 °C and finally nuclei were pelleted by centrifugation (500g, 5min, 4 °C). Nuclei pellet was resuspended in 500 µl of 1.1x DMF buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF) and nuclei were counted using a hemocytometer.

NeuN negative sorting. 10 mg adult forebrain tissue (P56) were resuspend in 500 µl lysis buffer (0.5% BSA, 0.1% Triton-X, cOmplete (Roche), 1 mM DTT in PBS) and incubated for 10 min at 4°C. After spinning down (5 min, 500 x g) sample was resuspended in 500 µl staining buffer (0.5% BSA in PBS). Nuclei suspension was incubated with anti-NeuN antibody (1:5000, MAB377, Lot 2806074, EMD Millipore) for 30 min at 4°C. After centrifugation nuclei were resuspend in 500 µl staining buffer (0.5% BSA in PBS) containing anti-mouse Alexa488-antibody (1:1000, A11001, Lot 1696425, Thermo Fisher Scientific). After incubating for 30 min at 4°C, nuclei were pelleted (5 min 500 x g) and resupended in 700 ul sort buffer (1% BSA, 1mM EDTA in PBS). After filtration into a FACS tube 5 ul DRAQ7 (Cell Signalling Technologies) was added and NeuN-negative nuclei were sorted using a SH800 sorter (Sony) into 5% BSA (Sigma) in PBS.

Bulk ATAC-seq. ATAC-seq was performed on 20,000 sorted nuclei as described previously with minor modifications[52]. After adding IGEPAL-CA630 (Sigma) in a final concentration of 0.1 % nuclei were pelleted for 15 min at 1000 x g. Pellet was resupended in 19 µl 1.1x DMF buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF). After addition of 1 µl Tn5 transposomes (0.5 µM) tagmentation was performed at 37°C for 60 min with shaking (500 rpm). Next, samples were purified using MinElute columns (Qiagen), PCR-amplified for 8-10 cycles with NEBNext® High-Fidelity 2X PCR Master Mix (NEB, 72°C 5 min, 98°C 30 s,[98°C 10 s, 63°C 30 s, 72°C 60 s] x cycles, 72°C ∞). Amplified libraries were purified using MinElute columns (Qiagen) and Ampure XP Bead (Beckmann Coulter). Sequencing was carried out on a HiSeq2500 or 4000 (50 bp PE, Illumina).

Single nuclei ATAC-seq data processing pipeline. Our in-house pipeline implements the following major steps:

- Step 1. Read alignment. Paired-end sequencing reads were aligned to mm10 reference genome using Bowtie2[53] in paired-end mode with following parameters “bowtie2 -p 5 -t -X2000 --no-mixed --no-discordant"
- Step 2. Alignment filtering. Non-uniquely mapped (MAPQ < 30) and improperly paired (flag = 1804) alignments were filtered.
- Step 3. Barcode error correction. Each barcode consists of four 8 bp long indexes (i5, i7, p5, p7). Reads with barcode combinations containing more than 1 mismatch (or 1

edit distance) for any index were removed. Index with less than 1 mismatch were changed to its closest index.

- Step 4. Reads separation. Reads were separated into individual cells based on the barcode combination.
- Step 5. Mark and remove PCR duplicates. For individual cells, we sorted reads based on the genomic coordinates using "samtools sort"[54], then marked and removed PCR duplicates using Picard tools (MarkDuplicates).
- Step 6. Mitochondrial reads removal. Reads mapped to the mitochondrial genome were filtered.
- Step 7. Adjusting position of Tn5 insertion. All reads aligning to the + strand were offset by +4 bp, and all reads aligning to the - strand were offset -5 bp.
- Step 8. Quality assessment of each single cell. Calculate coverage of constitutively accessible promoters (promoters that are accessible across all tissues/cell line from ENCODE DHS), number of reads and signal-over-noise ratio estimated by "reads in peaks" ratio for each cell.
- Step 9. Cell selection. We only kept cells that pass our threshold (1) coverage of constitutively accessible promoter > 10%; 2) number of reads > 1,000; 3) reads in peak ratio greater than estimation from corresponding bulk ATAC-seq level.
- Step 10. Replicates separation. Selected cells were separated into two replicates based on the predefined barcode combination.

Single nuclei ATAC-seq cluster analysis. Cluster analysis partitions cells into groups such that cells from the same group have higher similarity than cells from different

groups. Here, we developed a pipeline to obtain cell clusters (https://github.com/r3fang/snATAC). We first generated a catalogue of accessible chromatin regions using bulk ATAC-seq data and created a binary accessible matrix. Chromatin sites were 1 for a given cell if there was a read detected within the peak region. Next, we calculated paired-wise Jaccard index between every two cells on the basis of overlapping open chromatin regions. Next, we applied a non-linear dimensionality reduction method (t-SNE) to map the high-dimensional structure to a 3-D space[15]. This transforms high-dimensional structures to dense data clouds in a low-dimensional space, allowing partitioning of cells using a density-based clustering method[16]. We then identified the optimal number of cell clusters using the Dunn index[55]. Finally, we compared our cluster results to those of "shuffled" to further verify our cluster result is not driven by library complexity or other confounding factors.

- Step1. Determining accessible chromatin sites in single cells. To catalogue accessible chromatin sites in individual cells, we first created a reference map of open chromatin sites determined by bulk ATAC-seq. The chromatin accessibility maps from different time points (from E11.5 to P56) were merged into a single reference file using BEDtools[56]. For clustering of single cells, we have tested clustering performance using accessible promoters (2kb upstream of TSS) and distal elements, respectively, and found that clusters by distal elements outperformed promoters with lower Kullback-Leibler divergence (**Figure S1.5**). Therefore, we decided to only focus on distal genomic elements as features to perform clustering. Reads in individual cells overlapping with accessible sites were identified. We generated an accessible matrix

of the reads counts overlapping each individual accessible sites (columns) in each cell (row).

- Step 2. Binary Accessible Matrix. We next converted the chromatin accessibility matrix to a binary matrix $M_{N\times D}$ in which M_{ij} is 1 if any read in cell i mapped to region j.
- Step 3. Jaccard Index Matrix. Jaccard index matrix $J_{N\times N}$ were calculated between every two cells in which J_{ij} measures the commonly shared open chromatin regions between cell C_i and C_j as following:

$$J_{ij} = \frac{|M_i \cap M_j|}{|M_i \cup M_j|}$$

  Diagonal elements of $J_{N\times N}$ are set to be 0 as required by t-SNE analysis.
- Step 4. Dimensionality reduction using t-SNE. Using Jaccard index matrix $J_{N\times N}$ as input, we next applied t-SNE to map the N-dimensional data to a 3-D space[15]. Since t-SNE has a non-convex objective function, it is possible that different runs yield different solutions[15]. Thus, we ran t-SNE several times with different initiations and used the result with the lowest Kullback-Leibler divergence and best visualization. In a previous study sequencing depth was a confounding factor and highly correlated with the first principle component of PCA analysis (Pearson correlation >0.95)[12]. However, we did not observe correlation between sequencing depth and any of the t-SNE dimension. We expected that the coherent structure of the open chromatin landscape of cells with high similarity would rely on a continuous and smooth 3-D structure and cells for different groups would locate to distinct parts of the plot. We used t-SNE to transform the high-dimensional structures to dense data clouds in the

3-D space[15]. Finally, we applied a density-based clustering method to identify different cell populations within the embedded 3-D space[16].

- Step 5. Density-based clustering. We applied a density-based clustering method to partition cells into groups in the embedded 3-D space[16]. The method identifies cluster centres that are characterized by two properties: 1) high local density ρ_i and 2) large distance δ_i from points of higher density, which are centers of the clusters[16]. Any cells that showed values above defined thresholds (ρ_0, δ_0) were considered as centers of cluster. Next, the rest of cells were assigned to the center as described here[16]. Clearly, different thresholds (ρ_0, δ_0) will generate different number of clusters. To find the optimal number of clusters, we adopted the method developed by Habib et al to evaluate the quality of different cluster results[55].
- Step 6. Number of clusters. In detail, Habib's method applied the Dunn index to quantify the quality of cluster result as following[55]:

$$DB = \frac{min_{1 \le i < j \le n} \Delta(C_i, C_j)}{max_{1 \le k \le n} \Delta(C_k)}$$

in which $\Delta(C_i, C_j)$ represents the inter-cluster distance between cluster C_i and C_j, $\Delta(C_k)$ represents the intra-cluster distance of cluster C_k. We used the "MaxStep" distance developed by Habib et al to calculate the distance for Dunn index[55]. Finally, we iterated all possible (ρ_0, δ_0) combinations that yield different clusters and calculated its Dunn index. The clustering result with the highest Dunn index was chosen as final cluster.

- Step 7. "Shuffled" cells. Due to the limited genome coverage of each single cell, cells may cluster according to their sequencing depth rather than 'true' co-variation[12]. To

verify that our cluster results are not driven by such artefacts, we compared our results to a simulated data set. For this data set in which binary accessible sites within each cell were randomly shuffled across all accessible sites. In other words, we shuffled the data and removed the biological significance, but maintained the distribution of sequencing depth across cells. “Shuffled” cells were uniformly distributed as a “ball” in the embedded 3-D space without clear partition of cells. However, we did observe that there is a small portion of cells that tend to form a cluster but did not pass the cut-off (ρ_0, δ_0) used for the P56 forebrain data set[12].

Identification of cluster-specific features. We next developed a computational method which combines stability selection with LASSO[57] to identify genomic elements (features) that potentially distinguish cells belonging to different clusters. LASSO regression enables sparse feature selections through the use of L1 penalty. However, LASSO regression often does not result in a robust set of selected features and is sensitive to data perturbation. This is especially true when features are correlated. To overcome these limitations, we adopted stable lasso to robustly identify features that distinguish every two cell clusters (Algorithm 2). Finally, we combined all identified features that distinguish different cell types to identify genomic elements (features) that potentially distinguish cells belonging to different clusters.

Bulk ATAC-seq data analysis. Paired-end sequencing reads were aligned to the mm10 reference genome using Bowtie2 in paired-end mode with following parameters “bowtie2 -p 5 -t -X2000 --no-mixed --no-discordant[53] and PCR duplicates were removed

using samtools[54]. Next, mitochondrial reads were removed and the position of alignments adjusted[58]. For visualization the *bamCoverage* utility from deepTools2 was used[59].

Hierarchical clustering of ATAC-seq profiles in adult forebrain. DeepTools2 was used for correlation analysis and hierarchical clustering of ATAC-seq profiles from cell clusters and sorted cell-types in the adult forebrain[59]. First, we computed read coverage for each data set against the merged list of genomic elements that separate two cell clusters in the adult forebrain using the *multiBamSummary* utility. Next we used *plotCorrelation* to generate hierarchical clustering using Spearman correlation coefficient between two clusters[59].

Accessibility analysis and clustering of genomic elements. To cluster genomic elements based on their accessibility profile we used these promoter distal elements that were capable to distinguish two cell clusters. For each feature we extended the summits identified by MACS2[60] in both directions by 250 bp and generated a union set of elements using *mergeBED* functionality of BEDTools v2.17.0[56]. Next, we intersected cluster specific bam files with the peak list using the *coverageBED* functionality of BEDTools v2.17.0[56]. We discarded elements that had less than five reads on average. After adding a pseudocount of one we calculated cluster-specific RPM (reads per million sequenced reads) values for each genomic element. We divided the RPM value for a given cluster by the average value of all clusters (fold over mean) and finally log2 transformed the data. The generated matrix was used for k-means clustering of the elements using Ward's method. We performed this analysis for all adult clusters, the excitatory neuron clusters

and the 12 developmental cell clusters, respectively. To compare clusters of genomic elements in the adult forebrain with previously described single cell DNA methylation data[29], we calculated the fraction of cell-type specific differentially methylated regions (DMR) with each cluster using intersectBED functionality of BEDTools v2.17.0[56] and normalized it by the total number of elements. Since Luo et al.[29] focused on frontal cortex and specifically purified neurons, we centered the comparison on clusters associated with excitatory an inhibitory neuron.

Motif enrichment analysis. To identify potential regulators of chromatin accessibility we performed motif analysis using the AME utility of the MEME suite[49]. For enrichment of known motifs, one-tailed Fisher's Exact test was used to calculate significance. P-values were corrected by the Bonferroni method for multiple testing. A P-Value cut-off of $< 10^{-5}$ was chosen for known motifs from the JASPAR database (JASPAR_CORE_2016_vertebrates.meme)[61]. For identification of *de novo* motifs HOMER tools was used with default settings[50].

Annotation of genomic elements. The GREAT algorithm was used to annotate distal genomic elements using following settings to define the regulatory region of a gene: Basal+extension (constitutive 1 kb upstream and 0.1 kb downstream, up to 500 kb max extension)[32]. Gene ontology categories “Molecular Function” and “Biological Processes” were used.

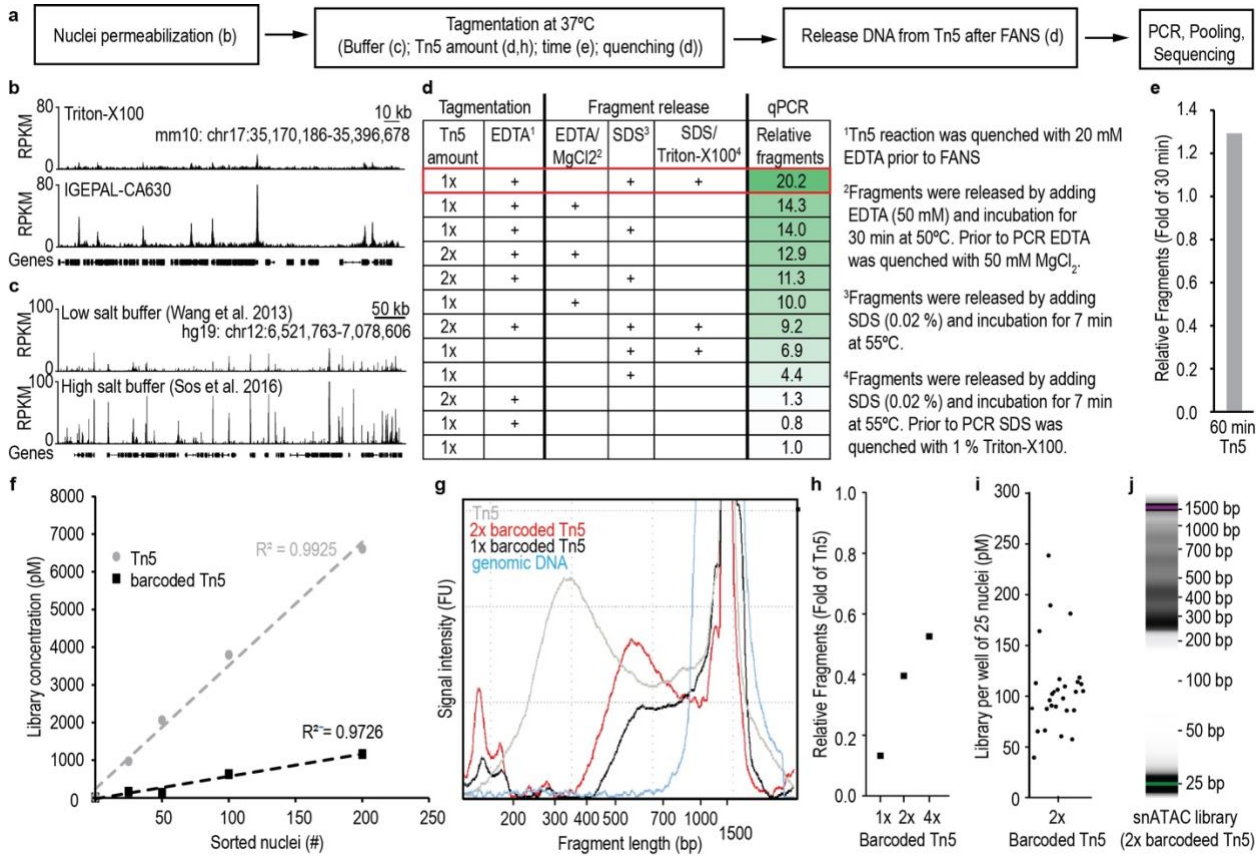
Analysis of dynamic chromatin accessibility within a cell cluster. First, the ATAC-seq reads were counted in all peaks for each stage, cell type and replicate. For each cell cluster, only stages with more than 250,000 reads overlapping ATAC-seq peaks and more than 50 nuclei were used for dynamic analysis. Peaks with greater than 1 read per million reads (RPM) in at least 2 samples were kept. We used edgeR[62] to assess the significance of difference between adjacent stages for cell clusters with at least 4 out of 7 stages passing filtering criteria. P-values were corrected using the Bonferroni method. Peaks with a Bonferroni p-value less than 0.05 were called dynamic peaks. The total number of dynamic peaks in each cell type are listed in (**Figure S1.11c**). For each cell type, the read counts in each peak were normalized into a unit vector (i.e values were divided by the square root of the sum of the squares of the values). K-means was used for clustering of cell clusters with more than 200 dynamic elements (K=3). Motif enrichment analysis was performed for each peak cluster using HOMER[50].

VISTA analysis. Genomic locations of 484 VISTA validated elements[45] were downloaded from https://enhancer.lbl.gov using the search term “forebrain”. Genomic locations were converted from mm9 to mm10 using the *liftOver* tool (minimum rematch ratio of 0.95). 91 of these were showed specific activity in the subpallium[46]. To identify developmental clusters that are enriched for subpallial enhancers we first calculated the ratio of elements per k-means cluster overlapping with the total forebrain enhancer list and the subpallial subset separately. Finally, we calculated the relative enrichment using the ratio of subpallial over the complete forebrain regions. For anatomical annotation of distinct clusters, we intersected these regions with enhancers that are active in specific

areas in the developing mouse forebrain[47]. After filtering clusters with less than 5 overlapping regions, we performed a binomial test to identify anatomical regions enriched for each cluster. The enrichment score is defined as - log10(binomial P-value).

1.9 Supplementary Figures

Figure S1.1. SnATAC-seq protocol optimization. (**a**) Overview of critical steps for the snATAC-seq procedure for nuclei from frozen tissues. (**b**) IGEPAL-CA630 but not Triton-X100 was sufficient for tagmentation of frozen tissues (n = 1 experiment). (**c**) Tagmentation was facilitated by high salt concentrations in reaction buffer (n = 1 experiment; Wang, Q. *et al. Nature protocols, 2013,* doi:10.1038/nprot.2013.118: Sos, B. C. *et al. Genome biology*, 2016, doi:10.1186/s13059-016-0882-7). (**d**) Maximum number of fragments per nucleus could be recovered when quenching Tn5 by EDTA prior to FANS and denaturation of Tn5 after FANS by SDS. Finally, SDS was quenched by Triton-X100 to allow efficient PCR amplification. (**e**) Increasing tagmentation time from 30 min to 60 min can result in more DNA fragments per nucleus (n = 1 experiment). (**f**) Number of sorted nuclei was highly correlated with the final library concentration. Tn5 loaded with barcoded adapters showed less efficient tagmentation as compared to Tn5 without barcodes. Wells were amplified for 13 cycles, purified and libraries quantified by qPCR using standards with known molarity (n = 1 experiment). (**g**) Tagmentation with barcoded Tn5 was less efficient and resulted in larger fragments than Tn5 (550 bp vs. 300 bp). Ratio for barcoded Tn5 was based on concentration of regular Tn5. (**h**) Doubling the concentration of barcoded Tn5 increased the number of fragments per nucleus by 3-fold. Further increase resulted only in minor improvements (n = 1 experiment). (**i**) Dot blot illustrating the amount of library from 25 nuclei per well. Each well was amplified for 11 cycles and quantified by qPCR. This output was used to calculate the number of required PCR cycles for snATAC-seq libraries to prevent overamplification (n = 28 wells). (**j**) Size distribution of a successful snATAC-seq library from a mixture of E15.5 forebrain and GM12878 cells shows a nucleosomal pattern. SnATAC-seq was performed including all the optimization steps described above with barcoded Tn5 in 96 well format (n = 1 experiment; snATAC libraries for forebrain samples showed comparable nucleosomal patterns: n = 16 experiments).


a
Nuclei permeabilization (b)
Tagmentation at 37ºC
(Buffer (c); Tn5 amount (d,h); time (e); quenching (d))
Release DNA from Tn5 after FANS (d)
PCR, Pooling, Sequencing
b
Triton-X100
10 kb
mm10: chr17:35,170,186-35,396,678
IGEPAL-CA630
RPKM
Genes
c
Low salt buffer (Wang et al. 2013)
50 kb
hg19: chr12:6,521,763-7,078,606
High salt buffer (Sos et al. 2016)
d
Tagmentation
Fragment release
qPCR
Tn5 amount
EDTA[1]
EDTA/ MgCl2[2]
SDS[3]
SDS/ Triton-X100[4]
Relative fragments
1x + + + 20.2
1x + + 14.3
1x + + 14.0
2x + + 12.9
2x + + 11.3
1x + 10.0
2x + + + 9.2
1x + + 6.9
1x + 4.4
2x + 1.3
1x + 0.8
1x 1.0
[1]Tn5 reaction was quenched with 20 mM EDTA prior to FANS
[2]Fragments were released by adding EDTA (50 mM) and incubation for 30 min at 50ºC. Prior to PCR EDTA was quenched with 50 mM MgCl2.
[3]Fragments were released by adding SDS (0.02 %) and incubation for 7 min at 55ºC.
[4]Fragments were released by adding SDS (0.02 %) and incubation for 7 min at 55ºC. Prior to PCR SDS was quenched with 1 % Triton-X100.
e
Relative Fragments (Fold of 30 min)
60 min Tn5
f
Library concentration (pM)
Tn5
barcoded Tn5
R² = 0.9925
R²= 0.9726
Sorted nuclei (#)
g
Tn5
2x barcoded Tn5
1x barcoded Tn5
genomic DNA
Signal intensity (FU)
Fragment length (bp)
h
Relative Fragments (Fold of Tn5)
1x 2x 4x
Barcoded Tn5
i
Library per well of 25 nuclei (pM)
2x
Barcoded Tn5
j
1500 bp
1000 bp
700 bp
500 bp
400 bp
300 bp
200 bp
100 bp
50 bp
25 bp
snATAC library
(2x barcodeed Tn5)

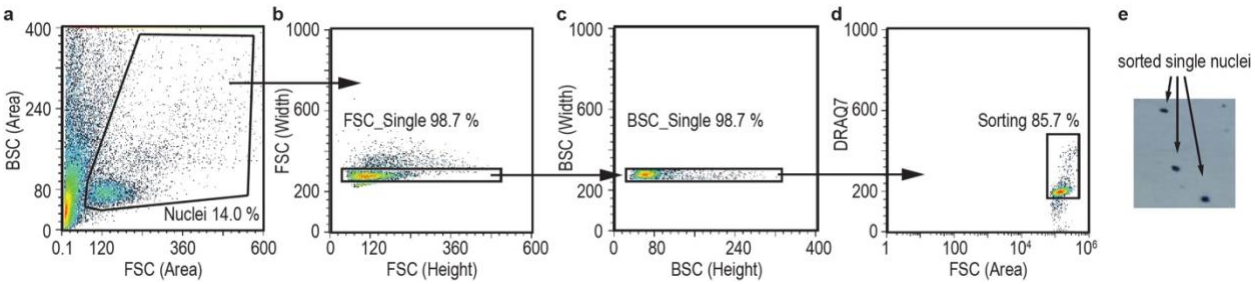


Figure S1.2. Isolation of single nuclei after tagmentation. (**a-d**) Density plots illustrating the gating strategy for single nuclei. First, big particles were identified (**a**), then duplicates were removed (**b**, **c**) and finally, nuclei were sorted based on high DRAQ7 signal (**d**), which stains DNA in nuclei. (**e**) Verification of single cell suspension after FANS was done with Trypan Blue staining under a microscope.

Figure S1.3. Overview of snATAC-seq sequencing data and quality filtering for single nuclei. (**a**) Distribution of insert sizes between reads pairs derived from sequencing of snATAC-seq libraries indicates nucleosomal patterning. (**b**) Individual barcode representation in the final library shows variability between barcodes. (**c**) To assess the probability of two nuclei sharing the same nuclei barcode, single nuclei ATAC-seq was performed on a 1:1 mixture of human GM12878 cells and mouse E15.5 forebrain nuclei. A collision was indicated by < 90% of all reads mapping to either the mouse genome (mm9) or the human genome (hg19). We identified 8.2% of these barcode collision events. (**d**) Read coverage per barcode combination after removal of potential barcodes with less than 1,000 reads. (**e**) Constitutive promoter coverage for each single cell. The red line indicates the constitutive promoter coverage in corresponding bulk ATAC-seq data sets from the same biological sample. Cells with less coverage than the bulk ATAC-seq data set were discarded. (**f**) Fraction of reads falling into peaks for each single nucleus. The red line indicates fraction of reads in peak regions in corresponding bulk ATAC-seq data sets from the same biological sample. Nuclei with lower reads in peak ratios coverage than the bulk ATAC-seq data set were discarded from downstream analysis.

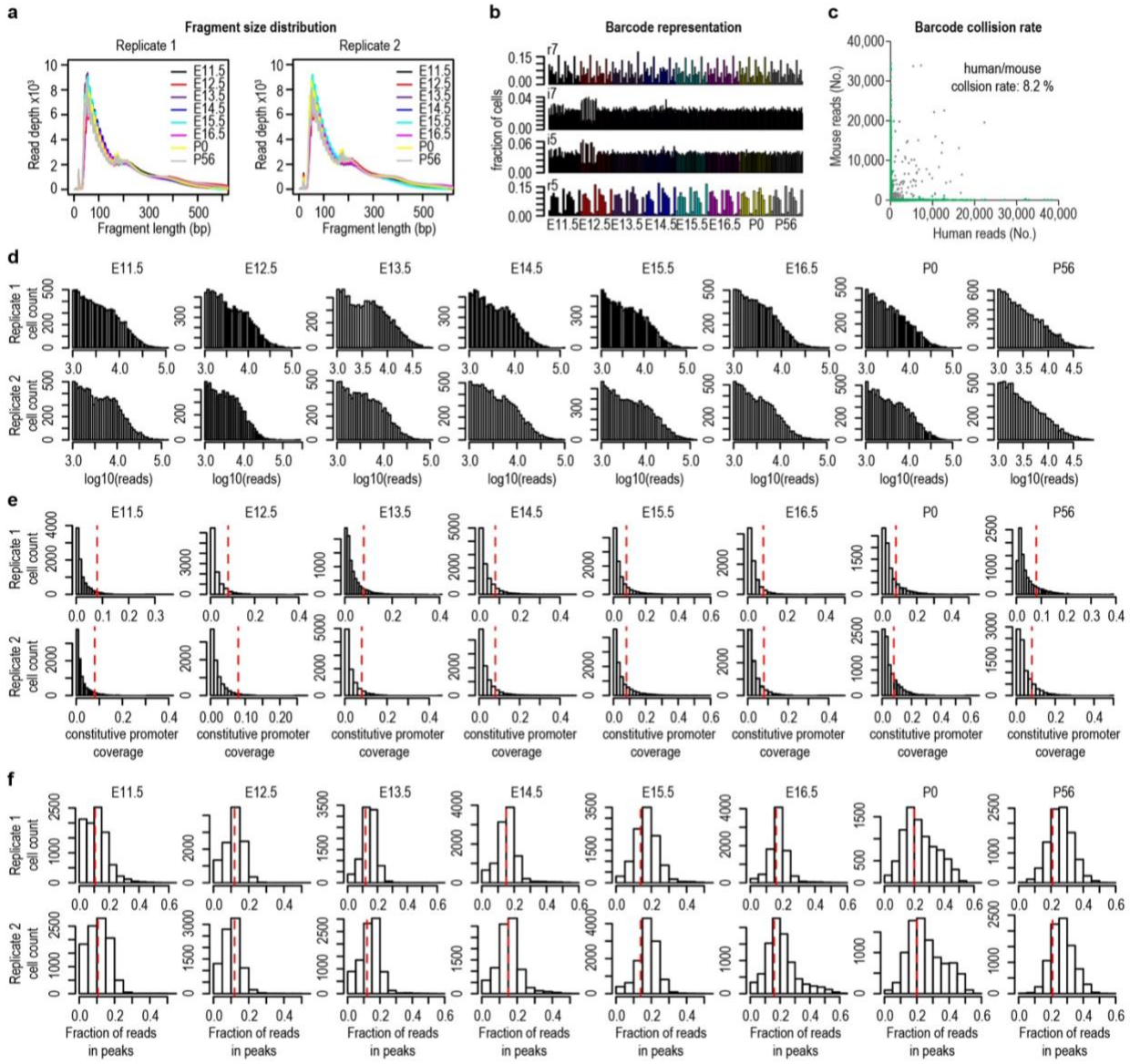

a
Fragment size distribution
Replicate 1
Replicate 2
Read depth x10³
Fragment length (bp)
E11.5
E12.5
E13.5
E14.5
E15.5
E16.5
P0
P56
b
Barcode representation
fraction of cells
r7
r5
c
Barcode collision rate
human/mouse
collsion rate: 8.2 %
Mouse reads (No.)
Human reads (No.)
d
Replicate 1
Replicate 2
cell count
log10(reads)
e
constitutive promoter
coverage
f
Fraction of reads
in peaks

Figure S1.4. SnATAC-seq data sets are robust and reproducible. Pearson correlation of chromatin accessibility profiles from two independent experiments derived from bulk ATAC-seq (left column) and from aggregate snATAC-seq after aggregating single nuclei profiles (middle column) is shown in each plot. In the right column the correlation between bulk ATAC-seq and aggregate snATAC-seq are displayed for the experiment on the first set of forebrain tissues. Data are displayed from forebrain tissues from following time points: **a.** E11.5, **b.** E12.5, **c.** E13.5, **d.** E14.5, **e.** E15.5, **f.** E16.5, **g.** P0, and **h.** P56. For bulk ATAC-seq data generated by the ENCODE consortium were processed.

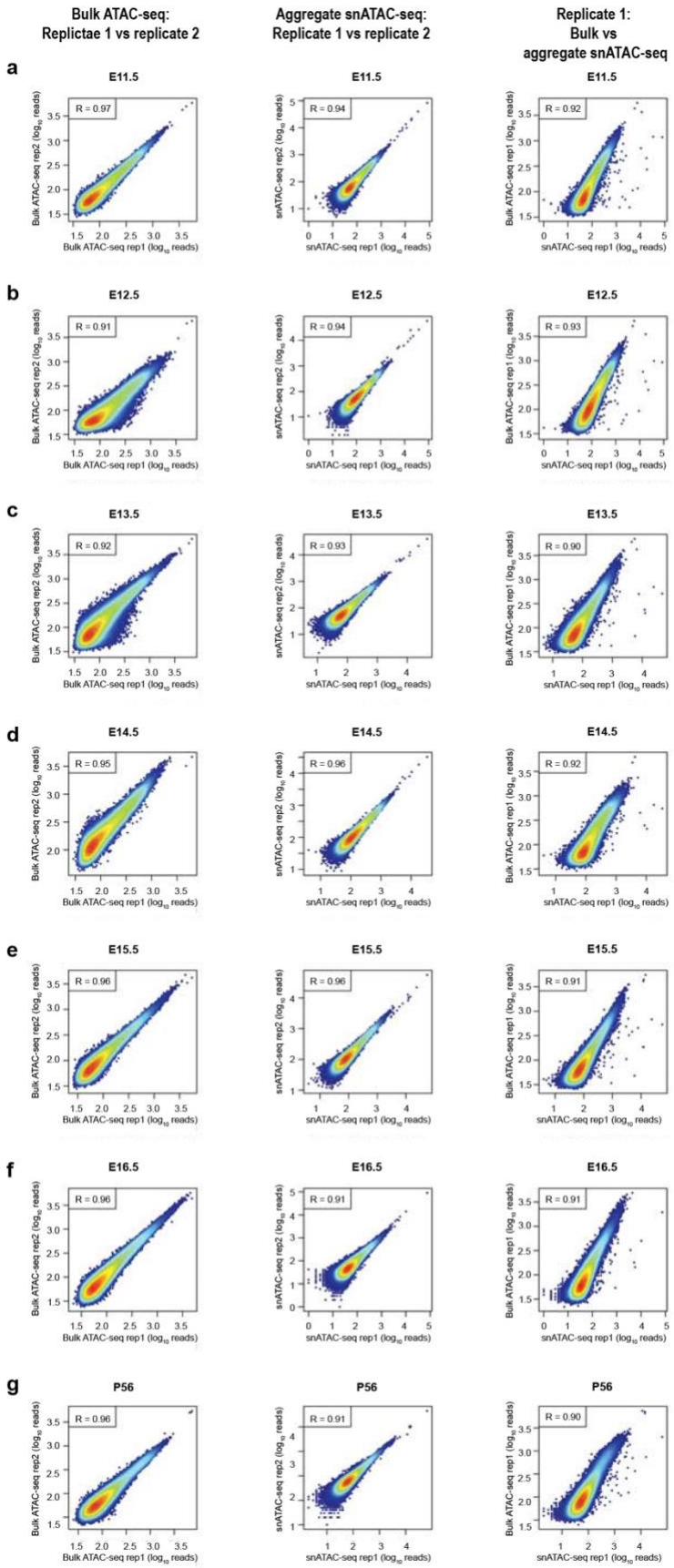
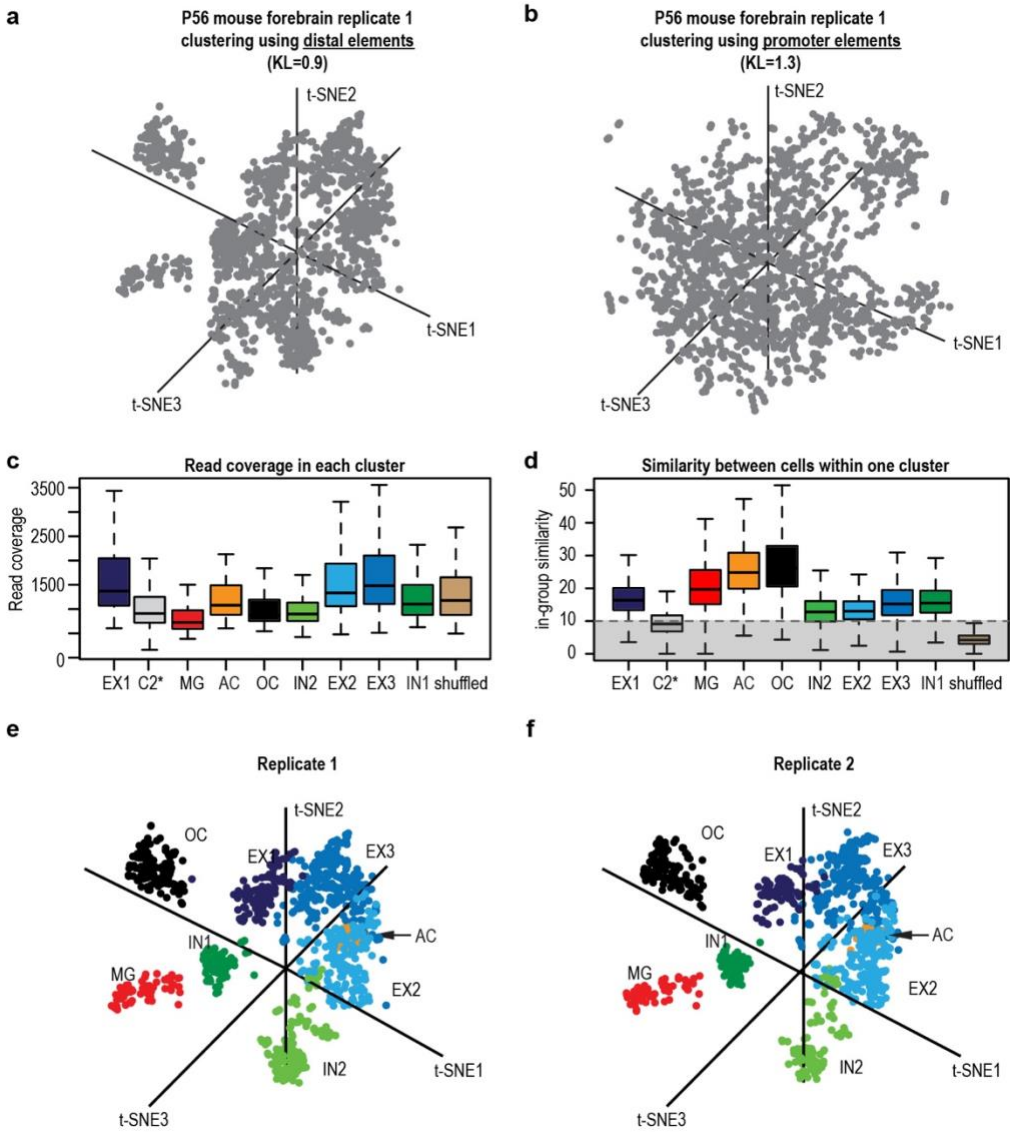

Bulk ATAC-seq:
Replictae 1 vs replicate 2
Aggregate snATAC-seq:
Replicate 1 vs replicate 2
Replicate 1:
Bulk vs
aggregate snATAC-seq
a
E11.5
R = 0.97
R = 0.94
R = 0.92
b
E12.5
R = 0.91
R = 0.94
R = 0.93
c
E13.5
R = 0.92
R = 0.93
R = 0.90
d
E14.5
R = 0.95
R = 0.96
R = 0.92
e
E15.5
R = 0.96
R = 0.96
R = 0.91
f
E16.5
R = 0.96
R = 0.91
R = 0.91
g
P56
R = 0.96
R = 0.91
R = 0.90
Bulk ATAC-seq rep2 (log10 reads)
Bulk ATAC-seq rep1 (log10 reads)
snATAC-seq rep2 (log10 reads)
snATAC-seq rep1 (log10 reads)

Figure S1.5. Clustering strategies, quality control of clusters and clustering result for individual experiments in adult forebrain. (**a, b**) T-SNE visualization of clustering using (**a**) distal element (regions outside 2 kb of refSeq transcriptional start sites) or (**b**) promoter regions (KL: Kullback-Leibler divergence reported by t-SNE). **c.** Box plot of read coverage for each cluster (sample size for cluster is EX1: 190, C2: 946, MG: 126, AC: 120, OC: 252, IN2: 320, EX2: 366, EX3: 519, IN1: 195, shuffled: 199; 25% quantile is EX1: 1076, C2: 665, MG: 595, AC: 884.25, OC: 755, IN2: 754, EX2: 106, EX3: 1104, IN1: 881, shuffled: 880; median value is EX1: 1372, C2: 855, MG: 726, AC: 1079, OC: 871, IN2: 899, EX2: 1334, EX3: 1482, IN1: 1102, shuffled: 1178; 75% quantile is EX1: 2045, C2: 1196, MG: 972, AC: 1489, OC: 1188, IN2: 1134, EX2: 1929, EX3: 2102, IN1: 1496, shuffled: 1652). (**d**) Box plot of similarity analysis between any two given cells in a cluster. Cluster C2 was discarded before downstream analysis due to low its intra-group similarity (median < 10). As a negative control, randomly shuffled cells were included in the analysis displaying exceptionally low in-group similarity (sample size is EX1: 190, C2:946, MG:126, AC:120, OC: 252, IN2: 320, EX2: 366, EX3: 519, IN1: 195, shuffled: 199; 25% quantile is EX1: 13.34, C2: 6.84, MG: 15.15, AC: 19.89, OC: 20.60, IN2: 9.88, EX2: 10.53, EX3: 11.81, IN1: 12.58, shuffled: 3.02; median is EX1: 16.34, C2: 9.12, MG: 19.68, AC: 24.835, OC: 26.23, IN2: 12.77, EX2: 13.00, EX3: 15.23, IN1: 15.50, shuffled: 4.20; 75% quantile is EX1: 20.07, C2: 11.74, MG: 25.58, AC: 30.860, OC: 32.95, IN2: 16.11, EX2: 16.02, EX3: 19.46, IN1: 19.25, shuffled: 5.56). (**e, f**) T-SNE visualization of single cells from (**e**) replicate 1 and (**f**) replicate 2. The projection and color coding are the same as in **Figure 1.2d**.


a
P56 mouse forebrain replicate 1
clustering using distal elements
(KL=0.9)
t-SNE2
t-SNE1
t-SNE3
b
P56 mouse forebrain replicate 1
clustering using promoter elements
(KL=1.3)
t-SNE2
t-SNE1
t-SNE3
c
Read coverage in each cluster
Read coverage
3500
2500
1500
0
EX1 C2* MG AC OC IN2 EX2 EX3 IN1 shuffled
d
Similarity between cells within one cluster
in-group similarity
50
40
30
20
10
0
EX1 C2* MG AC OC IN2 EX2 EX3 IN1 shuffled
e
Replicate 1
t-SNE2
OC
EX1
EX3
AC
IN1
MG
EX2
IN2
t-SNE1
t-SNE3
f
Replicate 2
t-SNE2
OC
EX1
EX3
AC
IN1
MG
EX2
IN2
t-SNE1
t-SNE3

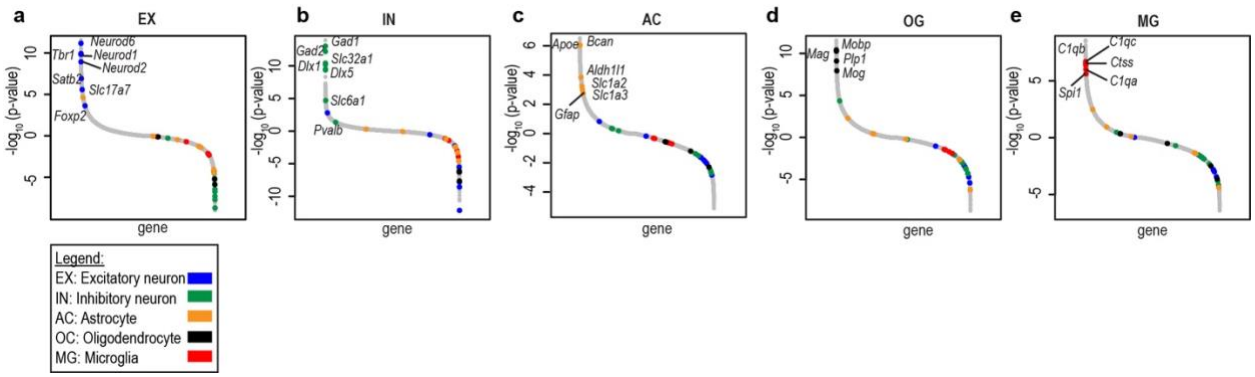


Figure S1.6. Ranking of gene loci (TSS ± 10kb) compared to other clusters in adult forebrain. Negative binomial test shows enrichment for (**a**) excitatory neuron markers (**b**) inhibitory neuron markers (**c**) astrocyte markers (**d**) oligodendrocyte markers and (**e**) microglia markers extending the examples shown in **Figure 1.2b**. Please note for general assignment accessibility profiles for Ex1-3 and IN1/2 were merged, respectively. For each cell type, data from two experiments (n = 2) were used to carry out the negative binomial test.

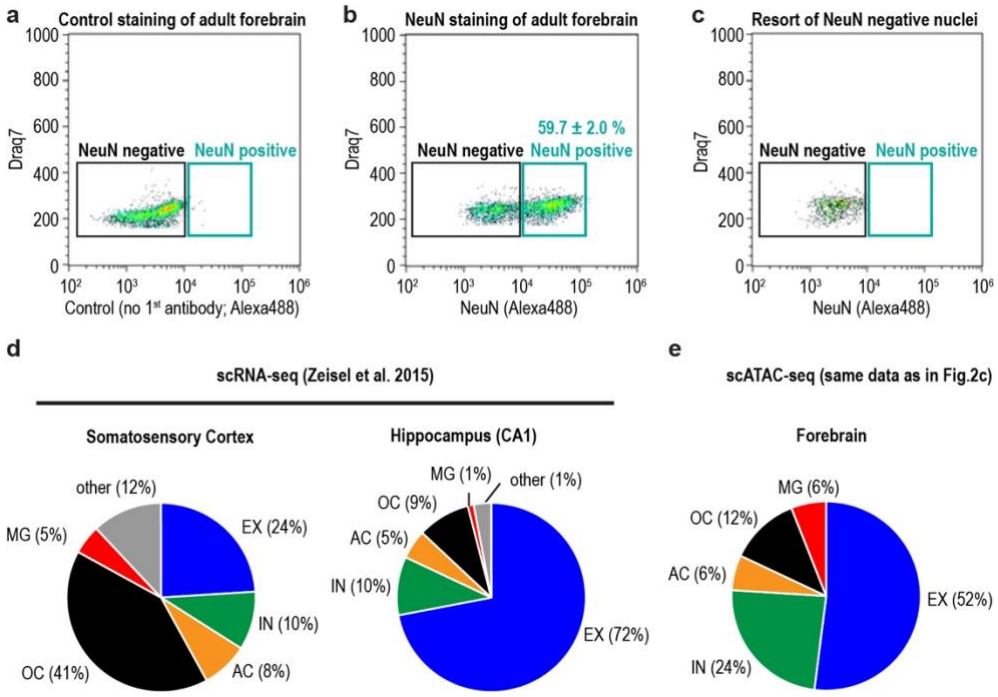


Figure S1.7: Flow cytometric analysis of adult mouse forebrain and comparison to single cell RNA-seq data from different brain regions. **a-c** Dot blots illustrating nuclei from adult forebrain stained for flow cytometry with Alexa488 conjugated secondary antibodies. (**a**) Displayed are representative blots for experiments without antigen specific primary antibody and (**b**) with antibodies recognizing the post-mitotic neuron marker NeuN[22] (n = 3, average ± SEM). (**c**) NeuN negative nuclei were sorted for ATAC-seq experiments and purity (> 98%) was confirmed by flow cytometry of the sorted population. (**d**) Relative composition of different forebrain regions derived from single cell RNA-seq shows region specific differences[19]. (**e**) Relative composition derived from snATAC-seq (compare to **Figure 1.2c**) of adult forebrain shows values in between.

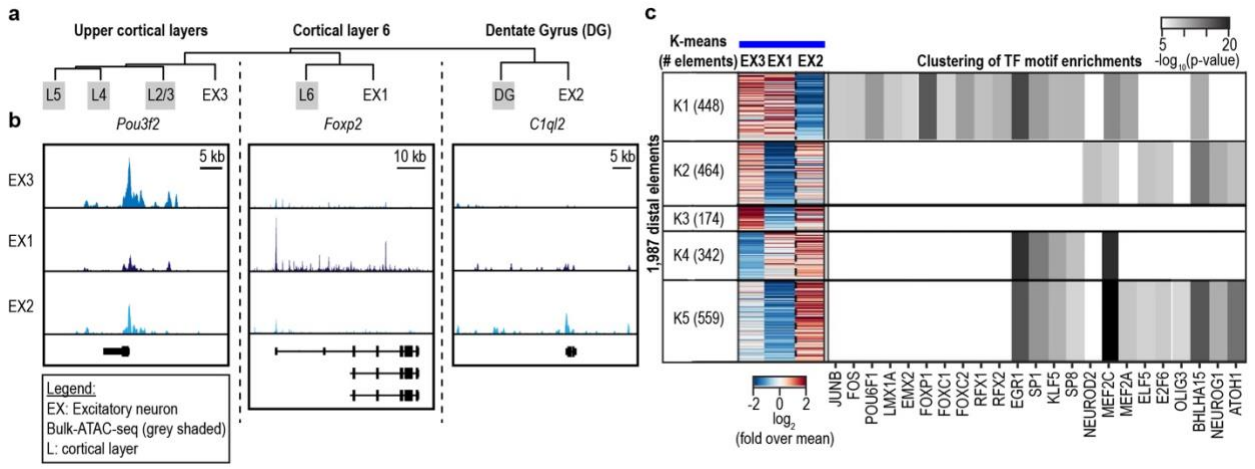


Figure S1.8. Sub-classification of excitatory neurons into hippocampal and cortical neuron types. (**a**) Hierarchical clustering of aggregate single cell data for excitatory neuron cluster and sorted bulk data sets corresponding to different anatomical regions (grey shaded). (**b**) Chromatin accessibility at marker gene loci. (**c**) K-means clustering of promoter distal genomic elements and enrichment analysis for transcription factor motifs. Statistical test for motif enrichment: One-tailed Fisher's Exact test; displayed p-values are Bonferroni corrected for multiple testing[59].

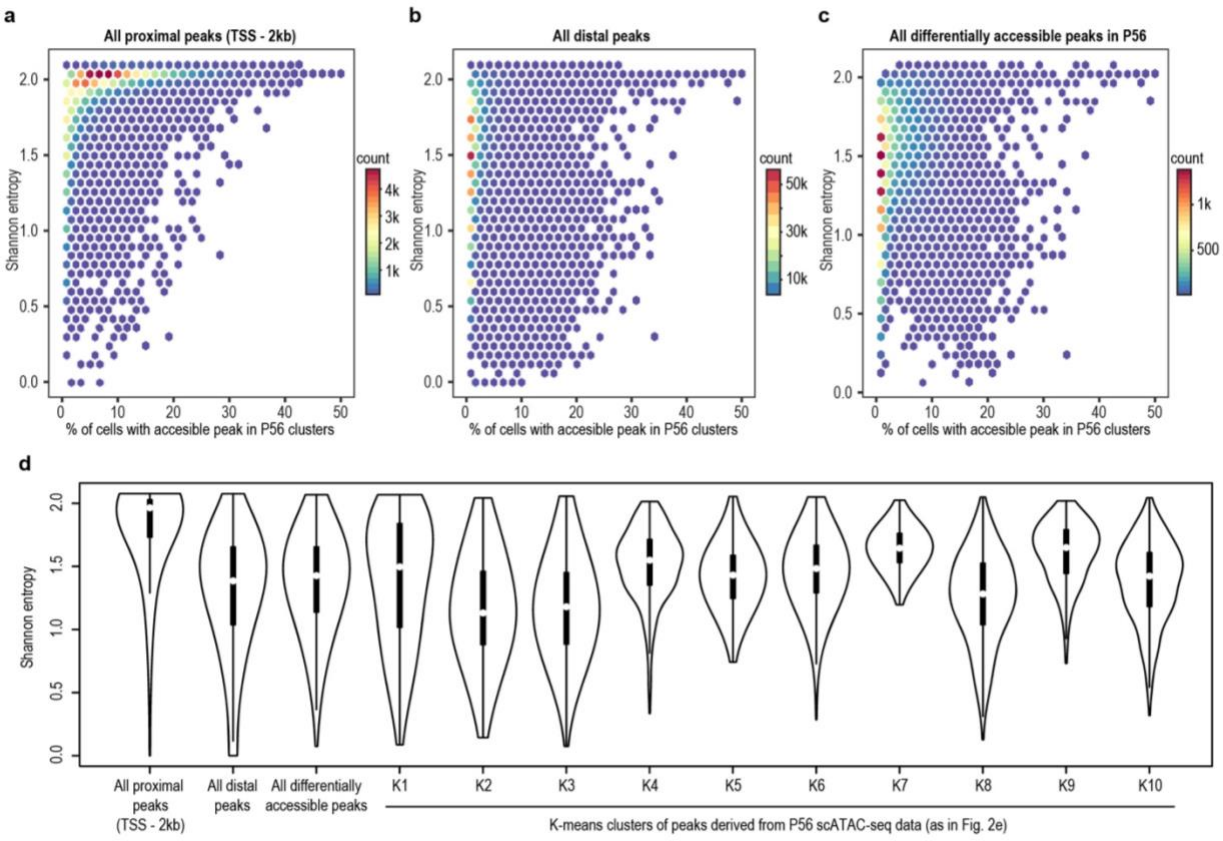


Figure S1.9. Cell-type specificity and coverage of the cis elements. (**a-c**) Graphs illustrate cell-type specificity of genomic elements as measured by Shannon entropy based on normalized read counts for each cell-type and percentage of nuclei in which a genomic element was called accessible as indicated by presence of at least 1 read overlapping with the element a peak. Analysis was performed for the adult forebrain (P56) against (**a**) TSS-proximal genomic elements (TSS - 2kb), (**b**) distal elements and (**c**) the subset of genomic elements that separated two cell clusters. **d.** Violin plots illustrate higher cell-type specificity for distal elements compared to proximal elements indicated by significantly lower Shannon entropy value ($p < 2.2e-16$). In addition, all genomic elements that separate two clusters as well as subsets identified from k-means clustering of genomic elements depending on chromatin accessibility in adult forebrain are displayed (related to **Figure 1.2e**). (all proximal peaks n = 14,262 (minimum/median/maximum; 0/1.96/2.08), all distal peaks n = 140,102 (0/1.38/2.08), all differentially accessible peaks n = 4,980 (0.07/1.4/2.06), K1 n = 529 (0.08/1.49/2.06), K2 n = 586 (0.14/1.13/2.04), K3 n = 737 (0.07/1.18/2.05), K4 n = 270 (0.33/1.55/2.01), K5 n = 601 (0.74/1.43/2.05), K6 n = 513 (0.28/1.48/2.05), K7 n = 538 (1.19/1.64/2.02), K8 n = 490 (0.13/1.28/2.05), K9 n = 282 (0.73/1.65/2.02), K10 n = 434 (0.32/1.42/2.04). TSS: transcriptional start site.

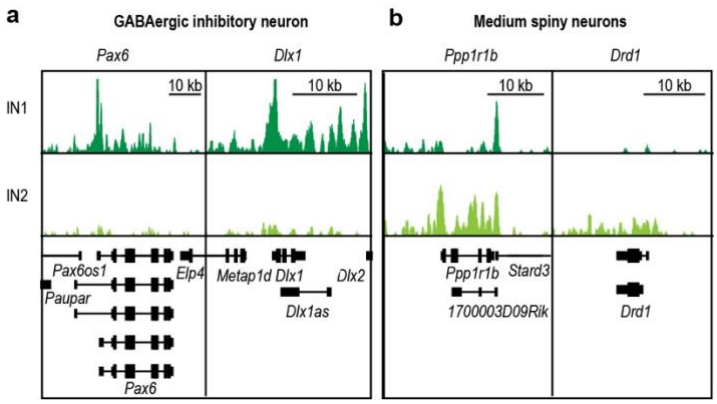


Figure S1.10. Distinct chromatin accessibility profiles of two GABAergic neuron clusters. IN2 is depleted for chromatin accessibility at the genes *Pax6* and *Dlx1* (**a**) but enriched for marker genes of medium spiny neurons as compared to IN1 cluster (**b**).

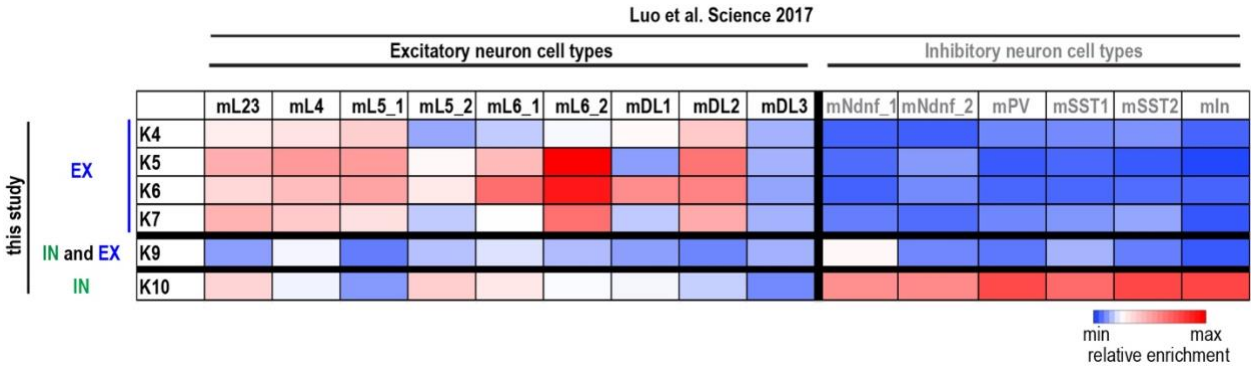


Figure S1.11. Comparison of chromatin accessibility and differentially methylated regions in neuronal subtypes. Displayed is the fraction of cell-type specific differentially methylated[29] that overlapped with genomic elements accessible in excitatory (EX) and inhibitory neurons (IN). This analysis illustrates that cis regulatory elements specific for inhibitory neurons and excitatory neurons, respectively, could be identified by both methods. Clusters (K) from this study are the same as in **Figure 1.2e** (m: mouse; L: layer; DL: deep layer).

Figure S1.12. Dynamics of chromatin accessibility within distinct cell groups. (**a**) Number of reads in peaks per developmental time point for a specific nuclei cluster. (**b**) Number of nuclei per time point for a specific nuclei cluster. For analysis of dynamics only cell clusters with > 3 stages with > 50 nuclei and > 250,000 reads in peaks were considered. (**c**) Overview of dynamic elements identified per cell cluster (see **supplementary methods**) (**d-g**)**.** K-means clustering and motif enrichment analysis for nuclei clusters with > 200 dynamic genomic elements. Statistical test for motif enrichment: hypergeometric test. P-values were not corrected for multiple testing[50]. (e: embryonic; RG: Radial glia; EX: Excitatory neuron; IN: Inhibitory neuron; EMP: Erythromyeloid progenitor cell; AC: Astrocyte).

a Reads in peaks per cell cluster at each time point

Cell cluster \ Time point	E11.5	e12.5	e13.5	e14.5	e15.5	e16.5	p0
RG1	1,077,714	925,916	253,274	166,469	125,898	20,118	1,401
RG2	34,368	212,006	767,951	1,262,027	1,186,931	356,975	7,909
eEX1	161,911	208,487	668,184	1,320,188	2,114,083	506,033	265,608
eEX2	13,135	67,121	382,525	953,300	4,811,953	858,302	1,550,411
RG3	2,512,140	1,759,240	697,854	446,290	448,451	56,178	39,685
RG4	122,932	239,670	378,461	463,678	336,554	84,909	51,263
eIN1	475,169	918,028	1,211,824	1,164,612	2,916,576	300,165	762,899
eIN2	81,280	203,870	323,656	474,479	587,100	387,242	346,968
eIN3	15,504	50,410	51,024	163,359	751,039	155,228	291,443
eIN4	23,537	83,326	275,184	400,720	935,362	285,780	320,223
eAC	0	0	35,917	79,090	148,759	79,918	1,168,309
EMP	174,341	49,393	20,082	11,265	14,659	5,549	6,050

b Nuclei per cell cluster at each time point

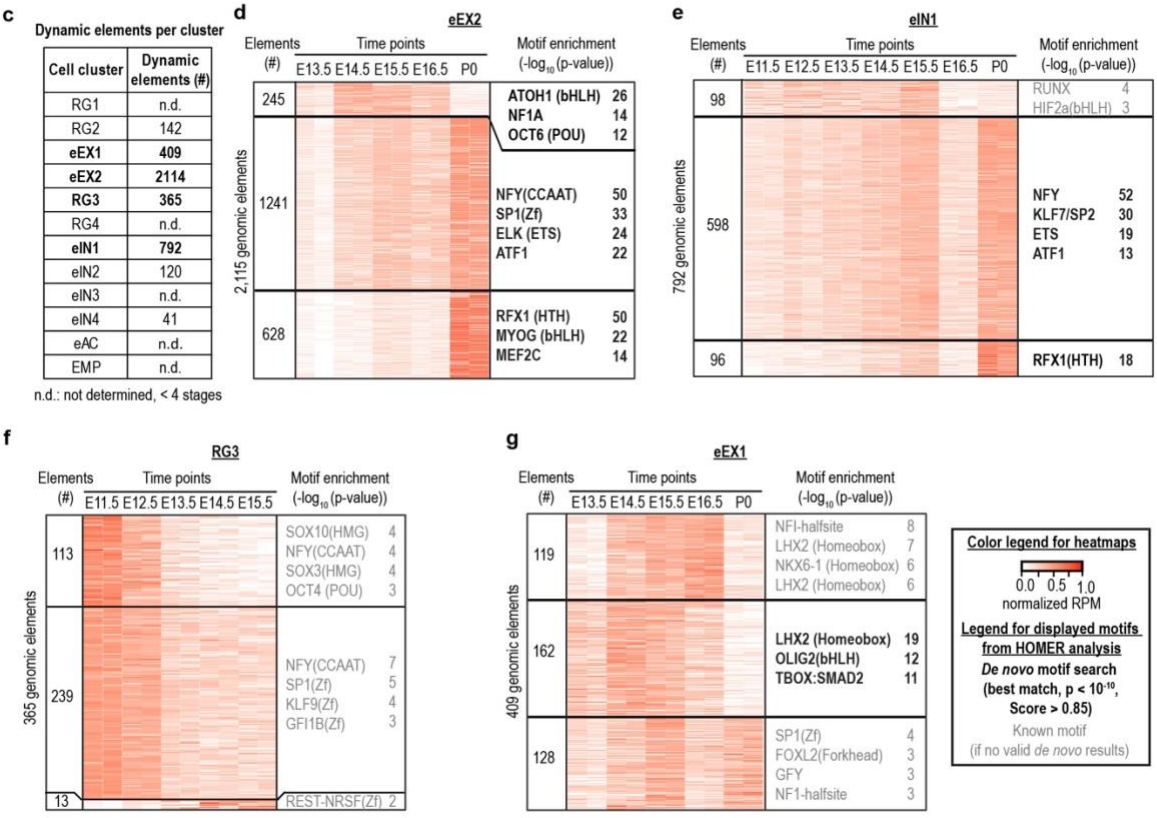
Cell cluster \ Time point	e11.5	e12.5	e13.5	e14.5	e15.5	e16.5	p0
RG1	216	189	62	49	41	9	1
RG2	7	32	161	312	303	109	3
eEX1	36	47	141	309	427	138	63
eEX2	4	14	87	290	1274	308	744
RG3	661	420	191	129	145	22	26
RG4	30	51	92	119	100	31	25
eIN1	137	224	370	354	854	118	412
eIN2	19	45	73	133	168	134	186
eIN3	4	11	14	44	184	53	165
eIN4	7	22	83	117	281	103	168
eAC	0	0	8	19	39	29	575
EMP	92	30	16	7	11	2	5

Color code for A, B: > 50 nuclei | > 50 nuclei and > 250,000 reads in peaks

c Dynamic elements per cluster

Cell cluster	Dynamic elements (#)
RG1	n.d.
RG2	142
eEX1	**409**
eEX2	**2114**
RG3	**365**
RG4	n.d.
eIN1	**792**
eIN2	120
eIN3	n.d.
eIN4	41
eAC	n.d.
EMP	n.d.

n.d.: not determined, < 4 stages



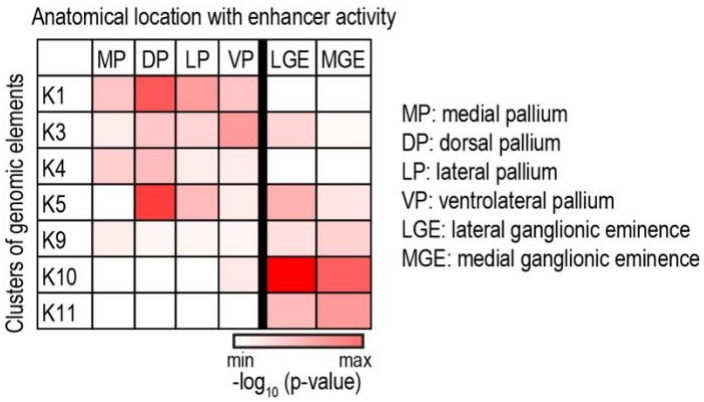


Figure S1.13. Distal genomic element clusters are associated with distinct anatomical locations in the developing forebrain. Displayed is the enrichment of clusters of open chromatin for enhancers that are active in distinct regions of the developing forebrain (n = 95)[47]. As expected, elements mainly associated with radial glia and excitatory neuron cell-types (**Figure 1.2e**, K1, 3, 4) were enriched for pallial subregions, whereas inhibitory neuron associated elements **(Figure 1.2e**, K9-11) were enriched in LGE and MGE regions. Clusters with less than 5 overlapping elements were excluded from the analysis. Binomial testing was used for statistical analysis. The p-values were not corrected. Anatomically annotated enhancers: n = 146[47]; open chromatin regions: K1: n = 880; K3: n = 1838; K4: n = 1015; K5: n = 1276; K9: n = 1042; K10: n = 1238; K11: n = 623.

1.11 References

1. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).

2. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

3. Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R. & Stamatoyannopoulos, J. A. Large-scale identification of sequence variants influencing human transcription factor occupancy in vivo. Nature Genetics 47, 1393–1401 (2015).

4. Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. & Stamatoyannopoulos, J. A. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).

5. Wu, J., Huang, B., Chen, H., Yin, Q., Liu, Y., Xiang, Y., Zhang, B., Liu, B., Wang, Q., Xia, W., Li, W., Li, Y., Ma, J., Peng, X., Zheng, H., Ming, J., Zhang, W., Zhang, J., Tian, G., Xu, F., Chang, Z., Na, J., Yang, X. & Xie, W. The landscape of accessible chromatin in mammalian preimplantation embryos. Nature 534, 652–657 (2016).

6. Mo, A., Mukamel, E. A., Davis, F. P., Luo, C., Henry, G. L., Picard, S., Urich, M. A., Nery, J. R., Sejnowski, T. J., Lister, R., Eddy, S. R., Ecker, J. R. & Nathans, J.

Epigenomic Signatures of Neuronal Diversity in the Mammalian Brain. Neuron 86, 1369–1384 (2015).

7. The Mouse ENCODE Consortium, Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., Shen, Y., Pervouchine, D. D., Djebali, S., Thurman, R. E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G. K., Williams, B. A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.-H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M. A., Zhang, M., Byron, R., Groudine, M. T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.-C., Rasmussen, M. D., Bansal, M. S., Kellis, M., Keller, C. A., Morrissey, C. S., Mishra, T., Jain, D., Dogan, N., Harris, R. S., Cayting, P., Kawli, T., Boyle, A. P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V. S., Cline, M. S., Erickson, D. T., Kirkup, V. M., Learned, K., Sloan, C. A., Rosenbloom, K. R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., James Kent, W., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P. J., Wilken, M. S., Reh, T. A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A. P., Neph, S., Humbert, R., Scott Hansen, R., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E. E., Orkin, S. H., Levasseur, D., Papayannopoulou, T., Chang, K.-H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M. J., Blobel, G. A., Cao, X., Zhong, S., Wang, T., Good, P. J., Lowdon, R. F., Adams, L. B., Zhou, X.-Q., Pazin, M. J., Feingold, E. A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S. M., Stamatoyannopoulos, J. A., Snyder, M. P., Guigo, R., Gingeras, T. R., Gilbert, D. M., Hardison, R. C., Beer, M. A. & Ren, B. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364 (2014).

8. Gray, L. T., Yao, Z., Nguyen, T. N., Kim, T. K., Zeng, H. & Tasic, B. Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. eLife 6, (2017).

9. Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghighi, F. G., Sejnowski, T. J., Behrens, M. M. & Ecker, J. R. Global Epigenomic Reconfiguration During Mammalian Brain Development. Science 341, 1237905 (2013).

10. Gilsbach, R., Preissl, S., Grüning, B. A., Schnick, T., Burger, L., Benes, V., Würch, A., Bönisch, U., Günther, S., Backofen, R., Fleischmann, B. K., Schübeler, D. & Hein, L. Dynamic DNA methylation orchestrates cardiomyocyte development, maturation and disease. Nature Communications 5, (2014).

11. Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., Koenig, J. L., Snyder, M. P., Pritchard, J. K., Kundaje, A., Greenleaf, W. J., Majeti, R. & Chang, H. Y. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nature Genetics 48, 1193–1203 (2016).

12. Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K.

L., Steemers, F. J., Trapnell, C. & Shendure, J. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348, 910–914 (2015).

13. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y. & Greenleaf, W. J. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015).

14. Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R. S., Stehling-Sun, S., Sabo, P. J., Byron, R., Humbert, R., Thurman, R. E., Johnson, A. K., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Giste, E., Haugen, E., Dunn, D., Wilken, M. S., Josefowicz, S., Samstein, R., Chang, K.-H., Eichler, E. E., De Bruijn, M., Reh, T. A., Skoultchi, A., Rudensky, A., Orkin, S. H., Papayannopoulou, T., Treuting, P. M., Selleri, L., Kaul, R., Groudine, M., Bender, M. A. & Stamatoyannopoulos, J. A. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 346, 1007–1012 (2014).

15. van der Maaten, L. Barnes-Hut-SNE. arXiv:1301.3342 [cs, stat] (2013). at <http://arxiv.org/abs/1301.3342>

16. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. Science 344, 1492–1496 (2014).

17. Zeisel, A., Munoz-Manchado, A. B., Codeluppi, S., Lonnerberg, P., La Manno, G., Jureus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J. & Linnarsson, S. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142 (2015).

18. La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L. E., Stott, S. R. W., Toledo, E. M., Villaescusa, J. C., Lönnerberg, P., Ryge, J., Barker, R. A., Arenas, E. & Linnarsson, S. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. Cell 167, 566-580.e19 (2016).

19. Rousseau, A., Nutt, C. L., Betensky, R. A., Iafrate, A. J., Han, M., Ligon, K. L., Rowitch, D. H. & Louis, D. N. Expression of Oligodendroglial and Astrocytic Lineage Markers in Diffuse Gliomas: Use of YKL-40, ApoE, ASCL1, and NKX2-2. Journal of Neuropathology and Experimental Neurology 65, 1149–1156 (2006).

20. Pernet, V., Joly, S., Christ, F., Dimou, L. & Schwab, M. E. Nogo-A and Myelin-Associated Glycoprotein Differently Regulate Oligodendrocyte Maturation and Myelin Formation. Journal of Neuroscience 28, 7435–7444 (2008).

21. Matcovitch-Natan, O., Winter, D. R., Giladi, A., Vargas Aguilar, S., Spinrad, A., Sarrazin, S., Ben-Yehuda, H., David, E., Zelada Gonzalez, F., Perrin, P., Keren-Shaul, H., Gury, M., Lara-Astaiso, D., Thaiss, C. A., Cohen, M., Bahar Halpern, K., Baruch, K., Deczkowska, A., Lorenzo-Vivas, E., Itzkovitz, S., Elinav, E., Sieweke, M. H., Schwartz, M. & Amit, I. Microglia development follows a stepwise program to regulate brain homeostasis. Science 353, aad8670–aad8670 (2016).

22. Huttner, H. B., Bergmann, O., Salehpour, M., Rácz, A., Tatarishvili, J., Lindgren, E., Csonka, T., Csiba, L., Hortobágyi, T., Méhes, G., Englund, E., Solnestam, B. W., Zdunek, S., Scharenberg, C., Ström, L., Ståhl, P., Sigurgeirsson, B., Dahl, A., Schwab, S., Possnert, G., Bernard, S., Kokaia, Z., Lindvall, O., Lundeberg, J. & Frisén, J. The age and genomic integrity of neurons after cortical stroke in humans. Nature Neuroscience 17, 801–803 (2014).

23. Su, Y., Shin, J., Zhong, C., Wang, S., Roychowdhury, P., Lim, J., Kim, D., Ming, G. & Song, H. Neuronal activity modifies the chromatin accessibility landscape in the adult brain. Nature Neuroscience 20, 476–483 (2017).

24. Kierdorf, K., Erny, D., Goldmann, T., Sander, V., Schulz, C., Perdiguero, E. G., Wieghofer, P., Heinrich, A., Riemke, P., Hölscher, C., Müller, D. N., Luckow, B., Brocker, T., Debowski, K., Fritz, G., Opdenakker, G., Diefenbach, A., Biber, K., Heikenwalder, M., Geissmann, F., Rosenbauer, F. & Prinz, M. Microglia emerge from erythromyeloid precursors via Pu.1- and Irf8-dependent pathways. Nature Neuroscience 16, 273–280 (2013).

25. Glasgow, S. M., Zhu, W., Stolt, C. C., Huang, T.-W., Chen, F., LoTurco, J. J., Neul, J. L., Wegner, M., Mohila, C. & Deneen, B. Mutual antagonism between Sox10 and NFIA regulates diversification of glial lineages and glioma subtypes. Nature Neuroscience 17, 1322–1329 (2014).

26. Nord, A. S., Pattabiraman, K., Visel, A. & Rubenstein, J. L. R. Genomic Perspectives of Transcriptional Regulation in Forebrain Development. Neuron 85, 27–47 (2015).

27. Yuan, F., Fang, K.-H., Cao, S.-Y., Qu, Z.-Y., Li, Q., Krencik, R., Xu, M., Bhattacharyya, A., Su, Y.-W., Zhu, D.-Y. & Liu, Y. Efficient generation of region-specific forebrain neurons from human pluripotent stem cells under highly defined condition. Scientific Reports 5, (2016).

28. Barbosa, A. C., Kim, M.-S., Ertunc, M., Adachi, M., Nelson, E. D., McAnally, J., Richardson, J. A., Kavalali, E. T., Monteggia, L. M., Bassel-Duby, R. & Olson, E. N. MEF2C, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function. Proceedings of the National Academy of Sciences 105, 9391–9396 (2008).

29. Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., Castanon, R., Lucero, J., Nery, J. R., Sandoval, J. P., Bui, B., Sejnowski, T. J., Harkins, T. T., Mukamel, E. A., Behrens, M. M. & Ecker, J. R. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. Science 357, 600–604 (2017).

30. Martynoga, B., Drechsel, D. & Guillemot, F. Molecular Control of Neurogenesis: A View from the Mammalian Cerebral Cortex. Cold Spring Harbor Perspectives in Biology 4, a008359–a008359 (2012).

31. Pollen, A. A., Nowakowski, T. J., Chen, J., Retallack, H., Sandoval-Espinosa, C., Nicholas, C. R., Shuga, J., Liu, S. J., Oldham, M. C., Diaz, A., Lim, D. A., Leyrat, A. A., West, J. A. & Kriegstein, A. R. Molecular Identity of Human Outer Radial Glia

during Cortical Development. Cell 163, 55–67 (2015).

32. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology 28, 495–501 (2010).

33. Subramanian, L., Sarkar, A., Shetty, A. S., Muralidharan, B., Padmanabhan, H., Piper, M., Monuki, E. S., Bach, I., Gronostajski, R. M., Richards, L. J. & Tole, S. Transcription factor Lhx2 is necessary and sufficient to suppress astrogliogenesis and promote neurogenesis in the developing hippocampus. Proceedings of the National Academy of Sciences 108, E265–E274 (2011).

34. Hsu, L. C.-L., Nam, S., Cui, Y., Chang, C.-P., Wang, C.-F., Kuo, H.-C., Touboul, J. D. & Chou, S.-J. Lhx2 regulates the timing of β-catenin-dependent cortical neurogenesis. Proceedings of the National Academy of Sciences 112, 12199–12204 (2015).

35. Castro, D. S., Skowronska-Krawczyk, D., Armant, O., Donaldson, I. J., Parras, C., Hunt, C., Critchley, J. A., Nguyen, L., Gossler, A., Göttgens, B., Matter, J.-M. & Guillemot, F. Proneural bHLH and Brn Proteins Coregulate a Neurogenic Program through Cooperative Binding to a Conserved DNA Motif. Developmental Cell 11, 831–844 (2006).

36. Castro, D. S., Martynoga, B., Parras, C., Ramesh, V., Pacary, E., Johnston, C., Drechsel, D., Lebel-Potter, M., Garcia, L. G., Hunt, C., Dolle, D., Bithell, A., Ettwiller, L., Buckley, N. & Guillemot, F. A novel function of the proneural factor Ascl1 in progenitor proliferation identified by genome-wide characterization of its targets. Genes & Development 25, 930–945 (2011).

37. Long, J. E., Cobos, I., Potter, G. B. & Rubenstein, J. L. R. Dlx1&2 and Mash1 Transcription Factors Control MGE and CGE Patterning and Differentiation through Parallel and Overlapping Pathways. Cerebral Cortex 19, i96–i106 (2009).

38. Heng, Y. H. E., McLeay, R. C., Harvey, T. J., Smith, A. G., Barry, G., Cato, K., Plachez, C., Little, E., Mason, S., Dixon, C., Gronostajski, R. M., Bailey, T. L., Richards, L. J. & Piper, M. NFIX Regulates Neural Progenitor Cell Differentiation During Hippocampal Morphogenesis. Cerebral Cortex 24, 261–279 (2014).

39. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T. & Taipale, J. DNA-Binding Specificities of Human Transcription Factors. Cell 152, 327–339 (2013).

40. Hori, K., Cholewa-Waclaw, J., Nakada, Y., Glasgow, S. M., Masui, T., Henke, R. M., Wildner, H., Martarelli, B., Beres, T. M., Epstein, J. A., Magnuson, M. A., MacDonald, R. J., Birchmeier, C. & Johnson, J. E. A nonclassical bHLH Rbpj transcription factor complex is required for specification of GABAergic neurons independent of Notch signaling. Genes & Development 22, 166–178 (2008).

41. Tian, X., Kai, L., Hockberger, P. E., Wokosin, D. L. & Surmeier, D. J. MEF-2 regulates activity-dependent spine loss in striatopallidal medium spiny neurons. Molecular and Cellular Neuroscience 44, 94–108 (2010).

42. Onorati, M., Castiglioni, V., Biasci, D., Cesana, E., Menon, R., Vuono, R., Talpo, F., Laguna Goya, R., Lyons, P. A., Bulfamante, G. P., Muzio, L., Martino, G., Toselli, M., Farina, C., Barker, R. A., Biella, G. & Cattaneo, E. Molecular and functional definition of the developing human striatum. Nature Neuroscience 17, 1804–1815 (2014).

43. Ghisletti, S., Barozzi, I., Mietton, F., Polletti, S., De Santa, F., Venturini, E., Gregory, L., Lonie, L., Chew, A., Wei, C.-L., Ragoussis, J. & Natoli, G. Identification and Characterization of Enhancers Controlling the Inflammatory Gene Expression Program in Macrophages. Immunity 32, 317–328 (2010).

44. Choksi, S. P., Lauter, G., Swoboda, P. & Roy, S. Switching on cilia: transcriptional networks regulating ciliogenesis. Development 141, 1427–1441 (2014).

45. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res 35, D88–D92 (2007).

46. Silberberg, S. N., Taher, L., Lindtner, S., Sandberg, M., Nord, A. S., Vogt, D., Mckinsey, G. L., Hoch, R., Pattabiraman, K., Zhang, D., Ferran, J. L., Rajkovic, A., Golonzhka, O., Kim, C., Zeng, H., Puelles, L., Visel, A. & Rubenstein, J. L. R. Subpallial Enhancer Transgenic Lines: a Data and Tool Resource to Study Transcriptional Regulation of GABAergic Cell Fate. Neuron 92, 59–74 (2016).

47. Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R. V., McKinsey, G. L., Pattabiraman, K., Silberberg, S. N., Blow, M. J., Hansen, D. V., Nord, A. S., Akiyama, J. A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., Kaplan, T., Kriegstein, A. R., Rubin, E. M., Ovcharenko, I., Pennacchio, L. A. & Rubenstein, J. L. R. A High-Resolution Enhancer Atlas of the Developing Telencephalon. Cell 152, 895–908 (2013).

48. Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., Wildberg, A., Gao, D., Fung, H.-L., Chen, S., Vijayaraghavan, R., Wong, J., Chen, A., Sheng, X., Kaper, F., Shen, R., Ronaghi, M., Fan, J.-B., Wang, W., Chun, J. & Zhang, K. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science 352, 1586–1590 (2016).

49. Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research 34, W369–W373 (2006).

50. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell 38, 576–589 (2010).

51. Amini, S., Pushkarev, D., Christiansen, L., Kostem, E., Royce, T., Turk, C., Pignatelli,

N., Adey, A., Kitzman, J. O., Vijayan, K., Ronaghi, M., Shendure, J., Gunderson, K. L. & Steemers, F. J. Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nature Genetics 46, 1343–1349 (2014).

52. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods 10, 1213–1218 (2013).

53. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25 (2009).

54. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

55. Habib, N., Li, Y., Heidenreich, M., Swiech, L., Avraham-Davidi, I., Trombetta, J. J., Hession, C., Zhang, F. & Regev, A. Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. Science 353, 925–928 (2016).

56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

57. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288 (1996).

58. Adey, A., Morrison, H. G., (no last name), A., Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X. & Shendure, J. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biology 11, R119 (2010).

59. Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F. & Manke, T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research 44, W160–W165 (2016).

60. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based Analysis of ChIP-Seq (MACS). Genome Biology 9, R137 (2008).

61. Mathelier, A., Fornes, O., Arenillas, D. J., Chen, C., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A. W., Parcy, F., Lenhard, B., Sandelin, A. & Wasserman, W. W. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research 44, D110–D115 (2016).

62. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26,

139–140 (2010).

CHAPTER 2: COMPREHENSIVE ANALYSIS OF SINGLE CELL ATAC-SEQ DATA

2.1 Abstract

Identification of the *cis*-regulatory elements controlling cell-type specific gene expression patterns is essential for understanding the origin of cellular diversity. Conventional assays to map *cis* regulatory elements via open chromatin analysis of primary tissues is hindered by heterogeneity of the samples. Single cell analysis of transposase-accessible chromatin (scATAC-seq) can overcome this limitation. However, the high-level noise of each single cell profile and the large volumes of data could pose unique computational challenges. Here, we introduce SnapATAC, a software package for analyzing scATAC-seq datasets. SnapATAC overcomes these challenges by employing diffusion maps, a non-linear dimensionality reduction algorithm that is highly robust to noise, to resolve the heterogeneity in complex tissues and map the trajectories of cellular states. Using the Nyström method, a sampling technique that generates the low rank embedding for large-scale dataset, SnapATAC can process data from a million cells. In addition, SnapATAC provides tools for integration of scATAC-seq and scRNA-seq, prediction of enhancer-promoter pairing, correction of batch effects and annotation of new datasets based on an existing reference cell atlas. As a demonstration of its utility, SnapATAC was applied to 55,592 single-nucleus ATAC-seq profiles from the mouse secondary motor cortex. The analysis results revealed ~370,000 candidate regulatory elements active in 31 distinct cell populations and inferred candidate transcriptional regulators in each of the cell types. These results demonstrate that SnapATAC is a systematic and powerful tool for analyzing single cell ATAC-seq datasets.

2.2 Introduction

Human body comprises of divergent cell types that are highly specialized to carry out distinct functions[1]. The identity of each cell type is established during development through complex gene expression programs, which are driven in part by sequence-specific transcription factors that interact with *cis*-regulatory sequences in a cell-type specific manner[2]. Thus, identifying the *cis*-elements and their cellular specificity is an essential step towards understanding the cell type specific gene expression programs

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

of a nano-well array (Takara Bio, ICELL8)[11]. More recently, single cell ATAC-seq analysis has been demonstrated on droplet-based platforms[12,13], enabling profiling of chromatin accessibility from even hundreds of thousands of cells in a single experiment[12]. Hereafter, these methods are referred to collectively as single cell ATAC-seq (scATAC-seq).

The growing volumes of scATAC-seq datasets and the sparsity of signals in each individual profile due to low detection efficiency (5-15% of peaks detected per cell)[5] present a unique computational challenge for resolving cellular heterogeneity. To address this challenge, a number of unsupervised algorithms have been developed. One approach, chromVAR[14], groups similar cells together by dissecting the variability of transcription factor (TF) motif occurrence in the open chromatin regions in each cell. Another approach employs the natural language processing techniques such as Latent Semantic Analysis (LSA)[8] and Latent Dirichlet Allocation (LDA)[15] to group cells together based on the similarity of chromatin accessibility. A third approach analyzes the variability of chromatin accessibility in cells based on the k-mer composition of the sequencing reads from each cell[12,16]. A fourth approach, Cicero[17], infers cell-to-cell similarities based on the gene activity scores predicted from their putative regulatory elements in each cell.

Because the current methods often require performing linear dimensionality reduction such as principle component analysis on a cell matrix of hundreds of thousands of dimensions, scaling the analysis to millions of cells remains very challenging or nearly impossible. In addition, the unsupervised identification of cell types or states in complex

tissues using scATAC-seq dataset does not match the power of scRNA-seq[18]. One possibility is that the current methods rely on the use of pre-defined accessibility peaks based on the aggregate signals. There are several limitations to this choice. First, the cell type identification could be biased toward the most abundant cell types in the tissues. Second, sufficient number of single cell profiles are required to create robust aggregate signal for calling peaks. Third, these techniques lack the ability to reveal regulatory elements in the rare cell populations, which are underrepresented in the aggregate signal.

To overcome these limitations, we develop a software package, Single Nucleus Analysis Pipeline for ATAC-seq – SnapATAC (https://github.com/r3fang/SnapATAC). SnapATAC does not require population-level peak annotation prior to clustering. Instead, it resolves cellular heterogeneity by directly comparing the genome-wide accessibility profiles between cells with the use of the diffusion maps algorithm[19,20], which is highly robust to noise and perturbation. Furthermore, with the use of a sampling technique, Nyström method[21,21,22], SnapATAC improves the computational efficiency and enables the analysis of scATAC-seq from a million cells on regular hardware. Additionally, SnapATAC provides a collection of frequently used features, including integration of scATAC-seq and scRNA-seq dataset, prediction of enhancer-promoter interaction, discovery of key transcription factors, identification of differentially accessible elements, construction of trajectories during cellular differentiation, correction of batch effect and classification of new dataset based on existing cell atlas. Thus, SnapATAC represents a comprehensive solution for scATAC-seq analysis.

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

2.3 Results

Overview of SnapATAC workflow. SnapATAC first performs pre-processing of sequencing reads including demultiplexing, reads alignments and filtering, duplicate removal and barcode selection using SnapTools (https://github.com/r3fang/SnapTools) (**Supplementary Methods**), and then generates a "snap" (Single-Nucleus Accessibility Profiles) file specially formatted for storing single cell ATAC-seq datasets (**Figure S2.1a**). SnapTools is substantially faster than another popular tool - CellRanger for preprocessing (**Figure S2.1b**). To remove potential doublets, SnapATAC adopts a recently reported algorithm Scrublet[23] (**Supplementary Methods** and **Figure S2.2**).

Next, SnapATAC resolves the heterogeneity of cell population by assessing the similarity of chromatin accessibility between cells. To achieve this goal, each single cell chromatin accessibility profile is represented as a binary vector, the length of which corresponds to the number of uniform-sized bins that segment the genome. Through systematic benchmarking, an optimal bin size of 5kb is chosen (**Supplementary Methods** and **Figure 2.3**). A bin with value "1" indicates that one or more reads fall within that bin, and the value "0" indicates otherwise. The set of binary vectors from all the cells are converted into a Jaccard similarity matrix, with the value of each element calculated from the fraction of overlapping bins between every pair of cells. Because the value of Jaccard Index could be influenced by sequencing depth of a cell (**Supplementary Methods**), a regression-based normalization method is developed to remove this confounding factor (**Supplementary Methods** and **Figure S2.4** and **S2.5**). Using the normalized similarity matrix, eigenvector decomposition is performed for dimensionality

reduction. Such procedure is known as the diffusion maps alogrithm[19,20]. This approach is chosen because it preserves the nonlinear structure of the data through a random-walk process on the data and is highly robust to perturbation and noise[19,20], which makes it particularly well suited for the sparse single cell ATAC-seq dataset. Finally, in the reduced dimension, SnapATAC uses Harmony[24] to remove potential batch effect between samples introduced by technical variability (**Supplementary Methods**).

The computational cost of the diffusion maps algorithm scales exponentially with the number of cells. To improve the scalability of SnapATAC, a sampling technique - the Nyström method[21] – is used to efficiently generate the low-rank diffusion maps embedding for large-scale datasets (**Supplementary Methods**). Nyström method contains two major steps: 1) it computes the diffusion maps embedding for a subset of selected cells (also known as landmarks); 2) it projects the remaining cells to the embedding learned from the landmarks. This achieves significant speedup considering that the number of landmarks could be substantially smaller than the total number of cells. Through benchmarking, we further demonstrate that this approach will not sacrifice the performance once the landmarks are carefully chosen (**Supplementary Methods** and **Figure S2.6**; **Figure S2.7**) as reported before[22].

Nyström method is stochastic and could yield different clustering results in each sampling. To overcome this limitation, a consensus approach is used that combines a mixture of low-dimensional manifolds learned from different sets of sampling (**Supplementary Methods**). This consensus algorithm naturally fits within the distributed

computing environments where their computational costs are roughly the same as that of the standard single sampling method.

As a standalone software package, SnapATAC also provides a number of commonly used functions for scATAC-seq analysis, as described below:

First, to facilitate the annotation of resulting cell clusters, SnapATAC provides three different approaches: i) SnapATAC annotates the clusters based on the accessibility score at the canonical marker genes (**Supplementary Methods**); ii) it infers cell type labels by integrating with corresponding single cell RNA-seq datasets (**Supplementary Methods** and **Figure 2.2a**); iii) it allows supervised annotation of new single cell ATAC-seq dataset based on an existing cell atlas (**Supplementary Methods**).

Second, SnapATAC allows identification of the candidate regulatory elements in each cluster by applying peak-calling algorithms to the aggregate chromatin profiles. Differential analysis is then performed to identify cell-type specific regulatory elements. Candidate master transcription factors in each cell cluster are discovered through motif enrichment analysis[14,25] of the differentially accessible regions in each cluster. SnapATAC further conducts Genomic Regions Enrichment of Annotation Tool (GREAT) analysis[26] to identify the biological pathways active in each cell type.

Third, SnapATAC incorporates a new approach to link candidate regulatory elements to their putative target genes. In contrast to previous method[17] that relies on

analysis of co-accessibility of distal elements and promoters, SnapATAC infers the linkage based on the association between gene expression and chromatin accessibility in single cells where scRNA-seq data is available (**Supplementary Methods**). First, SnapATAC integrates scATAC-seq and scRNA-seq in a way that significantly outperforms existing methods on the accuracy (Wilcox two-sided rank test P < 2.2e-16; **Figure S2.8a**). Second, for each scATAC-seq profile, a corresponding gene expression profile is imputed based on the weighted average of its *k*-nearest neighboring cells in the scRNA-seq dataset. Thus, a "pseudo" cell is created that contains the information of both chromatin accessibility and gene expression. Finally, logistic regression is performed to quantify the association between the gene expression and binarized accessibility state at distal elements (**Supplementary Methods**). This new approach is used to integrate ~15K peripheral blood mononuclear cells (PBMC) chromatin profiles and ~10K PBMC transcriptomic profiles (**Figure 2.2a**) and represent them in a joint t-SNE embedding space (**Figure 2.2a**). Over 98% of the single cell ATAC-seq cells can be confidently assigned to a cell type defined in the scRNA-seq dataset (**Figure S2.8b**). Enhancer-gene pairs are predicted for 3,000 genes differentially expressed between cell types in PBMC as determined by scRNA-seq using Seurat[18] (**Supplementary Methods**). The accuracy of these predictions is supported by several lines of evidence. First, the promoters exhibit the highest association with the gene expression (**Figure S2.8c**). Second, the association score exhibits a distance decay from the TSS, consistent with the distance decay of interaction frequency observed in chromatin conformation study[27] (**Figure S2.8c**). Finally, the predictions match well with the expression quantitative trait loci (*cis*-eQTLs) derived from interferon-γ and lipopolysaccharide stimulation of monocytes[28], with the gene-

enhancer pairs overlapping with 64% of cis-eQTLs, nearly two-fold of that is expected for genes located at the same distances (**Figure 2.2c** and **Supplementary Methods**). While the predictions require further experimental validation, statistical association between scATAC-seq and scRNA-seq provides another approach to symmetrically link enhancers to their putative target genes.

Fourth, SnapATAC has incorporated a function to construct cellular trajectories from single cell ATAC-seq with the use of the diffusion maps algorithm, previously used to define cellular trajectories from single cell RNA-seq dataset[29]. As a demonstration of this feature, SnapATAC is used to analyze a dataset that contains 4,259 cells from the hippocampus in the fetal mouse brain (E18). Immature granule cells originating in the dentate gyrus give rise to both mature granule cells (DG) and pyramidal neurons (CA3). Analysis of 4,259 cells with diffusion maps reveals a clear branching structure in the first two diffusion components (DC) (**Figure 2.3a**), the pattern of which is remarkably similar to the result previously obtained from single cell transcriptomic analysis[29] (**Figure S2.9b**). For instance, the DG-specific transcription factor *Prox1* is exclusively accessible in one branch whereas *Neurod6* and *Spock1* that is known to be specific to CA3 are accessible in the other branch. Markers of progenitors such as *Hes5* and *Mki67*, however, are differentially accessible before the branching point (**Figure S2.3b**). Further using lineage inference tool such as Slingshot[30], SnapATAC defines the trajectories of cell states for pseudo-time analysis (**Figure 2.3a**). These results demonstrate that SnapATAC can also reveal lineage trajectories with high accuracy.

Performance evaluation. To compare the accuracy of cell clustering between SnapATAC and published scATAC-seq analysis methods, a simulated dataset of scATAC-seq profiles are generated with varying coverages, from 10,000 (high coverage) to 1,000 reads per cell (low coverage) by down sampling from 10 previously published bulk ATAC-seq datasets (**Supplementary Methods**). The performance of each method in identifying the original cell types is measured by Adjusted Rank Index (ARI). This comparison shows that SnapATAC is the most robust and accurate method across all ranges of data sparsity (Wilcoxon signed-rank test, $P < 0.01$; **Figure 2.4a**; **Figure S2.10**). SnapATAC performs especially well on the sparse datasets (**Figure 2.4a**), likely due to the fact that the diffusion maps algorithm is highly robust to noise and perturbation. Next, a set of 1,423 human cells corresponding to 10 distinct cell types generated using C1 Fluidigm platform, where the ground truth is known[14] , is analyzed by SnapATAC and other methods. Again, SnapATAC correctly identifies the cell types with higher accuracy than alternative approaches (**Figure S2.11**).

To compare the sensitivity of SnapATAC to detect rare cell types to that of previously published methods, we analyzed three scATAC-seq datasets representing different types of bio-samples. The first dataset contains 9,529 single nucleus open chromatin profiles generated from the mouse secondary motor cortex. SnapATAC uncovers 22 distinct cell populations (**Figure 2.4b** and **Figure S2.12**) whereas alternative methods fail to distinguish the rare neuronal subtypes including Sst (Gad2+ and Sst+), Vip (Gad2+ and Vip+), L6b (Sulf1- and Tl4e+) and L6.CT (Sulf1+ and Foxp2+) (**Figure S2.13**). The second dataset includes 4,098 cells from the adult mouse brain (10X

genomics). SnapATAC again uncovers more well-known neuronal populations than alternative approaches (**Figure S2.14-S2.15**). The third dataset contains 4,792 PBMC. SnapATAC successfully separates the pre-B cells from B cell progenitor cells, while alternative methods fail to distinguish these two subtypes (**Figure S2.16-S2.17**). These results suggest that SnapATAC outperforms existing methods in sensitivity.

To compare the scalability of SnapATAC to that of existing methods, a previous scATAC-seq dataset that contains over 80k cells from 13 different mouse tissues[8] is used. This dataset is down sampled to different number of cells, ranging from 20,000 to 80,000 cells. For each sampling, SnapATAC and other methods are performed, and the CPU running time of dimensionality reduction is monitored (**Supplementary Methods**). The running time of SnapATAC scales linearly and increases at a significantly lower slope than alternative methods (**Figure 2.4c**). Using the same computing resource, when applied to 100k cells, SnapATAC is much faster than existing methods (**Figure 2.4c**). For instance, when applied to 100k cells, SnapATAC is nearly 10 times faster than LSA and more than 100 times faster than cisTopic. More importantly, because SnapATAC avoids the loading of the full cell matrix in the memory and can naturally fit within the distributed computing environments (**Supplementary Methods**), the running time and memory usage for SnapATAC plateau after 20,000 cells, making it possible for analyzing datasets of even greater volumes. To test this, we simulate one million cells of the same coverage with the above dataset (**Supplementary Methods**) and process it with SnapATAC, LSA and cisTopic. Using the same computing resource, SnapATAC is the only method that is able to process this dataset (**Figure 2.4c** and **Supplementary Methods**). These results

demonstrate that SnapATAC provides a highly scalable approach for analyzing large-scale scATAC-seq dataset.

To evaluate the clustering reproducibility, the above mouse scATAC-seq dataset is down-sampled to 90% of the original sequencing depth in 5 different iterations. Each down sampled dataset is clustered using SnapATAC and other methods. Clustering results are compared between sampled datasets to estimate the stability. SnapATAC has a substantially higher reproducibility of clustering results between different down-sampled datasets than other methods (**Figure 2.4d**; two-side t-test Pvalue < 1e-2).

The improved performance of SnapATAC likely results from the fact that it considers all reads from each cell, not just the fraction of reads within the peaks defined in the population. To test this hypothesis, clustering is performed after removing reads overlapping with the predefined peak regions. The outcome largely recapitulates the majority of cell types obtained from the full dataset (**Figure S2.18**). This holds true for all three datasets tested (**Figure S2.18**). One possibility is that the off-peak reads may be enriched for the euchromatin (or compartment A) that strongly correlates with active genes[27] and varies considerably between cell types[31]. Consistent with this hypothesis, the density of the non-peak reads in scATAC-seq library is highly enriched for the euchromatin (compartment A) as defined using genome-wide chromatin conformation capture analysis (i.e. Hi-C) in the same cell type[32] (**Figure S2.19**). These observations suggest that the non-peak reads discarded by existing methods can actually contribute to distinguish different cell types.

Including the off-peak reads, however, raises a concern regarding whether SnapATAC is sensitive to technical variations (also known as batch effect). To test this, SnapATAC is applied to four datasets generated using different technologies. Each dataset contains at least two biological replicates produced by the same technology. In all cases, the biological replicates are well mixed in the t-SNE embedding space showing no batch effect (**Figure S2.20a-d**), suggesting that SnapATAC is robust to the technical variations. To test whether SnapATAC is robust to technical variation introduced by different technological platforms, it is used to integrate two mouse brain datasets generated using plate and droplet-based scATAC-seq technologies. In the joint t-TSNE embedding space, these two datasets are separated based on the technologies (**Figure S2.21a**). To remove the platform-to-platform variations, Harmony[24], a single cell batch effect correction tool, is incorporated into the SnapATAC pipeline (**Supplementary Methods**). After applying Harmony, these two datasets are fully mixed in the joint t-SNE embedding (**Figure S2.21b**) and clusters are fairly represented by both datasets (**Figure S2.21c**).

A high-resolution cis-regulatory atlas of the mouse motor cortex. To demonstrate the utility of SnapATAC in resolving cellular heterogeneity of complex tissues and identify candidate *cis*-regulatory elements in diverse cell type, it is applied to a single nucleus ATAC-seq dataset generated from the secondary mouse motor cortex in the adult mouse brain as part of the BRAIN Initiative Cell Census Consortium[29] (**Figure S2.22a**). This dataset includes two biological replicates, each pooled from 15 mice to

minimize potential batch effects. The aggregate signals show high reproducibility between biological replicates (Pearson correlation = 0.99; **Figure S2.22b-d**) and a significant enrichment for transcription start sites (TSS), indicating a high signal-to-noise ratio (**Figure S2.22e**). After filtering out the low-quality nuclei (**Figure S2.22a**) and removing putative doublets using Scrublet[23] (**Figure S2.23b**), a total of 55,592 nuclear profiles with an average of ~5,000 unique fragments per nucleus remain and are used for further analysis. To our knowledge, this dataset represents the largest single cell chromatin accessibility dataset generated for the mouse brain to date.

SnapATAC identifies initially a total of 20 major clusters using the consensus clustering approach (**Figure S2.24**). The clustering result is highly reproducible between biological replicates (Pearson correlation=0.99; **Figure S2.25a**) and is resistant to sequencing depth effect (**Figure S2.25b**). Based on the gene accessibility score at the canonical marker genes (**Figure S2.26**), these clusters are classified into 10 excitatory neuronal subpopulations (Snap25+, Slc17a7+, Gad2-; 52% of total nuclei), three inhibitory neuronal subpopulations (Snap25+, Gad2+; 10% of total nuclei), one oligodendrocyte subpopulation (Mog+; 8% of total nuclei), one oligodendrocyte precursor subpopulation (Pdgfra+; 4% of total nuclei), one microglia subpopulation (C1qb+; 5% of total nuclei), one astrocyte subpopulation (Apoe+; 12% of total nuclei), and additional populations of endothelial, and somatic muscle cells accounting for 6% of total nuclei (**Figure 2.5a**).

In mammalian brain, GABAergic interneurons exhibit spectacular diversity that shapes the spatiotemporal dynamics of neural circuits underlying cognition[33]. To examine whether iterative analysis could help tease out various subtypes of GABAergic neurons, SnapATAC is applied to the 5,940 GABAergic nuclei (CGE, Sst and Vip) identified above, finding 17 distinct sub-populations (**Figure S2.27a**) that are highly reproducible between biological replicates (Pearson correlation = 0.99; **Figure S2.27b**). Based on accessibility level at the marker genes (**Figure S2.28**), these 17 clusters are classified into five Sst subtypes (Chodl+, Cbln4+, Igfbp6+, Myh8+ and C1ql3+), two Pv subtypes (Tac1+ and Ntf3+), two Lamp5 subtypes (Smad3+ and Ndnf+), four Vip subtypes (Mybpc1+, Chat+, Gpc3+, Crhr2+), Sncg and putative doublets (**Figure 2.5b**). These clusters include a rare type Sst-Chodl (0.1%) previously identified in single cell RNA analysis[34]. This represents the first time this population is recapitulated by single cell chromatin accessibility analysis. While the identity and function of these subtypes require further experimental validation, our results demonstrate the exquisite sensitivity of SnapATAC in resolving distinct neuronal subtypes with only subtle differences in the chromatin landscape.

A key utility of single cell chromatin accessibility analysis is to identify regulatory sequences in the genome. By pooling reads from nuclei in each major cluster (**Figure 2.5a**), cell-type specific chromatin landscapes can be obtained (**Figure 2.5b** and **Supplementary Methods**). Peaks are determined in each cell type, resulting in a total of 373,583 unique candidate *cis*-regulatory elements. Most notably, 56% (212,730/373,583) of these open chromatin regions cannot be detected from bulk ATAC-seq data of the same brain region (**Supplementary Methods**). The validity of these additional open

chromatin regions identified from scATAC-seq data are supported by several lines of evidence. First, these open chromatin regions are only accessible in minor cell populations (**Figure S2.29a**) that are undetectable in the bulk ATAC-seq signal. Second, these sequences show significantly higher conservation than randomly selected genomic sequences with comparable mappability scores (**Figure S2.29c**). Third, these open chromatin regions display an enrichment for transcription factor (TF) binding motifs corresponding to the TFs that play important regulatory roles in the corresponding cell types. For example, the binding motif for Mef2c is highly enriched in novel candidate *cis*-elements identified from Pvalb neuronal subtype (P-value = 1e-363; **Figure S2.29d**), consistent with previous report that Mef2c is upregulated in embryonic precursors of Pv interneurons[35]. Finally, the new open chromatin regions tend to test positive in transgenic reporter assays. Comparison to the VISTA enhancer database[36] shows that enhancer activities of 256 of the newly identified open chromatin regions have been previously tested using transgenic reporter assays in e11.5 mouse embryos. Sixty five percent (167/256; 65%) of them drive reproducible reporter expression in at least one embryonic tissue, which was substantially higher than background rates (9.7%) estimated from regions in the VISTA database that lack canonical enhancer mark[37]. Four examples are displayed (**Figure S2.29e**).

SnapATAC identifies 294,304 differentially accessible elements between cell types (**Supplementary Methods** and **Figure 2.5e**). Motif enrichment analysis (**Figure 2.5g**) and GREAT analysis (**Figure 2.5f**) then identify the master regulators and transcriptional pathways active in each of the cell types. For instance, the binding motif for ETS-factor

PU.1 is highly enriched in microglia-specific candidate CREs, motifs for SOX proteins are enriched in Ogc-specific elements, and bHLH motifs are enriched in excitatory neurons-specific CREs (**Figure 2.5g**). Interestingly, motifs for candidate transcriptional regulators, including NUCLEAR FACTOR 1 (NF1), are also enriched in candidate CREs detected in rare neuronal populations such as two inhibitory neuron subtypes (Lamp5.Ndnf and Lamp5.Smad3). Motif for CTCF, a multifunctional protein in genome organization and gene regulation[38], is highly enriched in Sst-Chodl, indicating that CTCF may also play a distinct role in neurogenesis. Finally, motifs for different basic-helix-loop-helix (bHLH) family transcription factors, known determinants of neural differentiation[39], show enrichment for distinct Sst subtypes. For instance, E2A motif is enriched in candidate CREs found in Sst.Myh8 whereas AP4 motif is specifically enriched in peaks found in Sst.Cbln4, suggesting specific role that different bHLH factors might play in different neuronal subtypes.

SnapATAC enables supervised annotation of new scATAC-seq dataset. Unsupervised clustering of scATAC-seq datasets frequently requires manual annotation, which is labor-intensive and limited to prior knowledge. To overcome this limitation, SnapATAC provides a function to project new single cell ATAC-seq datasets to an existing cell atlas to allow for supervised annotation of cells. First, the diffusion maps algorithm is used to project the query cells to the low-dimension manifold pre-computed from the reference cells (Supplementary Methods). In the joint manifold, a neighborhood-based classifier is used to determine the cell type of each query cell based on the label of its *k* nearest neighboring cells in the reference dataset (Supplementary Methods). The

accuracy of this method is determined by five-fold cross validation using the mouse motor cortex atlas. On average, 98% (±1%) of the cells can be correctly classified, suggesting a high accuracy of the method (Figure 2.6a).

To demonstrate that SnapATAC could be applied to datasets generated from distinct technical platforms, it is used to annotate 4,098 scATAC-seq profiles from mouse brain cells generated using a droplet-based platform. After removing batch effect introduced by different platforms using Harmony, the query cells are well mixed with the reference cells in the joint diffusion maps embedding (**Figure 2.30**). The predicted cluster labels are also consistent with the cell types defined using unbiased clustering analysis (NMI=0.85, ARI=0.68; **Figure 2.6b**).

To investigate whether SnapATAC could recognize cell types in the query dataset that are not present in the reference atlas, multiple query data sets are sampled from the above mouse motor cortex dataset and a perturbation is introduced to each sampling by randomly dropping a cell cluster. When this resulting query dataset is analyzed by SnapATAC against the original cell atlas, the majority of the cells that are left out from the original atlas are filtered out due to the low prediction score (**Figure 2.31**), again suggesting that our method is not only accurate but also robust to the novel cell types in the query dataset.

2.4 Discussion

In summary, SnapATAC is a comprehensive bioinformatic solution for single cell ATAC-seq analysis. The open-source software runs on regular hardware, making it accessible to a broad spectrum of researchers. Through extensive benchmarking, we have demonstrated that SnapATAC outperforms existing tools in sensitivity, accuracy, scalability and robustness of identifying cell types in complex tissues.

SnapATAC differs from previous methods in at least seven aspects. First, SnapATAC represents the only comprehensive solution for single cell ATAC-seq data analysis to date. In addition to clustering analysis, SnapATAC provides preprocessing, annotation, trajectory analysis, peak calling, differential analysis, batch effect correction and motif discovery all in one package. Second, SnapATAC identifies cell types in an unbiased manner without the need for population-level peak annotation, leading to superior sensitivity for identifying rare cell types in complex tissues. Third, SnapATAC employs the diffusion maps algorithm to identify cell types in heterogeneous tissues and map cellular trajectories, which is ideally suited for the sparse and noisy scATAC-seq datasets. Fourth, with Nyström sampling method, SnapATAC significantly reduces both CPU and memory usage, enabling analysis of large-scale dataset of a million cells or more. Fifth, SnapATAC not only integrates scATAC-seq with scRNA-seq dataset but also provides a new method to predict promoter-enhancer pairing relations based on the statistical association between gene expression and chromatin accessibility in single cells. Sixth, our method achieves high clustering reproducibility using a consensus

clustering approach. Finally, SnapATAC also enables supervised annotation of a new scATAC-seq dataset based on an existing reference cell atlas.

It is important to note that a different strategy has been used to overcome the bias introduced by population-based peak annotation[8]. This approach involves iterative clustering, with the first round defining the “crude” clusters in complex tissues followed by identifying peaks in these clusters, which are then used in subsequent round(s) of clustering. However, several limitations still apply. First, the “crude” clusters represent the most dominate cell types in the tissues; therefore, peaks in the rare populations may still be underrepresented. Indeed, when applied to the 10X mouse brain dataset, this approach is only able to reveal ~150,000 peaks in the adult mouse brain, less than half of the total peaks defined in the mouse brain from a current study[12]. Second, using these extended peaks as features for clustering does not improve the sensitivity of identifying rare cell populations compared to that using population-defined peak list (**Figure S2.32**). This is likely due to the fact that this method ignores the off-peak reads that contribute significantly to cell type identification as demonstrated in this study. Third, this approach requires multiple rounds of clustering, reads aggregation and peak calling, limiting its application to large scale dataset. Finally, peak-based methods hinder multi-sample integrative analysis where each sample has its own unique peak reference.

SnapATAC is applied to a new in-house dataset including 55,592 high quality single nucleus ATAC-seq profiles from mouse secondary motor cortex, producing a single cell atlas of candidate *cis*-regulatory elements for this mouse brain region. The cellular

diversity identified by chromatin accessibility is at an unprecedented resolution and is consistent with mouse neurogenesis and taxonomy revealed by single cell transcriptome data. Besides characterizing the constituent cell types, SnapATAC identifies candidate *cis*-regulatory sequences in each of the major cell types and infers the likely transcription factors that regulate cell-type specific gene expression programs. Importantly, a large fraction (56%) of the candidate *cis*-elements identified from the scATAC-seq data are not detected in bulk analysis. While further experiments to thoroughly validate the function of these additional open chromatin regions are needed, the ability for SnapATAC to uncover *cis*-elements from rare cell types of a complex tissue will certainly help expand the catalog of *cis*-regulatory sequences in the genome.

2.5 Acknowledgments

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

Chapter 2, in full, is currently being prepared for submission of the material as "Comprehensive Analysis of Single Cell ATAC-seq Data". Rongxin Fang, Sebastian Preissl, Yang Li, Xiaomeng Hou, Jacinta Lucero, Xinxin Wang, Amir Motamedi, Andrew K. Shiau, Kai Zhang, Xinzhu Zhou, Fangming Xie, Eran A. Mukamel, Yanxiao Zhang, M. Margarita Behrens, Joseph Ecker, and Bing Ren. The dissertation author was the primary investigator and author of this paper.

2.6 Author Contributions

This study was conceived and designed by R.F. and B.R.; Pipeline developed 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.; Data analysis performed by R.F. and Y. L.; Tn5 enzymes synthesized and provided by A.M. and A.S.; Manuscript written by R.F. and B.R. with input from all authors.

2.7 Figures

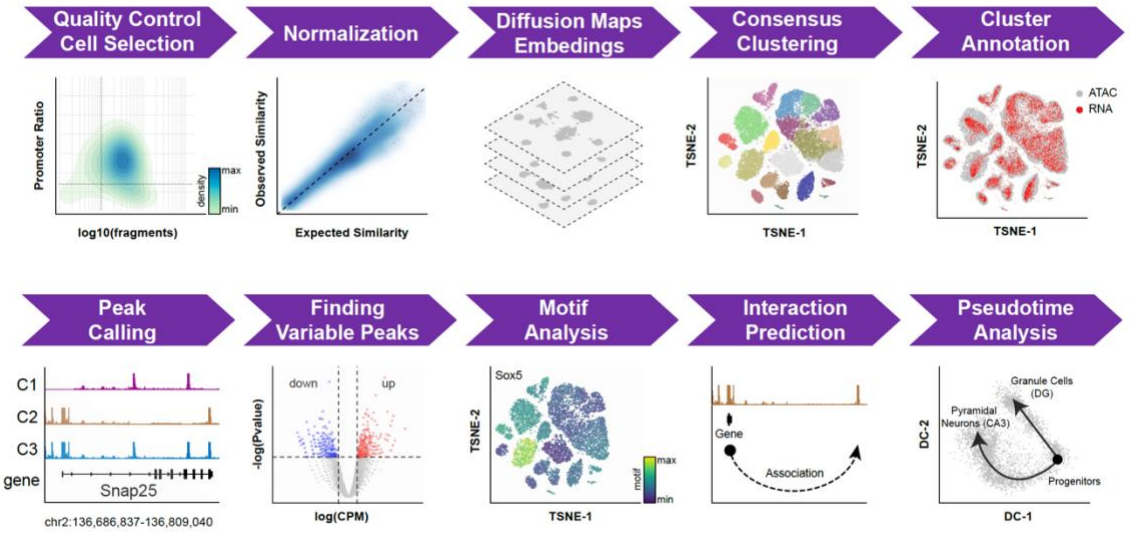
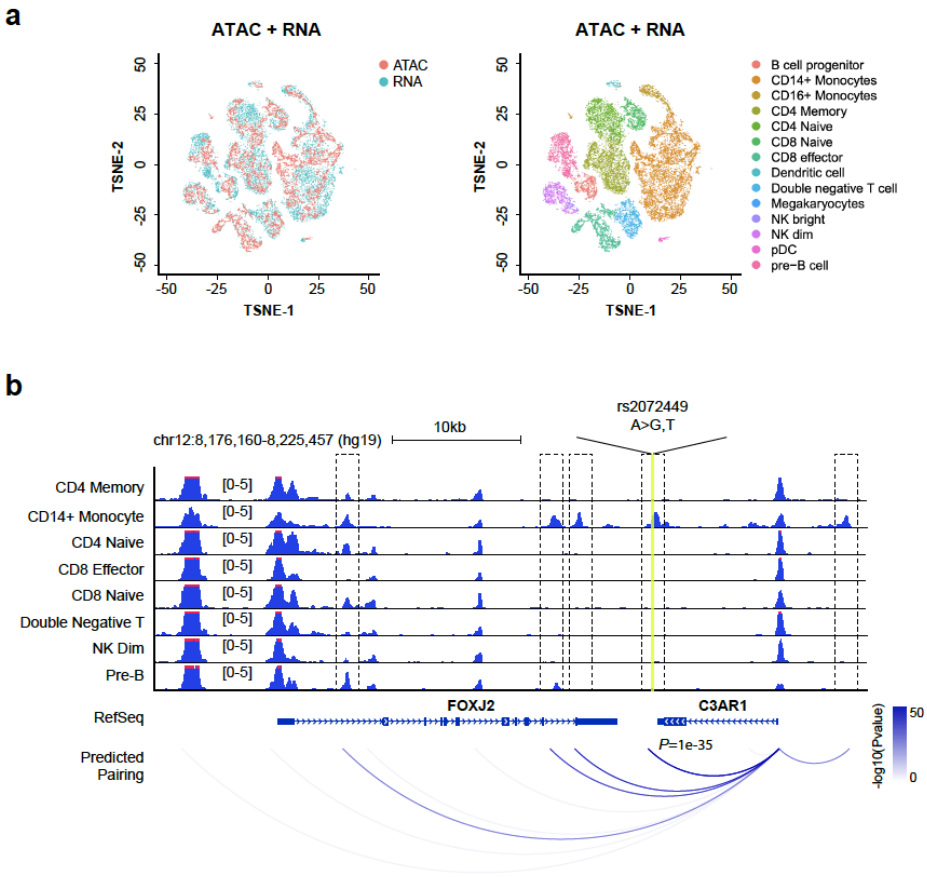


Figure 2.1. Schematic overview of SnapATAC analysis workflow. See main text for description of each step.

Figure 2.2. SnapATAC links distal regulatory elements to putative target genes. (**a**) Joint t-SNE visualization of scATAC-seq and scRNA-seq datasets from peripheral blood mononuclear cells (PBMC). Cells are colored by modality (left) and predicted cell types (right). (**b**) Cell-type specific chromatin landscapes are shown together with the association score between gene expression of C3AR1 and accessibility at its distal regulatory elements. Dash lines highlight the significant gene-enhancer pairs. Yellow line represents the SNP (rs2072449) that is associated with C3AR1 expression[28]. (**c**) Distribution of the scATAC/scRNA association *P*-value for 456 cis-eQTL pairs (left) and 456 negative control pairs matched for distances (**Supplementary Methods**).


a
ATAC + RNA
ATAC
RNA
TSNE-2
TSNE-1
ATAC + RNA
B cell progenitor
CD14+ Monocytes
CD16+ Monocytes
CD4 Memory
CD4 Naive
CD8 Naive
CD8 effector
Dendritic cell
Double negative T cell
Megakaryocytes
NK bright
NK dim
pDC
pre−B cell
b
chr12:8,176,160-8,225,457 (hg19)
10kb
rs2072449
A>G,T
CD4 Memory
CD14+ Monocyte
CD4 Naive
CD8 Effector
CD8 Naive
Double Negative T
NK Dim
Pre-B
[0-5]
RefSeq
FOXJ2
C3AR1
Predicted Pairing
P=1e-35
-log10(Pvalue)
c
cis-eQTLs pairs
negative control
Frequency
(298/456) 64%
(168/456) 36%
log10(P-value)

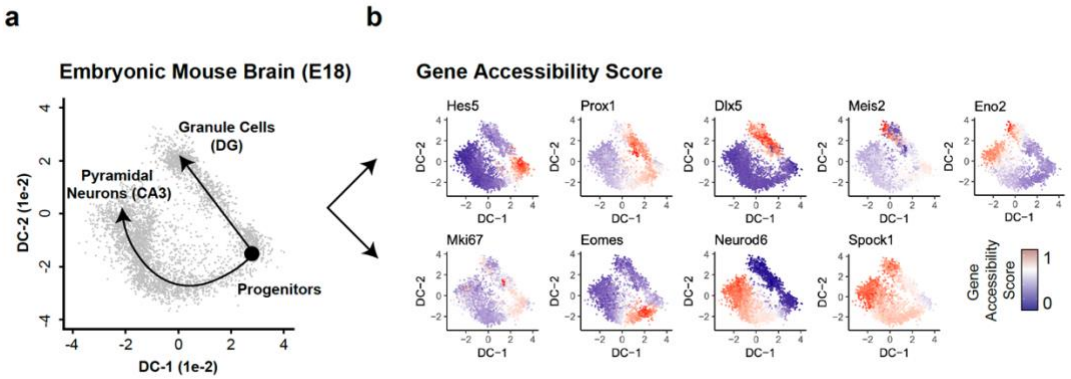


Figure 2.3. SnapATAC constructs cellular trajectories for the developing mouse brain. (**a**) Two-dimensional diffusion component visualization of a dataset that contains 4,259 single cell chromatin profiles from the hippocampus and ventricular zone in embryonic mouse brain (E18) reveals two-branch differentiation trajectories from progenitor cells to Granule Cells (DG) and Pyramidal Neurons (CA3) (left). The cellular trajectory is determined by Slingshot[30]. (**b**) Gene accessibility score of canonical marker genes is projected onto the diffusion component embedding. See also **Figure S2.9b** for dentate gyrus cell lineage identified using single nucleus RNA-seq.

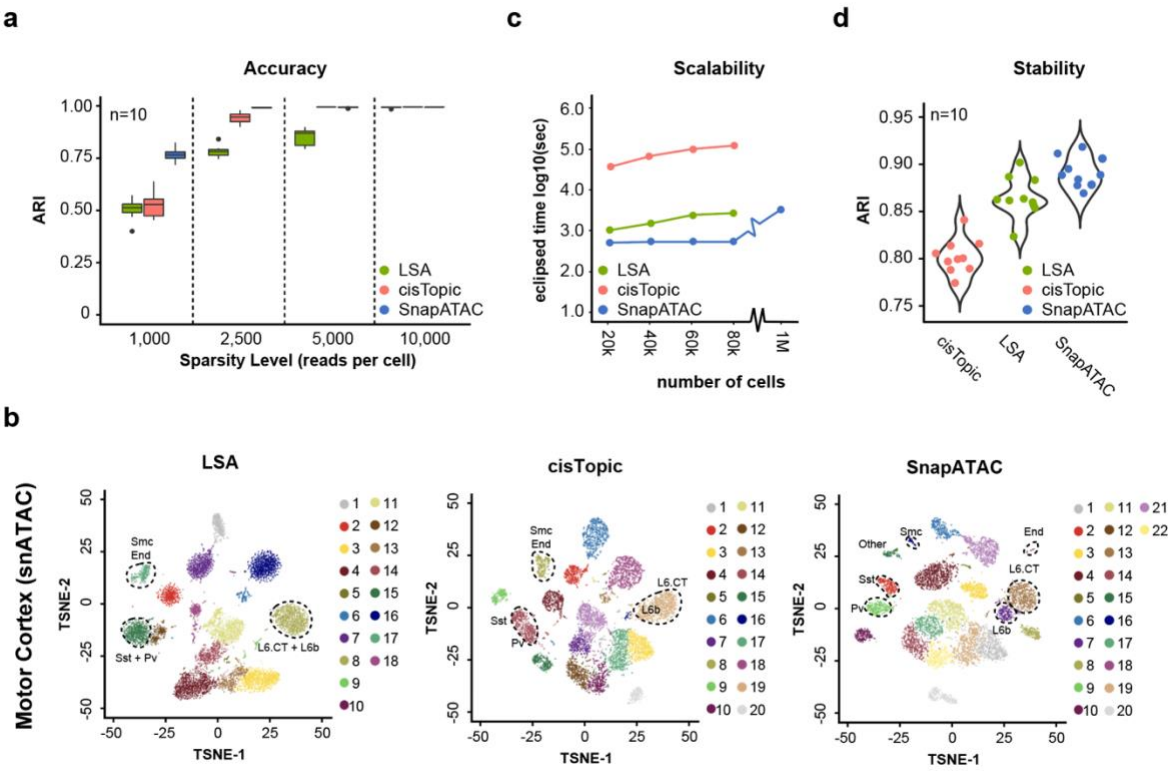
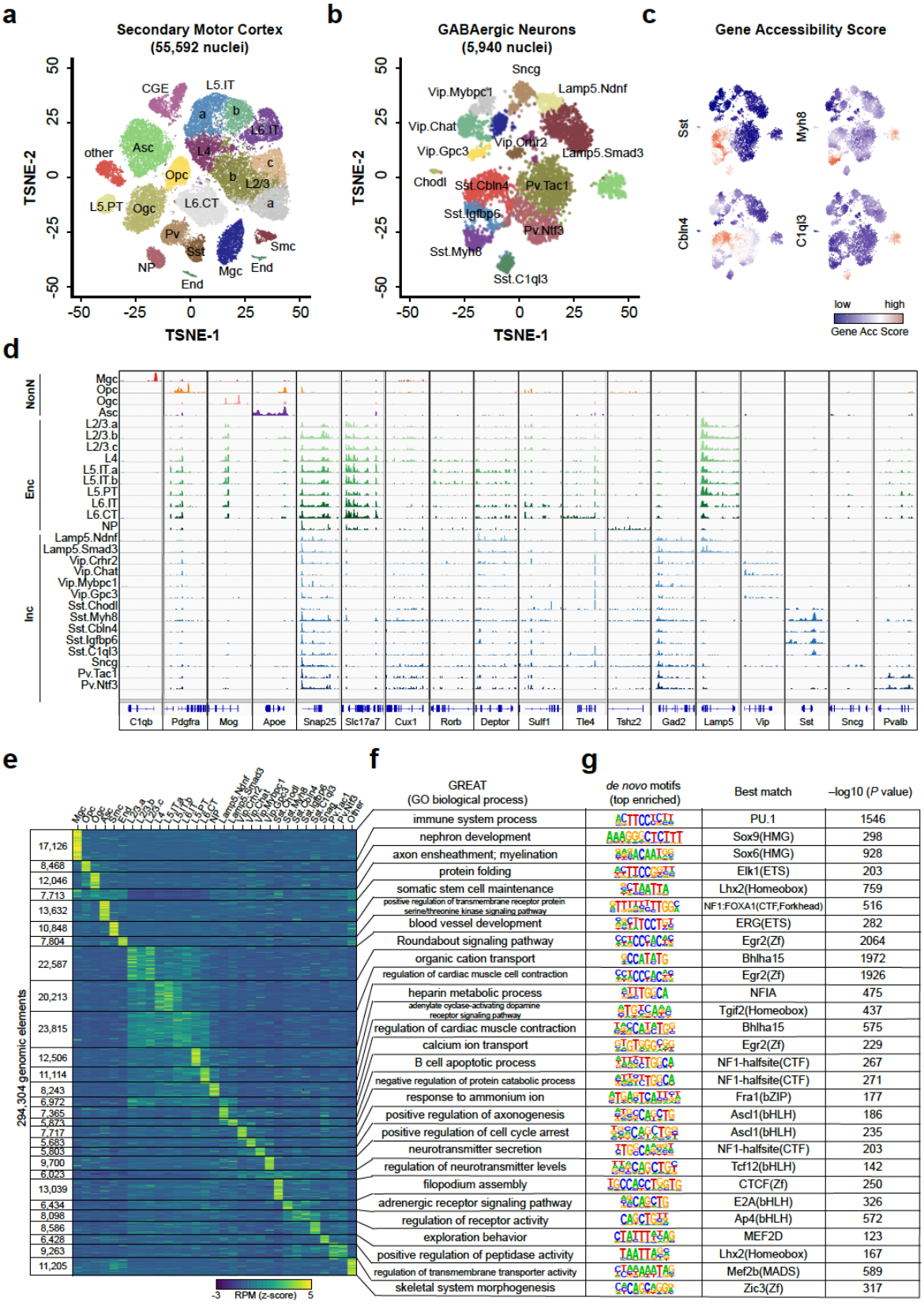


Figure 2.4. SnapATAC outperforms current methods in accuracy, sensitivity, scalability and stability of identifying cell types in complex tissues. (**a**) A set of simulated datasets are generated with varying coverage ranging from 1,000 to 10,000 reads per cell cells (**Supplementary Methods**). For each coverage, n=10 random replicates are simulated, and clustering accuracy measurement is based on Adjusted Rank Index (ARI). (**b**) T-SNE representation of an in-house dataset that contains 9,529 single nucleus ATAC-seq profiles from the mouse secondary motor cortex analyzed by LSA (left), cisTopic (middle) and SnapATAC (right). The black circles highlight the cell types only identified by SnapATAC. See also **Figure S2.12** for gene accessibility score at canonical marker genes and **Figure S2.13** for pairwise comparison of three methods. (**c**) Mouse dataset8 is sampled to different number of cells ranging from 20k to 1M. For each sampling, we compared the CPU running time of different methods for dimensionality reduction (**Supplementary Methods**). SnapATAC is the only method that is able to process a dataset of one million (1M) cells. (**d**) A set of perturbations (n=5) are introduced to the mouse dataset by down sampling to 90% of the original sequencing depth. Clustering outcomes are compared between different down sampled datasets to estimate the reproducibility.

Figure 2.5. A high-resolution cis-regulatory atlas of mouse secondary motor cortex (MOs). (**a**) T-SNE visualization of 20 cell types in MOs identified using SnapATAC. (**b**) Fourteen GABAergic subtypes revealed by iterative clustering of 5,940 GABAergic neurons (Sst, Pv and CGE). (**c**) Gene accessibility score of canonical marker genes for GABAergic subtypes projected onto the t-SNE embedding. Marker genes were identified from previous scRNA-seq analysis[34]. (**d**) Genome browser view of aggregate signal for each of the cell types. (**e**) *k*-means clustering of 294,304 differentially accessible elements based on chromatin accessibility. (**g**) Gene ontology analysis of each cell type predicted using GREAT analysis[26]. (**e**) Transcription factor motif enriched in each cell group identified using Homer[25].


a
Secondary Motor Cortex
(55,592 nuclei)
CGE
L5.IT
a
b
L6.IT
other
Asc
L4
c
Opc
b
L2/3
L5.PT
Ogc
L6.CT
a
Pv
Smc
Sst
NP
Mgc
End
End
TSNE-2
TSNE-1
b
GABAergic Neurons
(5,940 nuclei)
Sncg
Vip.Mybpc1
Lamp5.Ndnf
Vip.Chat
Vip.Crhr2
Vip.Gpc3
Lamp5.Smad3
Chodl
Sst.Cbln4
Pv.Tac1
Sst.Igfbp6
Pv.Ntf3
Sst.Myh8
Sst.C1ql3
TSNE-2
TSNE-1
c
Gene Accessibility Score
Sst
Myh8
Cbln4
C1ql3
low
high
Gene Acc Score
d
NonN
Enc
Inc
Mgc
Opc
Ogc
Asc
L2/3.a
L2/3.b
L2/3.c
L4
L5.IT.a
L5.IT.b
L5.PT
L6.IT
L6.CT
NP
Lamp5.Ndnf
Lamp5.Smad3
Vip.Crhr2
Vip.Chat
Vip.Mybpc1
Vip.Gpc3
Sst.Chodl
Sst.Myh8
Sst.Cbln4
Sst.Igfbp6
Sst.C1ql3
Sncg
Pv.Tac1
Pv.Ntf3
C1qb
Pdgfra
Mog
Apoe
Snap25
Slc17a7
Cux1
Rorb
Deptor
Sulf1
Tle4
Tshz2
Gad2
Lamp5
Vip
Sst
Sncg
Pvalb
e
294,304 genomic elements
17,126
8,468
12,046
7,713
13,632
10,848
7,804
22,587
20,213
23,815
12,506
11,114
8,243
6,972
7,365
5,873
7,717
5,683
5,803
9,700
6,023
13,039
6,434
8,098
8,586
6,428
9,263
11,205
-3
RPM (z-score)
5
f
g
GREAT
(GO biological process)
de novo motifs
(top enriched)
Best match
–log10 (P value)
immune system process
PU.1
1546
nephron development
Sox9(HMG)
298
axon ensheathment; myelination
Sox6(HMG)
928
protein folding
Elk1(ETS)
203
somatic stem cell maintenance
Lhx2(Homeobox)
759
positive regulation of transmembrane receptor protein serine/threonine kinase signaling pathway
NF1:FOXA1(CTF,Forkhead)
516
blood vessel development
ERG(ETS)
282
Roundabout signaling pathway
Egr2(Zf)
2064
organic cation transport
Bhlha15
1972
regulation of cardiac muscle cell contraction
Egr2(Zf)
1926
heparin metabolic process
NFIA
475
adenylate cyclase-activating dopamine receptor signaling pathway
Tgif2(Homeobox)
437
regulation of cardiac muscle contraction
Bhlha15
575
calcium ion transport
Egr2(Zf)
229
B cell apoptotic process
NF1-halfsite(CTF)
267
negative regulation of protein catabolic process
NF1-halfsite(CTF)
271
response to ammonium ion
Fra1(bZIP)
177
positive regulation of axonogenesis
Ascl1(bHLH)
186
positive regulation of cell cycle arrest
Ascl1(bHLH)
235
neurotransmitter secretion
NF1-halfsite(CTF)
203
regulation of neurotransmitter levels
Tcf12(bHLH)
142
filopodium assembly
CTCF(Zf)
250
adrenergic receptor signaling pathway
E2A(bHLH)
326
regulation of receptor activity
Ap4(bHLH)
572
exploration behavior
MEF2D
123
positive regulation of peptidase activity
Lhx2(Homeobox)
167
regulation of transmembrane transporter activity
Mef2b(MADS)
589
skeletal system morphogenesis
Zic3(Zf)
317

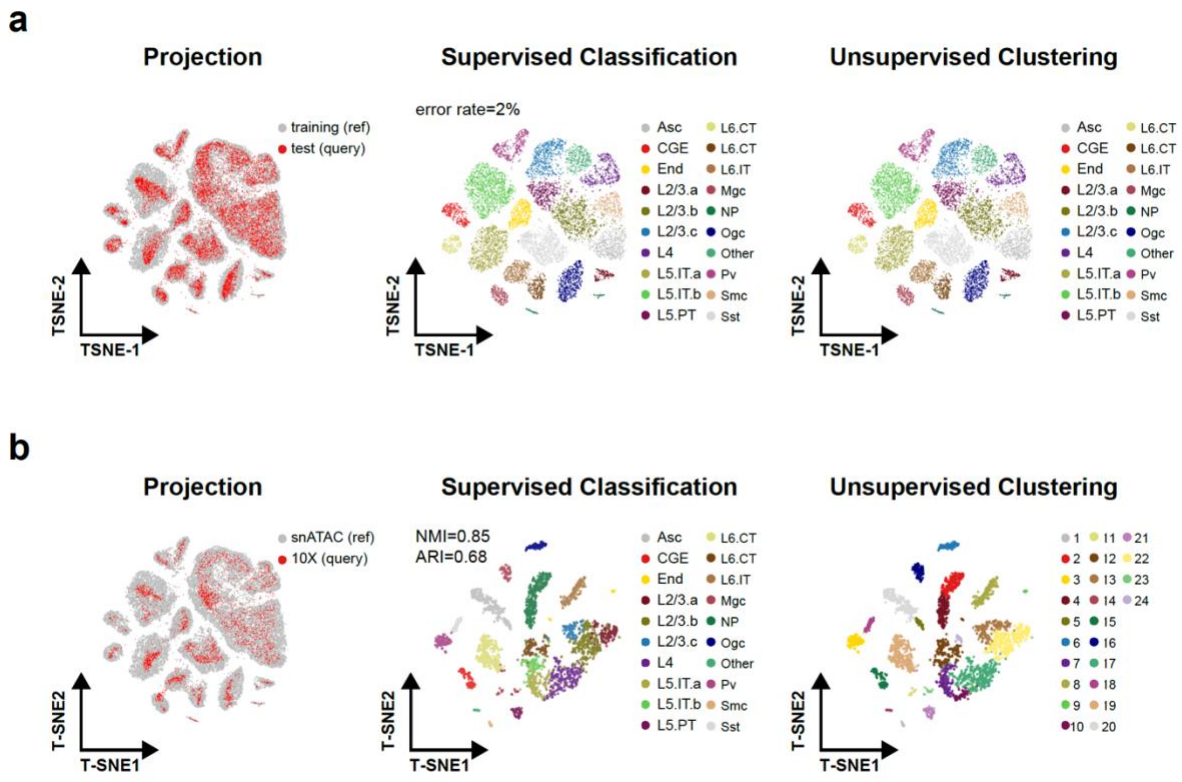


Figure 2.6. SnapATAC enables supervised annotation of new scATAC-seq dataset using reference cell atlas. (**a**) MOs snATAC-seq dataset is split into 80% and 20% as training and test dataset. A predictive model learned from the training dataset predicts cell types on the test dataset of high accuracy (error rate = 2%) as compared to the original cell type labels (right). (**b**) A predictive model learned from the reference dataset - MOs (snATAC) – accurately predicts the cell types on a query dataset from mouse brain – that is generated using a different technological platform, the 10X scATAC-seq. The t-SNE embedding is inferred from the reference cell atlas (left) or generated by SnapATAC in an unbiased manner from 10X mouse brain dataset (middle and right). Cells are visualized using t-SNE and are colored by the cell types predicted by supervised classification (middle) compared to the cluster labels defined using unsupervised clustering (right).

2.8 Supplementary Methods

Outline of the SnapATAC Pipeline. Barcode Demultiplexing. Using a custom python script, we first de-multicomplex FASTQ files by integrating the cell barcode into the read name in the following format:

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

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

Fragmentation & Filtering. Pair-end reads are converted into fragments and only those that meet the following criteria are kept: 1) properly paired (according to SMA flag value); 2) uniquely mapped (MAPQ > 30); 3) insert distance within [50-1000bp]. PCR duplicates (fragments sharing exactly the same genomic coordinates) are removed for each cell separately. Tn5 offset is then adjusted for each fragment.

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

contains snap-file version, date, alignment and reference genome information. BD session contains all unique barcodes and corresponding meta data. BM session contains cell-by-bin matrices of different resolutions. PM session contains cell-by-peak count matrix. GM session contains cell-by-gene count matrix. FM session contains all usable fragments for each cell. Fragments are indexed based on barcodes that enables fast retrieval of reads based on the barcodes.

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

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

Doublets Detection & Removal Using Scrublet (Optional). To identify doublets from single cell ATAC-seq datasets, we use doublets detection algorithm Scrublet[23]. We have found that cell-by-bin matrix can identify doublets with higher sensitivity and accuracy than cell-by-peak matrix (**Figure S2.2**). Thus, we choose to use 5kb cell-by-bin matrix as

input to identify doublets in single cell ATAC-seq dataset. Doublets detection is performed in this study when noted.

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

For each dataset, we first define the "landmark" cell types in a supervised manner. First, we perform cisTopic[15] for dimensionality reduction and identify cell clusters using graph-based algorithm Louvain[40] with k=15. Second, we manually define the major cell types in each dataset by examining the gene accessibility score at the canonical marker genes. Third, clusters sharing the same marker genes are manually merged and those failing to show unique signatures are discarded. In total, we define nine cell types in PBMC 5K (10X), 14 types in Mouse Brain 5K (10X) and 14 types in MOs M1 (snATAC). Among these cell types, 14 cell populations that account for less than 2% of the total population are considered as rare cell populations (**Figure 2.3a**)

We next evaluate the performance of bin size using three metrics: 1) cluster connectivity index (CI) which estimate the degree of connectedness of the landmark cell types; a lower CI represents a better separation; 2) coverage bias which estimates the read depth distribution in the two-dimensional embedding space; 3) sensitivity to identify

rare populations. Overall, we observe that regardless of sequencing depth and technological platform, bin size of 5kb results in optimal separation of different cell types (**Figure 2.3b-c**) and successfully identifies all rare cell populations in the dataset (**Figure 2.3d**). Therefore, we choose 5kb bins the optimal bin size in this study.

Matrix Binarization. We found the vast majority of the elements in the cell-by-bin count matrix is “0”, indicating either closed chromatin or missing value. Among the non-zero elements, some has abnormally high coverage (> 200) perhaps due to the alignment errors. These items usually account for less than 0.1% of total non-zero items in the matrix. Thus, we remove the top 0.1% items in the matrix to eliminate potential alignment errors. We next convert the remaining non-zero elements to “1”.

Bin Filtering. We next filter out any bins overlapping with the ENCODE blacklist downloaded from http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/. Second, we remove reads mapped to the X/Y chromosomes and mitochondrial DNA. Third, we observe that the bin coverage roughly obeys a log-normal distribution. We sort the bins based on the coverage and filter out the top 5% to remove the invariant features such as housekeeping gene promoters. For a dataset that has low coverage (average fragment number less than 5,000), we find the log-normal distribution does not apply, therefore, we do not perform coverage-based bin filtering.

Diffusion Maps Algorithm. We next apply diffusion maps algorithm, a nonlinear dimensionality reduction technique that discovers low dimensional manifolds by

performing harmonic analysis of a random walk in the data. A typical diffusion maps algorithm contains the following steps:

Now, let us express the diffusion maps algorithm in matrix notation. Let $\mathrm{X} \in \mathcal{R}^{n \times m}$ be a dataset with n cells and m bins and $X = \{0,1\}$. For diffusion maps algorithm, the first step is to compute a similarity matrix between the m high-dimensional data points to construct the n-by-n pairwise similarity matrix using a kernel function k that is an appropriate similarity metric. A popular choice is Gaussian kernel:

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

where $||.||$ is a distance metric to measure the distance between observations i and j.

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

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

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

$$jaccard(x_i, x_j) = jaccard(x_j, x_i)$$ (symmetric)

$$jaccard(x_i, x_j) \geq 0$$ (positivity preserving)

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

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

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

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

because $P_i \times P_j > 0$ (no empty cells allowed), then

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

The increase of either P_i or P_j will result in an increase of E_{ij} which suggests the Jaccard similarity between cells is highly affected by the read depth.

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

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

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

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

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

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

where d is the density estimate of all log10-transformed cell fragment count and x_i is the mean fragment count for cell i. Similar approach was first introduced in SCTranscform[41] to speed up the normalization of single cell RNA-seq.

We then proceed to normalize the full Jaccard coefficient matrix $J \in \mathcal{R}^{n \times n}$ using the regression model learned from y cells and compared the results to the case where all cells are used in the initial estimation step as well. We use the correlation of normalized Jaccard coefficient to compare this partial analysis to the full analysis. We observe that using as few as 2000 cells in the estimation gave rise to virtually identical estimates. We therefore use 2,000 cells in the initial model-fitting step. To remove outliers in the normalized similarity, we use the 0.99 quantile to cap the maximum value of the normalized matrix.

Next, using normalized Jaccard coefficient matrix N , we form a row-normalized matrix by:

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

where $D \in \mathcal{R}^{n \times n}$ is a diagonal matrix which is composed as $D_{i,i} = \sum_j N_{i,j}$. This allows us to compute the eigen decomposition

$$A = U\Lambda U^T$$

The columns $\varphi_i \in \mathcal{R}^n$ of $U \in \mathcal{R}^{n\times n}$ are the orthonormal eigenvectors. The diagonal matrix $\Lambda \in \mathcal{R}^{n\times n}$ has the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ in descending order as its entries.

Removing batch effects using Harmony. When the technical variability is at a larger scale than the biological variability, we apply batch effect corrector – Harmony – to eliminate such confounding factor. Given two datasets $X = \{X^1, X^2\}$ generated using different technologies, we first calculate the joint low-dimension manifold $U = \{U^1, U^2\}$ using diffusion maps as described above. We next apply Harmony[24] to U to regress out batch effect, resulting in a new harmonized embedding U^H. This is implemented as a function “runHarmony” in SnapATAC package.

Selection of Eigenvector and Eigenvalues. We next determine how many eigenvectors to include for the downstream analysis. Here we use an *ad hoc* approach for choosing the optimal number of components. We look at the scatter plot between every two pairs of eigenvectors and choose the number of eigenvectors that start exhibiting “blob”-like structure in which no obvious biological structure is revealed.

Nyström Landmark Diffusion Map. The computational cost of the diffusion maps algorithm scales exponentially with the increase of number of cells. For instance, calculating and normalizing the pair-wise kernel Matrix N becomes computationally

infeasible for large-scale dataset. To overcome this limitation, here we combine the Nyström method (a sampling technique) and diffusion maps to present Nyström Landmark diffusion map to overcome this limitation.

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

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

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

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

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

in which $N^{kk} \in \mathcal{R}^{k \times k}$ is the pairwise kernel matrix between k landmarks and $N^{kv} \in \mathcal{R}^{(n-k) \times k}$ is the similarity matrix between $(n-k)$ cells and k landmarks. Using N^{kk}, we perform diffusion map to obtain the r-rank diffusion map embedding $U^{kk} \in \mathcal{R}^{k \times r}$ by:

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

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

where $D^{kk} \in \mathcal{R}^{k \times k}$ is a diagonal matrix which is composed as $D^{kk}_{i,i} = \sum_j N^{kk}_{ij}$.

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

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

where $D^{kv} \in \mathcal{R}^{(n-k)\times(n-k)}$ is a diagonal matrix which is composed as $D_{i,i}^{kv} = \sum_j N_{i,j}^{kv}$.

$$U^{kv} = A^{kv}U^{kk}/\Lambda^{kk}$$

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

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

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

Optimizing the Number of Landmarks. To evaluate the effect of number of landmarks, we apply our method to a complex dataset that contains over 80k cells from 13 different mouse tissues. We employ the following three metrics to evaluate the performance. First, using different number of landmarks (k) ranging from 1,000 to 10,000, we compare the clustering outcome to the cell type label defined in the original study. The

goal of this is to identify the “elbow” point that performance drops abruptly. Second, for each sampling, we repeat for five times using different set of landmarks to evaluate stability between sampling. Third, we spiked in 1% Patski cells to assess the sensitivity of identifying rare cell types. We choose Patski cells because these cells were profiled using the same protocol by the same group[5] to minimize the batch effect

We observe that using as few as 5,000 landmarks can largely recapitulate the result obtained using 10,000 landmarks (**Figure 2.6a**), and 10,000 landmarks can achieve highly robust embedding between sampling (**Figure 2.6b**) and successfully recover spiked-in rare populations (**Figure 2.6c**) without showing batch effect between replicates (**Figure 2.6d**). To obtain a reliable low-dimensional embedding, we use 10,000 landmarks for all the analysis performed in this study. We next apply our method to another three large-scale datasets (**Figure 2.7**). SnapATAC can identify substantial heterogeneity, suggesting the generality of our method.

Ensemble Nyström Method. Nyström method is stochastic in its nature, different sampling will result in different embedding and clustering outcome. To improve the robustness of the clustering method, we next employ Ensemble Nyström Algorithm which combines a mixture of Nyström approximation to create an ensemble representation. Supported by theoretical analysis, this Ensemble approach has been demonstrated to guarantee a convergence and in a faster rate in comparison to standard Nyström method. Moreover, this ensemble algorithm naturally fits within distributed computing

environments, where their computational costs are roughly the same as that of the standard Nyström single sampling method.

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

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

$$\tilde{G}^{en} = \sum_{j=1}^{p} \mu^j \tilde{G}^j$$

where μ^j is the mixture weights that can be defined in many ways. Here we choose to use the most straightforward method by assigning an equal weight to each of the KNN-graph obtained from different samplings, $\mu^j = 1/p, r \in [1, p]$. While this choice ignores the relative quality of each Nyström approximation, it is computational efficient and already generates a solution superior to any one of the approximations used in the

combination. Using the ensemble weighted KNN graph $\tilde{G}^{en}$, we next apply community finding algorithm to identify cell clusters.

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

Gene Accessibility Score. To annotate the identified clusters, SnapATAC calculated the gene-body accessibility matrix G using "calGmatFromMat" function in SnapATAC packge where $G_{i,j}$ is the number of fragments overlapping with j-th genes in i-th cell. $G_{i,j}$ is then normalized to CPM (count-per-million reads) as $\tilde{G}$. The normalized accessibility score is then smoothed using Markov affinity-graph based method:

$$\hat{G}=\tilde{G}A^{t}$$

where A is the adjacent matrix obtained from K nearest neighbor graph and t is number of steps taken for Markov diffusion process. We set $t = 3$ in this study.

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

Motif Analysis. SnapATAC incorporates chromVAR to estimate the motif variability and Homer for *de novo* motif discovery. This is implemented as function “runChromVAR” and “runHomer” in SnapATAC package.

Identification of differentially accessible peaks. For a given group of cells C_i, we first look for their neighboring cells C_j ($|C_i| = |C_j|$) in the diffusion component space as “background” cells to compare to. If C_i accounts for more than half of the total cells, we use the remaining cells as local background. Next, we aggregate C_i and C_j to create two raw-count vectors as V_{ci} and V_{cj} .We then perform differential analysis between V_{ci} and V_{cj} using exact test as implemented in R package edgeR (v3.18.1) with BCV=0.1. P-value is then adjusted into False Discovery Rate (FDR) using Benjamini-Hochberg correction. Peaks with FDR less than 0.01 are selected as significant DARs. However, the statically significance is under powered for small clusters.

GREAT analysis. SnapATAC incorporates GREAT analysis to infer the candidate biological pathway active in each cell populations. This is implemented as function “runGREAT” SnapATAC package.

Integration with single cell RNA-seq. We use canonical correlation analysis (CCA) embedded in Seurat V3[18] to integrate single cell RNA-seq and single cell ATAC-seq. We first calculate the gene accessibility account at variable genes identified using single cell RNA-seq dataset. This can be done using a function called

"createGmatFromMat" in SnapATAC package. Next, SnapATAC converts the snap object to a Seurat v3 object using a function called "SnapToSeurat" in preparation for integration. Different from integration method in Seurat, we use the diffusion maps embedding as the dimensionality reduction method in the Seurat object. We next follow the vignette in Seurat website (https://satijalab.org/seurat/v3.0/atacseq_integration_vignette.html) to integrate these two modalities. The cell type for scATAC-seq is predicted using function "TransferData" in Seurat V3.

Finally, for each single cell ATAC profile, we infer its gene expression profile by calculating the weighted average expression profile of its nearest neighboring cells in the single cell RNA-seq dataset[18]. By doing so, we create pseudo-cells that contain information of both chromatin accessibility and gene expression profiles. The imputation of gene expression profile is done by "TransferData" function in Seurat V3.

Linking distal elements to putative target genes. Using the "pseudo" cells, we next sought to predict the putative target genes for regulatory elements based on the association between expression of a gene and chromatin accessibility at its distal elements. Given a gene G, we first identify its surrounding regulatory elements within 1MB window flanking G. Let Y^G be the imputed gene expression value for gene G among n cells. We perform logistic regression using Y^G as variable to predict the binary state for each of peaks surrounding G. The idea behind using logistic regression is that if there is a relationship between the gene expression (continuous variable) and chromatin

accessibility (categorical variable), we should be able to predict chromatin accessibility from the gene expression. Logistic regression does not make many of the key assumptions such as normality of the continuous variables. In addition, since we only have one variable (gene expression) for prediction every time, there is no problem of multicollinearity.

We next fit logistic regression between each of flanking peak and gene expression using “glm” function in R with binomial(link='logit') as the family function. By doing so, we obtain the regression coefficient β_1 and its corresponding P-value for each peak separately. Here we used 5e-8, a standard P-value cutoff for human genome-wise association study to determine the significant association. While this cutoff is less sample or gene specific compared to more complicated methods such as permutation test, it is computational efficient and already generates a reasonable set of gene-enhancer pairings.

To evaluate the performance of our methods, we compare our prediction with cis-eQTL derived from interferon-γ and lipopolysaccharide stimulation of monocytes. Significant cis-eQTL associations are downloaded from supplementary material in Fairfax (2014). We filter cis-eQTL based on two criteria: 1) only cis-eQTLs that overlap with the peaks identified in PBMC dataset are considered; 2) In addition, we only keep the cis-eQTLs whose genes overlap with the variable genes determined by scRNA-seq. This filtering reduced the cis-eQTL list to 456 hits.

Next, we estimate the association for each of cis-eQTLs by preforming logistic regression test as described above. To make a comparison, we derive a set of negative pairs matched for the distance. For instance, given a SNP at 100kb upstream of its target gene, we look for another pair that has the same distance but downstream to this gene.

Simulation of scATAC-seq datasets. First, we download the alignment files (bam files) for ten bulk ATAC-seq experiment from ENCODE. From each bam file, we simulate 1,000 single cell ATAC-seq datasets by randomly down sampling to a variety of coverages ranging from 1,000 to 10,000 reads per cells. We next create a cell-by-bin matrix of 5kb which is used for SnapATAC clustering. Merging peaks identified from each bulk experiment, we create cell-by-peak matrix used for LSA, Cis-Topic, Cicero and chromVAR for clustering. We repeat the sampling for n=10 times to estimate the variability of the clustering.

Comparison of scalability. To compare the scalability between SnapATAC to other methods, we next simulate multiple datasets of different number of cells ranging from 20k to 1M. We simulate these datasets in the following manner. Using the 80k mouse atlas dataset, we randomly sample this dataset to different number of cells ranging from 20k to 1M cells. For the sampling that has cells more than 80K, we sample with replacement and introduce perturbation to each cell by randomly removing 1% of the “1”s in each of the cells. This removes the duplicate cells and largely maintains the density of the matrix.

For each sampling, we then perform dimensionality reduction using LSA and cisTopic and compare their CPU running time. Specifically, we monitor the running time for 1) TF-IDF transformation and Singular Value Decomposition (SVD) for LSA, 2) function “runModels” with topics = c(2, 5, 10, 15, 20, 25, 30, 35, 40) and “selectModel” function in cisTopic. The time for matrix loading is not counted.

SnapATAC is implemented in the following way to allow processing large-scale dataset. We apply SnapATAC to the sampled datasets using a custom R script. Given 1M cells, we first randomly split the 1M cells into 100 chunks with each containing 10K cells. This can be done during the preprocessing by splitting the master bam file into multiple small bam files and generating multiple “snap” files. We next randomly sample 10K cells from 1M cells as landmarks using density-based sampling approach as described above. We next perform diffusion maps embedding to landmarks using function “runDiffusionMaps” in SnapATAC package. The “snap” object for landmarks is saved as “rds” file to the disk. We then compute the low dimension embedding for each of the 10K cells by projecting onto the diffusion maps embedding learned from the landmarks using function “runDiffusionMapsExtension” in SnapATAC package. This streaming-like process 1) avoids loading of the entire large cell matrix of 1M cells and 2) requires limited memory for each processor, representing a symmetrical approach for analyzing very large-scale datasets. The time for matrix loading and preprocessing is not counted for the comparison. All the comparisons were tested on a machine with 5 AMD Operon (TM) Processor 6276 CPUs.

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

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

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

$$U^l = A^l U^k / \Lambda^k$$

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

In the joint embedding space $[U^k, U^l]$, we next identify the mutual nearest neighbors between query and landmark cells. For each cell $i_1 \in \mathrm{X}^k$ belonging to the landmarks, we find the $k.nearest$ (5) cells in the query dataset with the smallest distances

to i_1. We do the same for each cell in query cell dataset to find its $k.nearest$ (5) neighbors in the landmark dataset. If a pair of cells from each dataset is contained in each other's nearest neighbors, those cells are considered to be mutual nearest neighbors or MNN pairs (or "anchors"). We interpret these pairs as containing cells that belong to the same cell type or state despite being generated in both landmark and query cells. Thus, any differences between cells in MNN pairs should theoretically represent the non-overlapping cell types. Here we removed any query cells that failed to identify an MNN pair correspondence in the reference dataset.

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

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

where $dist(c,i)$ is the Euclidean distance in the joint diffusion maps embedding space and S_{ai} is the weight for the corresponding MNN pair (anchor). We then apply a Gaussian kernel:

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

where sd is set to 1 by default. Finally, we normalize across all s anchors:

$$W_{c,i} = \frac{\widetilde{D_{c,\imath}}}{\sum_1^{j=s} \widetilde{D_{c,J}}}$$

we set $s = 50$.

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

$$P^l = LW^T$$

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

distance between cells in the inferred t-SNE coordinate does not neccessarily reflect the cell-to-cell relationship.

Tissue collection & nuclei isolation. Adult C57BL/6J male mice were purchased from Jackson Laboratories. Brains were extracted from P56-63 old mice and immediately sectioned into 0.6 mm coronal sections, starting at the frontal pole, in ice-cold dissection media. The secondary motor cortex (MOs) region was dissected from the first three slices along the anterior-posterior axis according to the Allen Brain reference Atlas (http://mouse.brain-map.org/). Slices were kept in ice-cold dissection media during dissection and immediately frozen in dry ice for posterior pooling and nuclei production. For nuclei isolation, the MOs dissected regions from 15-23 animals were pooled, and two biological replicas were processed for each slice. Nuclei were isolated as described in previous studies, except no sucrose gradient purification was performed. Flow cytometry analysis of brain nuclei was performed as described in Luo et al.

Tn5 transposase purification & loading. Tn5 transposase was expressed as an intein chitin-binding domain fusion and purified using an improved version of the method first described by Picelli et al. T7 Express lysY/I (C3013I, NEB) cells were transformed with the plasmid pTXB1-ecTn5 E54K L372P (#60240, Addgene). An LB Ampicillin culture was inoculated with three colonies and grown overnight at 37°C. The starter culture was diluted to an OD of 0.02 with fresh media and shaken at 37°C until it reached an OD of 0.9. The culture was then immediately chilled on ice to 10°C and expression was induced by adding 250 μM IPTG (Dioxane Free, CI8280-13, Denville Scientific). The culture was

shaken for 4 hours at 23°C after which cells were harvested in 2 L batches by centrifugation, flash frozen in liquid nitrogen and stored at -80°C. Cell pellets were resuspended in 20 ml of ice cold lysis buffer (20 mM HEPES 7.2-KOH, 0.8 M NaCl, 1 mM EDTA, 10% Glycerol, 0.2% Triton X-100) with protease inhibitors (cOmplete, EDTA-free Protease Inhibitor Cocktail Tablets, 11873580001, Roche Diagnostics) and passed three times through a Microfluidizer (lining covered with ice water, Model 110L, Microfluidics) with a 5 minute cool down interval in between each pass. Any remaining sample was purged from the Microfluidizer with an additional 25 ml of ice-cold lysis buffer with protease inhibitors (total lysate volume ~50ml). Samples were spun down for 20 min in an ultracentrifuge at 40K rpm (L-80XP, 45 Ti Rotor, Beckman Coulter) at 4°C. ~45 ml of supernatant was combined with 115 ml ice cold lysis buffer with protease inhibitors in a cold beaker (total volume = 160 ml) and stirred at 4°C. 4.2ml of 10% neutralized polyethyleneimine-HCl (pH 7.0) was then added dropwise. Samples were spun down again for 20 min in an ultracentrifuge at 40K rpm (L-80XP, 45 Ti Rotor, Beckman Coulter) at 4°C. The pooled supernatant was loaded onto ~10ml of fresh Chitin resin (S6651L, NEB) in a chromatography column (Econo-Column (1.5 × 15 cm), Flow Adapter: 7380015, Bio-Rad). The column was then washed with 50-100 ml lysis buffer. Cleavage of the fusion protein was initiated by flowing ~20ml of freshly made elution buffer (20 mM HEPES 7.2-KOH, 0.5 M NaCl, 1 mM EDTA, 10% glycerol, 0.02% Triton X-100, 100mM DTT) onto the column at a speed of 0.8ml/min for 25 min. After the column was incubated for 63 hrs at 4°C, the protein was recovered from the initial elution volume and a subsequent 30 ml wash with elution buffer. Protein-containing fractions were pooled and diluted 1:1 with buffer [20 mM HEPES 7.2-KOH,1 mM EDTA, 10% glycerol, 0.5mM TCEP)

to reduce the NaCl concentration to 250mM. For cation exchange, the sample was loaded onto a 1ml column HiTrap S HP (17115101, GE), washed with Buffer A (10mM Tris 7.5, 280 mM NaCl, 10% glycerol, 0.5mM TCEP) and then eluted using a gradient formed using Buffer A and Buffer B (10mM Tris 7.5, 1M NaCl, 10% glycerol, 0.5mM TCEP) (0% Buffer B over 5 column volumes, 0-100% Buffer B over 50 column volumes, 100% Buffer B over 10 column volumes). Next, the protein-containing fractions were combined, concentrated via ultrafiltration to ~1.5 mg/mL and further purified via gel filtration (HiLoad 16/600 Superdex 75 pg column (28989333, GE)) in Buffer GF (100mM HEPES-KOH at pH 7.2, 0.5 M NaCl, 0.2 mM EDTA, 2mM DTT, 20% glycerol). The purest Tn5 transposase-containing fractions were pooled and 1 volume 100% glycerol was added to the preparation. Tn5 transposase was stored at -20°C.

To generate Tn5 transposomes for combinatorial barcoding assisted single nuclei ATAC-seq, barcoded oligos were first annealed to pMENTs oligos (95 °C for 5 min, cooled to 14 °C at a cooling rate of 0.1 °C/s) separately. Next, 1 µl barcoded transposon (50 µM) was mixed with 7 ul Tn5 (~7 µM). The mixture was incubated on the lab bench at room temperature for 30 min. Finally, T5 and T7 transposomes were mixed in a 1:1 ratio and diluted 1:10 with dilution buffer (50 % Glycerol, 50 mM Tris-HCl (pH=7.5), 100 mM NaCl, 0.1 mM EDTA, 0.1 % Triton X-100, 1 mM DTT). For combinatorial barcoding, we used eight different T5 transposomes and 12 distinct T7 transposomes, which eventually resulted in 96 Tn5 barcode combinations per sample.

Bulk ATAC-seq data generation. ATAC-seq was performed on 30,000-50,000 nuclei as described previously with modifications. Nuclei were thawed on ice and pelleted for 5 min at 500 x g at 4 °C. Nuclei pellets were resuspended in 30 µl tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM K-acetate, 11 mM Mg-acetate, 17.6 % DMF) and counted on a hemocytometer. 30,000-50,000 nuclei were used for tagmentation and the reaction volume was adjusted to 19 µl using tagmentation buffer. After addition of 1 µl TDE1 (Illumina FC-121-1030), tagmentation was performed at 37°C for 60 min with shaking (500 rpm). Tagmented DNA was purified using MinElute columns (Qiagen), PCR-amplified for 8 cycles with NEBNext® High-Fidelity 2X PCR Master Mix (NEB, 72°C 5 min, 98°C 30 s, [98°C 10 s, 63°C 30 s, 72°C 60 s] x 8 cycles, 12°C held). Amplified libraries were purified using MinElute columns (Qiagen) and SPRI Beads (Beckmann Coulter). Sequencing was carried out on a NextSeq500 using a 150-cycle kit (75 bp PE, Illumina).

Bulk ATAC-seq data analysis. ATAC-seq reads were mapped to reference genome mm10 using BWA and *samtools* version 1.2 to eliminate PCR duplicates and mitochondrial reads. The paired end read ends were converted to fragments. Using fragments, MACS2 version 2.1.2 was used for generating signal tracks and peak calling with the following parameters: --nomodel --shift 100 --ext 200 --qval 1e-2 -B –SPMR.

Single-nucleus ATAC-seq data generation. Combinatorial ATAC-seq was performed as described previously with modifications. For each sample two biological replicates were processed. Nuclei were pelleted with a swinging bucket centrifuge (500 x

g, 5 min, 4°C; 5920R, Eppendorf). Nuclei pellets were resuspended in 1 ml nuclei permeabilization buffer (5 % BSA, 0.2 % IGEPAL-CA630, 1mM DTT and cOmpleteTM, EDTA-free protease inhibitor cocktail (Roche) in PBS) and pelleted again (500 x g, 5 min, 4°C; 5920R, Eppendorf). Nuclei were resuspended in 500 µL high salt tagmentation buffer (36.3 mM Tris-acetate (pH = 7.8), 72.6 mM potassium-acetate, 11 mM Mg-acetate, 17.6% DMF) and counted using a hemocytometer. Concentration was adjusted to 4500 nuclei/9 µl, and 4,500 nuclei were dispensed into each well of a 96-well plate. Glycerol was added to the leftover nuclei suspension for a final concentration of 25 % and nuclei were stored at -80°C. For tagmentation, 1 µL barcoded Tn5 transposomes were added using a BenchSmart™ 96 (Mettler Toledo), mixed five times and incubated for 60 min at 37 °C with shaking (500 rpm). To inhibit the Tn5 reaction, 10 µL of 40 mM EDTA were added to each well with a BenchSmart™ 96 (Mettler Toledo) and the plate was incubated at 37 °C for 15 min with shaking (500 rpm). Next, 20 µL 2 x sort buffer (2 % BSA, 2 mM EDTA in PBS) were added using a BenchSmart™ 96 (Mettler Toledo). All wells were combined into a FACS tube and stained with 3 µM Draq7 (Cell Signaling). Using a SH800 (Sony), 20 nuclei were sorted per well into eight 96-well plates (total of 768 wells) containing 10.5 µL EB (25 pmol primer i7, 25 pmol primer i5, 200 ng BSA (Sigma). Preparation of sort plates and all downstream pipetting steps were performed on a Biomek i7 Automated Workstation (Beckman Coulter). After addition of 1 µL 0.2% SDS, samples were incubated at 55 °C for 7 min with shaking (500 rpm). We added 1 µL 12.5% Triton-X to each well to quench the SDS and 12.5 µL NEBNext High-Fidelity 2× PCR Master Mix (NEB). Samples were PCR-amplified (72 °C 5 min, 98 °C 30 s, (98 °C 10 s, 63 °C 30 s, 72 °C 60 s) × 12 cycles, held at 12 °C). After PCR, all wells were combined. Libraries

were purified according to the MinElute PCR Purification Kit manual (Qiagen) using a vacuum manifold (QIAvac 24 plus, Qiagen) and size selection was performed with SPRI Beads (Beckmann Coulter, 0.55x and 1.5x). Libraries were purified one more time with SPRI Beads (Beckmann Coulter, 1.5x). Libraries were quantified using a Qubit fluorimeter (Life technologies) and the nucleosomal pattern was verified using a Tapestation (High Sensitivity D1000, Agilent). The library was sequenced on a HiSeq2500 sequencer (Illumina) using custom sequencing primers, 25% spike-in library and following read lengths: 50 + 43 + 40 + 50 (Read1 + Index1 + Index2 + Read2).

2.9 Supplementary Figures

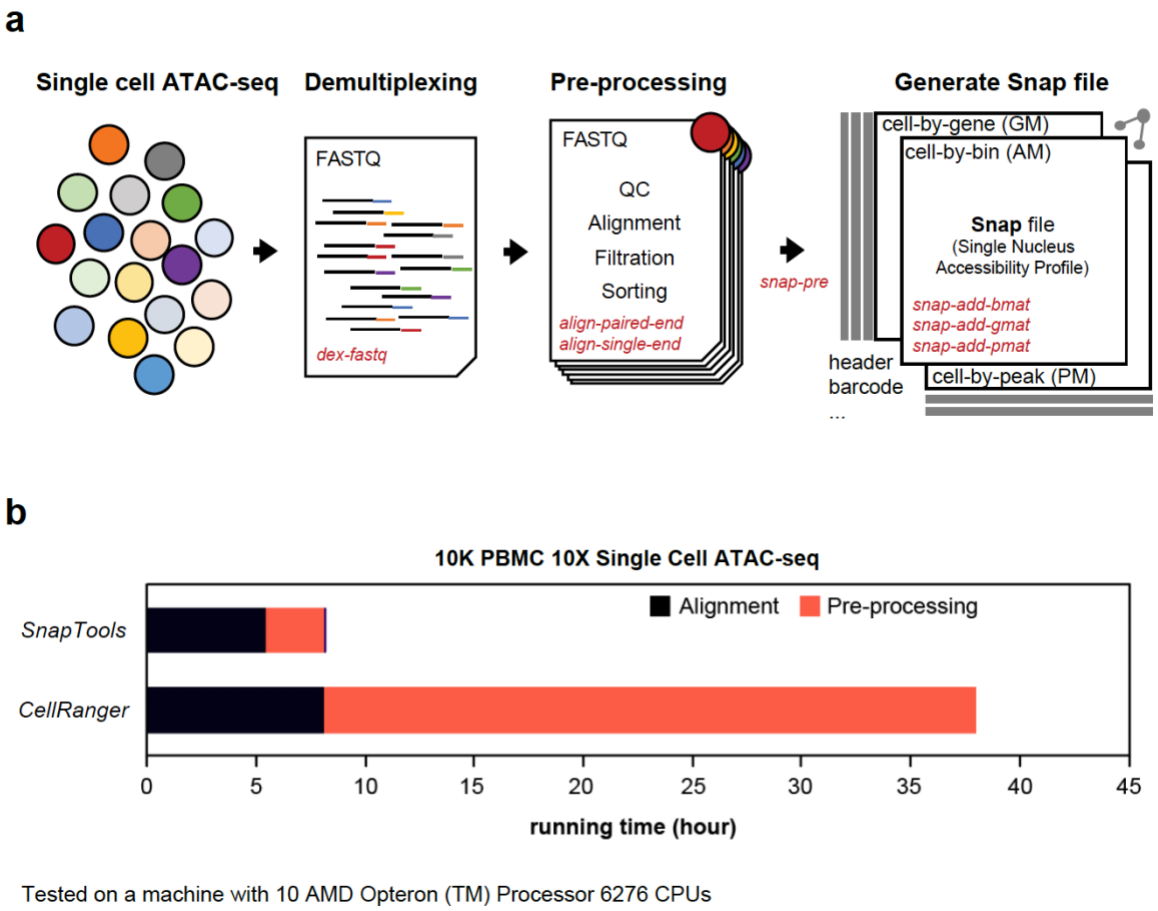


Figure S2.1. Overview of SnapTools workflow. (**a**) Demultiplexing: SnapTools first demultiplexed the fastq files by adding the cell barcodes to the beginning of each read name; Pre-processing: raw sequencing reads were aligned to the reference genome using BWA followed by filtration of erroneous alignments. A snap file was generated to store indexed reads and multiple cell matrices including cell-by-peak, cell-by-gene and cell-by-bin matrix. (**b**) Running time comparison between SnapTools and alternative method – cellRanger for alignment and preprocessing. Both methods were tested on a machine with 10 AMD Opteron (TM) Processor 6276 CPUs using 10K PBMC dataset (10X v1) from 10X genomics.

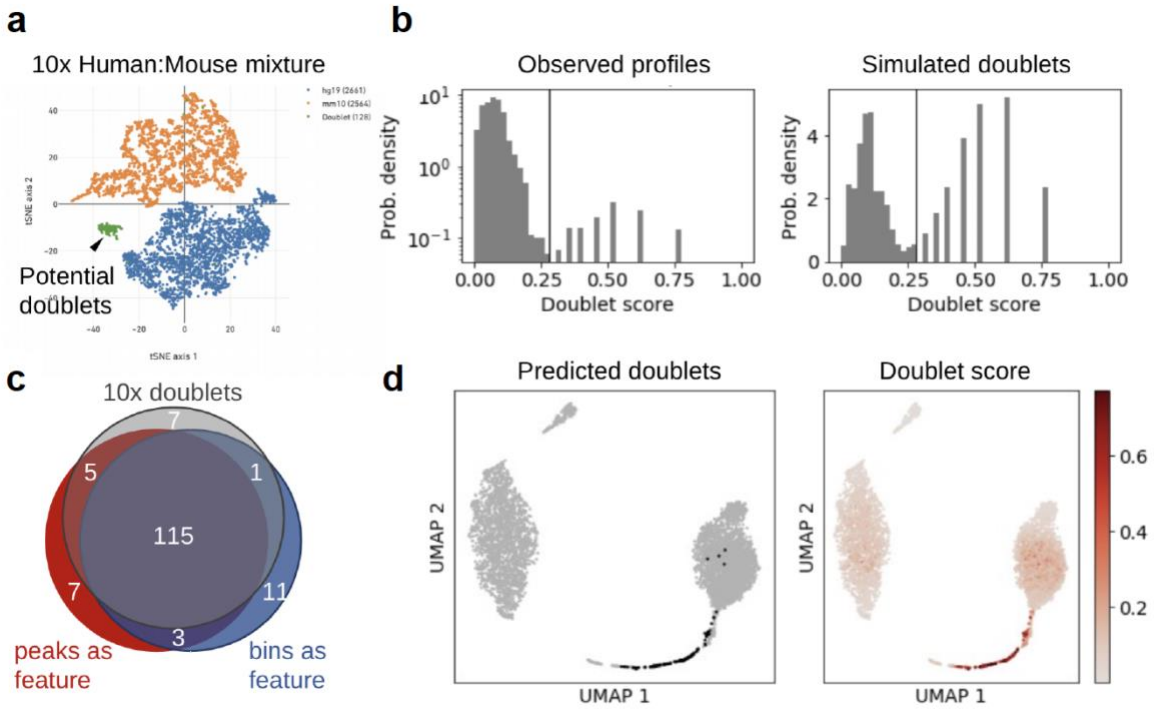
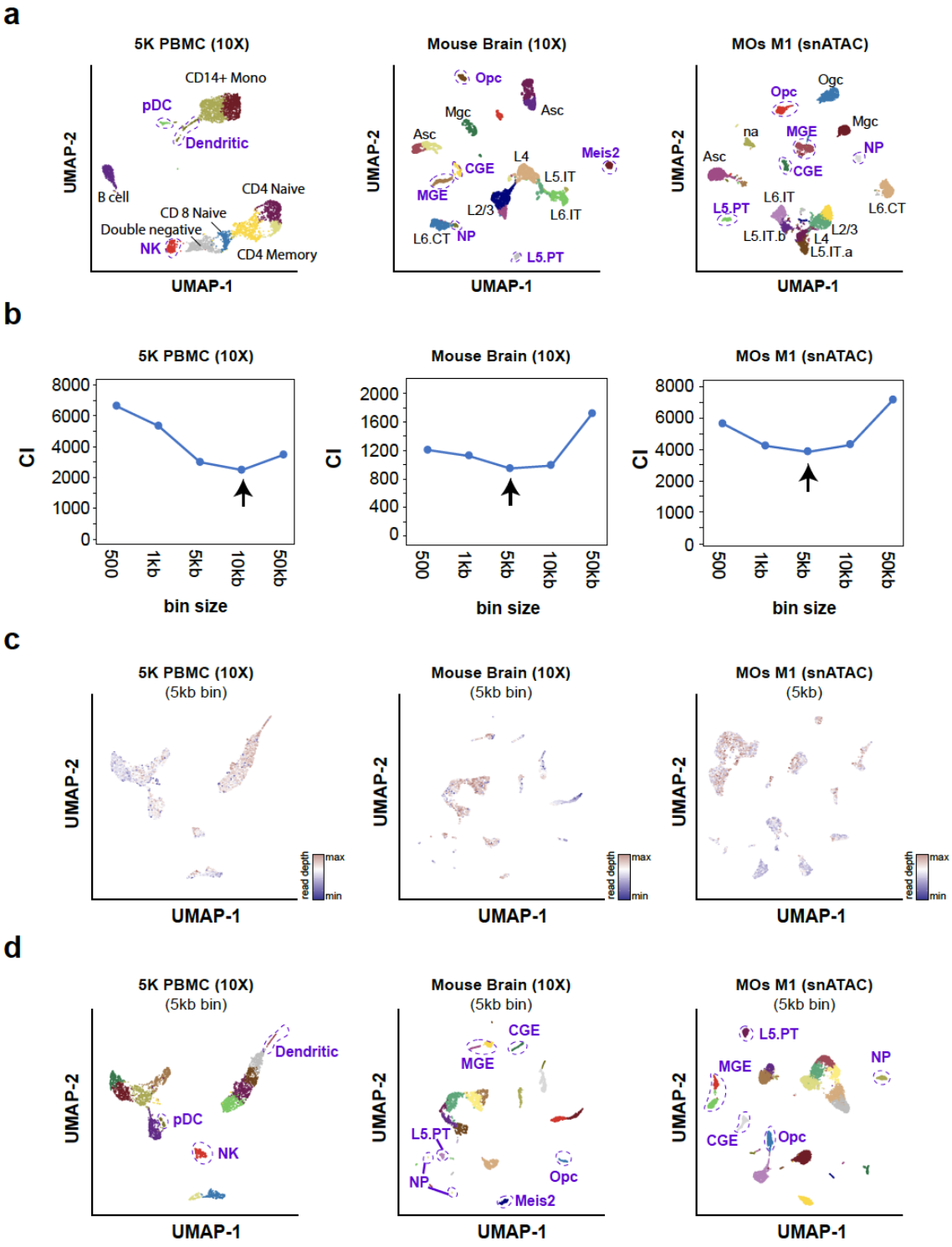


Figure S2.2. SnapATAC removes putative doublets using Scrublet. (**a**) T-SNE representation of a dataset (hgmm_1k 10X) that contained 1,000 human (GM12878) and mouse (A20) cells. Cells are colored by species determined based on the alignment ratio between human and mouse genome. Orange: A20; blue: GM12878; green: putative doublets. (**b**) Distribution of doublet score for putative doublets and simulated doublets estimated using Scrublet. (**c**) Doublets are predicted using cell-by-peak and cell-by-bin matrix separately. Venn diagram show the overlap between Scrublet-predicted doublets using peak or bin matrix and doublets identified based on alignment ratio. (**d**) Doublets scores projected onto the UMAP embedding.

Figure S2.3. Choosing the optimal bin size. (**a**) UMAP visualization of landmark cell types identified in three benchmarking datasets. UMAP embedding was computed using cisTopic and cell types were manually annotated based on the gene accessibility score at canonical marker genes (**Supplementary Methods**). See also **Figure S2.12, S2.14, S2.16** for corresponding gene accessibility score plot. Blue dash line highlights the rare cell populations that account for less than 2% of the total population. (**b**) Relationship between connectivity index (CI) and bin sizes. Connectivity index were calculated between landmark cell types in the reduced decimation using function “connectivity” in R package “clv”. A lower CI indicates a better separation of landmark cell types. (**c**) UMAP representation of three benchmarking datasets generated using SnapATAC using 5kb bin size. Cells colored by read depth to illustrate the sequencing depth effect. (**d**) Cells are colored by cluster labels identified by SnapATAC.



* blue circles highlight rare cell populations account for less than 2% of total population

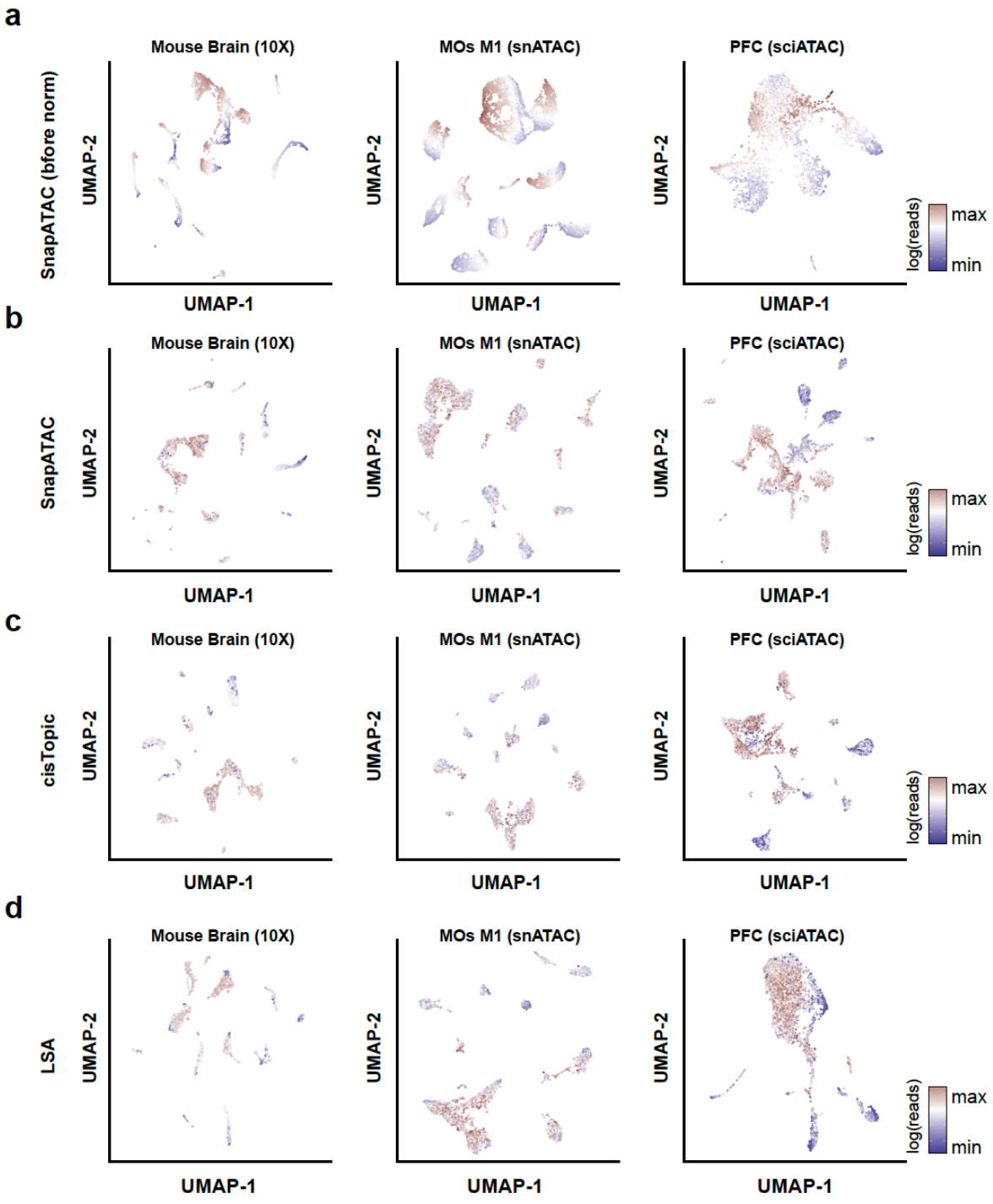


Figure S2.4. SnapATAC is robust to sequencing depth. Two dimensional UMAP representation of three benchmarking datasets analyzed by four methods (**a**) SnapATAC without normalization; (**b**) SnapATAC with normalization; (**c**) cisTopic and (**d**) Latent Sematic Analysis (LSA). Cells are color by log-scaled read depth.

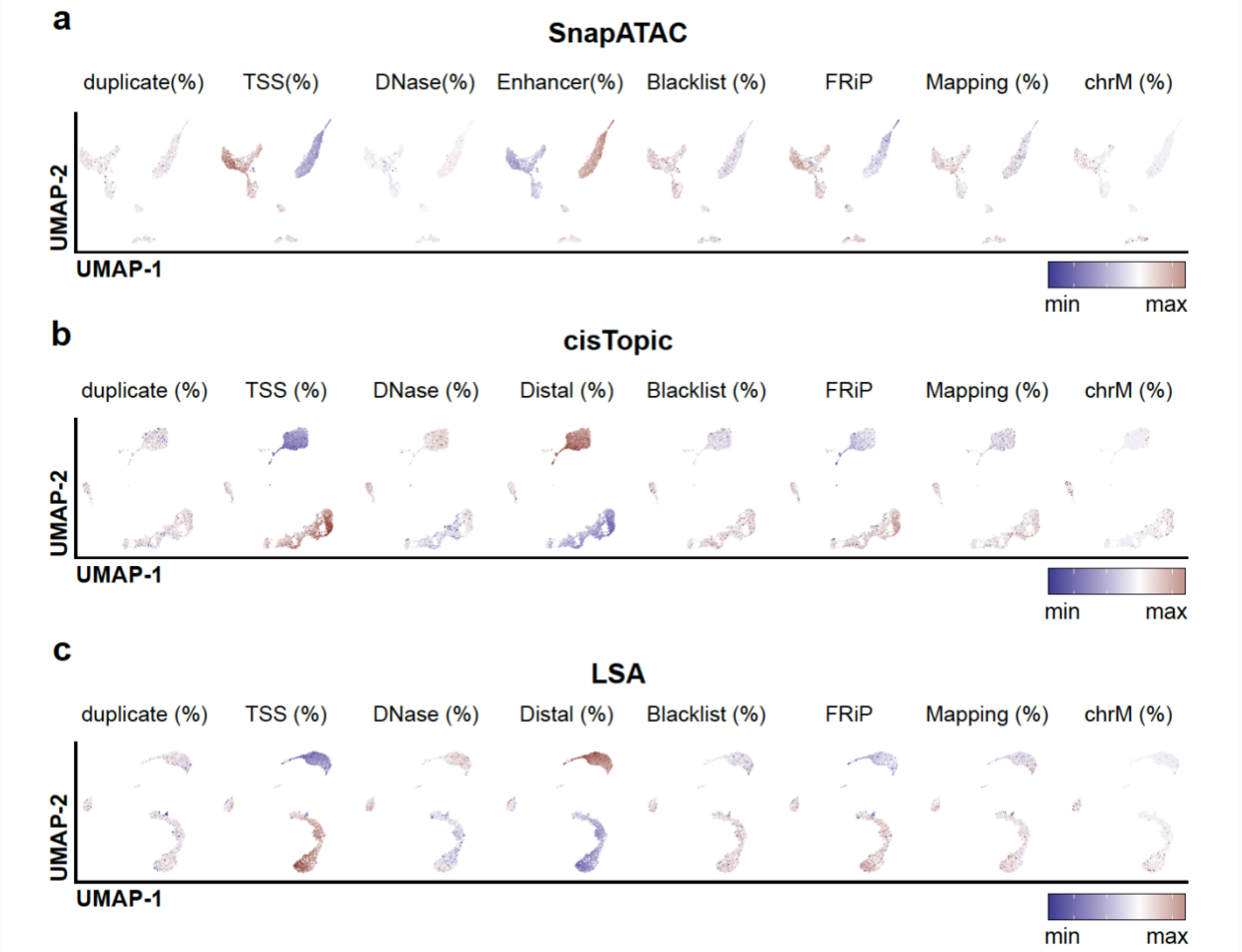


Figure S2.5. SnapATAC is robust to other biases. Potential bias in single cell ATAC-seq dataset projected onto the UMAP visualization generated using different analysis methods (**a**) SnapATAC (**b**) cisTopic and (**c**) LSA. Duplicate: percentage of fragments that are PCR duplicates. TSS: percentage of fragments overlapping or are within 1kb of a TSS. TSS position is based on the GENECODE V28 (Ensemble 92). DNase: the percentage of fragments overlapping a master DNase peak list. The DNase peak list is created by combining all ENCODE[1] DNase peaks from hg19. Blacklist: the percentage of fragments overlapping with the ENCODE blacklist. FRiP: the percentage of fragments overlapping with the peaks defined from the aggregate signal. Mapping: the percentage of fragments that are uniquely mapped. chrM: the percentage of fragments mapped to mitochondria DNA.

Figure S2.6. Nyström sampling improves the scalability without sacrificing the performance. (**a**) A line plot comparing the performance of clustering using various sampling parameters. The performance is evaluated using Adjusted Rank Index (ARI). SnapATAC was applied to the mouse atlas dataset that contained over 80k cells using different number of landmark cells (*k*) ranging from 1k to 10k. For each *k*, we performed clustering for n=5 times using different sets of randomly selected landmarks. (**b**) A line plot comparing the stability of clustering results between five samplings (pairwise comparison n=10). (**c**) To evaluate the sensitivity of identifying rare cell types, we spiked in 1% mouse Pastki cells generated using the same protocol in Cusanovich 2015[5] and this rare cell population was recapitulated using 10,000 landmarks (right). (**d**) Two-dimensional t-SNE representation of 80,000 mouse atlas cells colored by cluster labels identified using SnapATAC (left) and biological replicates (right).

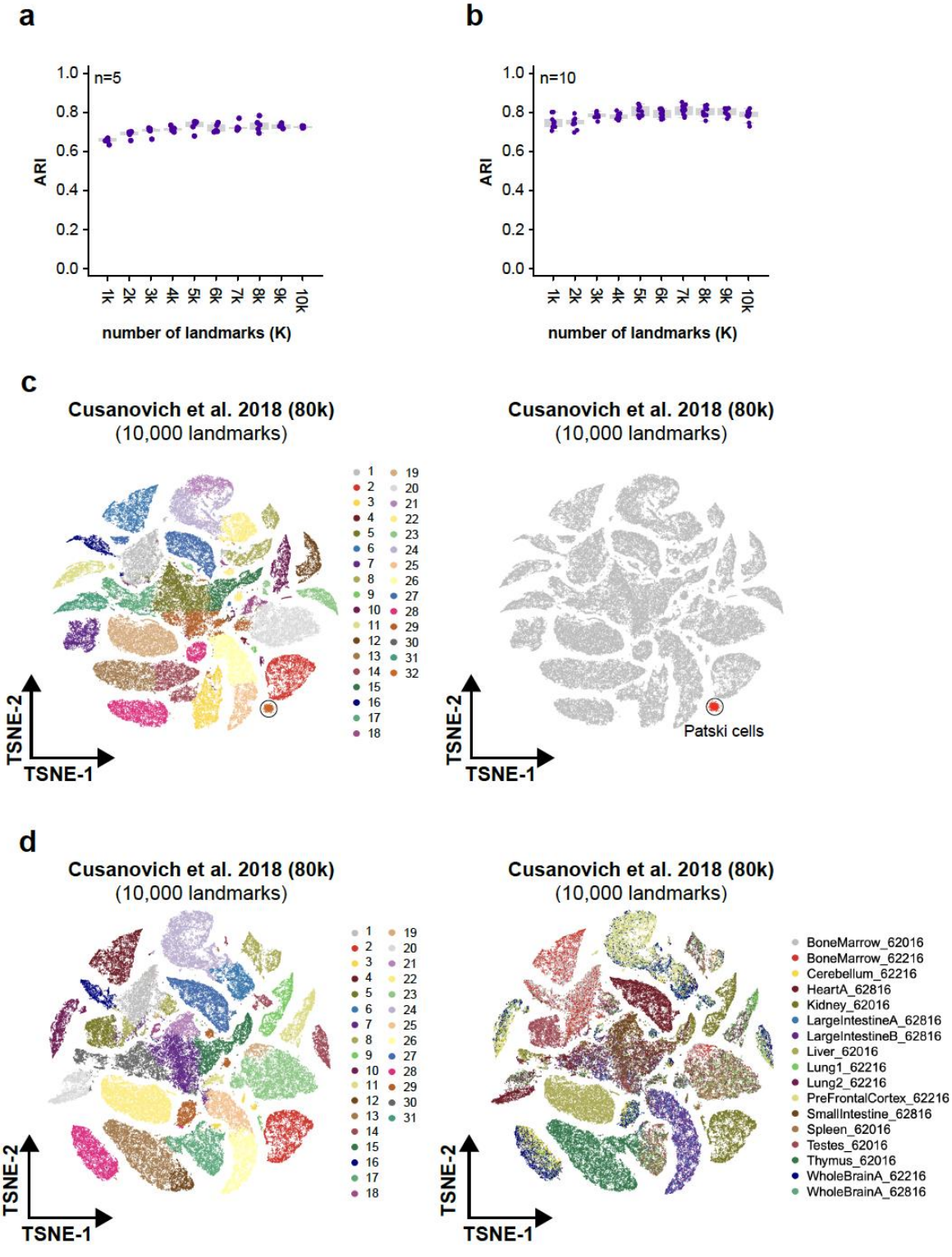
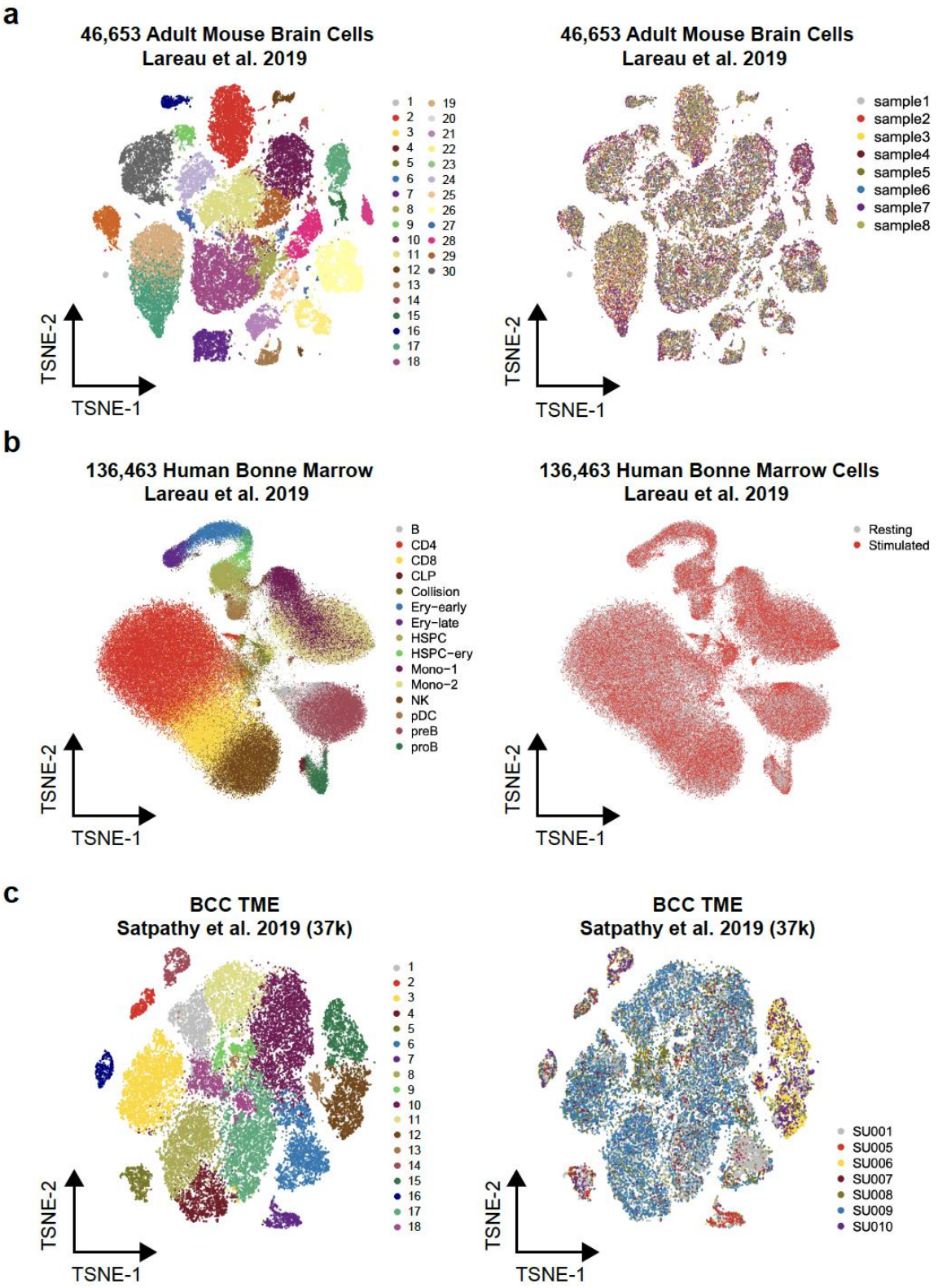

a
n=5
ARI
number of landmarks (K)
b
n=10
ARI
number of landmarks (K)
c
Cusanovich et al. 2018 (80k)
(10,000 landmarks)
TSNE-2
TSNE-1
Cusanovich et al. 2018 (80k)
(10,000 landmarks)
Patski cells
TSNE-2
TSNE-1
d
Cusanovich et al. 2018 (80k)
(10,000 landmarks)
TSNE-2
TSNE-1
Cusanovich et al. 2018 (80k)
(10,000 landmarks)
BoneMarrow_62016
BoneMarrow_62216
Cerebellum_62216
HeartA_62816
Kidney_62016
LargeIntestineA_62816
LargeIntestineB_62816
Liver_62016
Lung1_62216
Lung2_62216
PreFrontalCortex_62216
SmallIntestine_62816
Spleen_62016
Testes_62016
Thymus_62016
WholeBrainA_62216
WholeBrainA_62816
TSNE-2
TSNE-1

Figure S2.7. SnapATAC delineates cellular heterogeneity in published large-scale scATAC-seq datasets. Using 10,000 landmarks, SnapATAC is applied to three recently published large-scale scATAC-seq datasets and reveals substantial heterogeneity in the adult mouse brain[12] (**a**), human bone marrow[12] (**b**) and BCC TME[13] (**c**). Harmony is applied when analyzing human bone marrow[12] (**b**) and BCC TME[13] (**c**) because batch effect was observed and reported in these datasets in the original study.


a
46,653 Adult Mouse Brain Cells
Lareau et al. 2019
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
TSNE-2
TSNE-1
46,653 Adult Mouse Brain Cells
Lareau et al. 2019
sample1
sample2
sample3
sample4
sample5
sample6
sample7
sample8
TSNE-2
TSNE-1
b
136,463 Human Bonne Marrow
Lareau et al. 2019
B
CD4
CD8
CLP
Collision
Ery−early
Ery−late
HSPC
HSPC−ery
Mono−1
Mono−2
NK
pDC
preB
proB
TSNE-2
TSNE-1
136,463 Human Bonne Marrow Cells
Lareau et al. 2019
Resting
Stimulated
TSNE-2
TSNE-1
c
BCC TME
Satpathy et al. 2019 (37k)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
TSNE-2
TSNE-1
BCC TME
Satpathy et al. 2019 (37k)
SU001
SU005
SU006
SU007
SU008
SU009
SU010
TSNE-2
TSNE-1

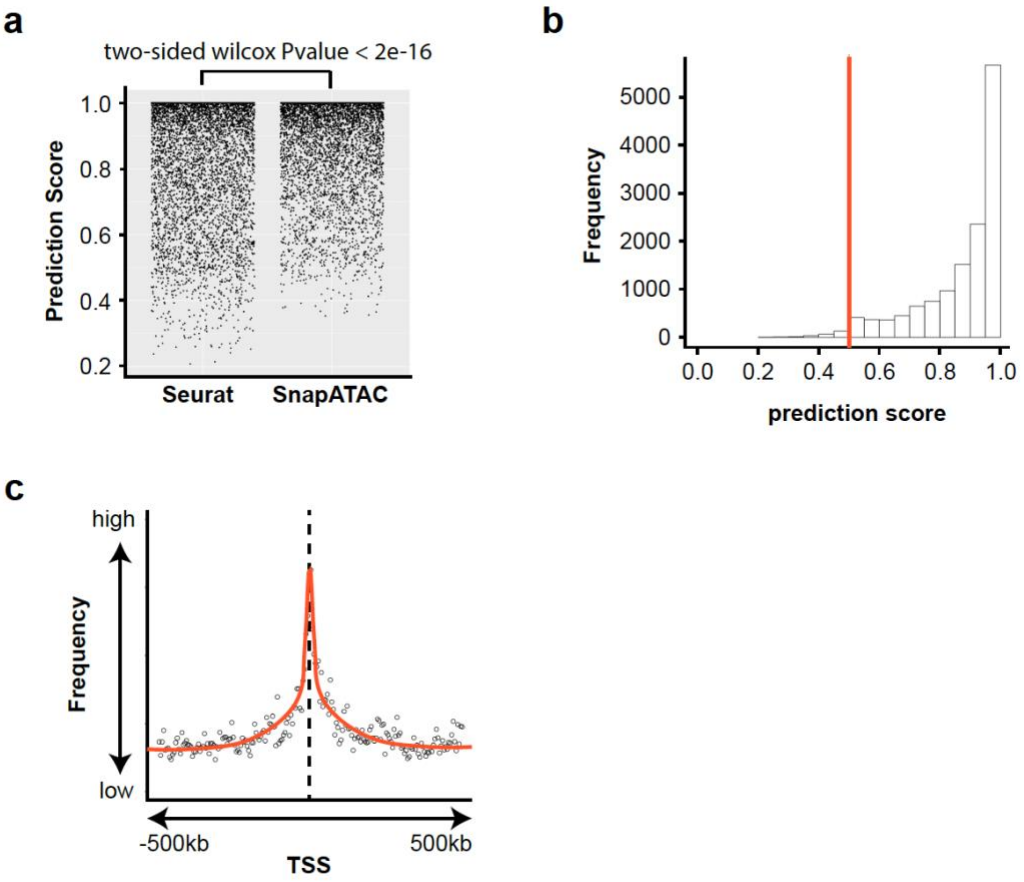


Figure S2.8. SnapATAC predicts gene and enhancer pairing by integrating scATAC-seq and scRNA-seq. (**a**) Prediction score distribution for single cell ATAC-seq (5K PBMC 10X) by Seurat (left) and SnapATAC (right). When predicting the cell type for scATAC-seq using corresponding scRNA-seq dataset (10K PBMC 10X), each cell in scATAC-seq was assigned with a prediction score indicating the confidence of the prediction. It ranges from 0 to 1, a higher score indicates a higher confidence. (**b**) Prediction score distribution for SnapATAC on 15K PBMC scATAC-seq. (**c**) Distance decay curve for the association (-logPvalue) between regulatory elements and the TSS of their putative target genes.

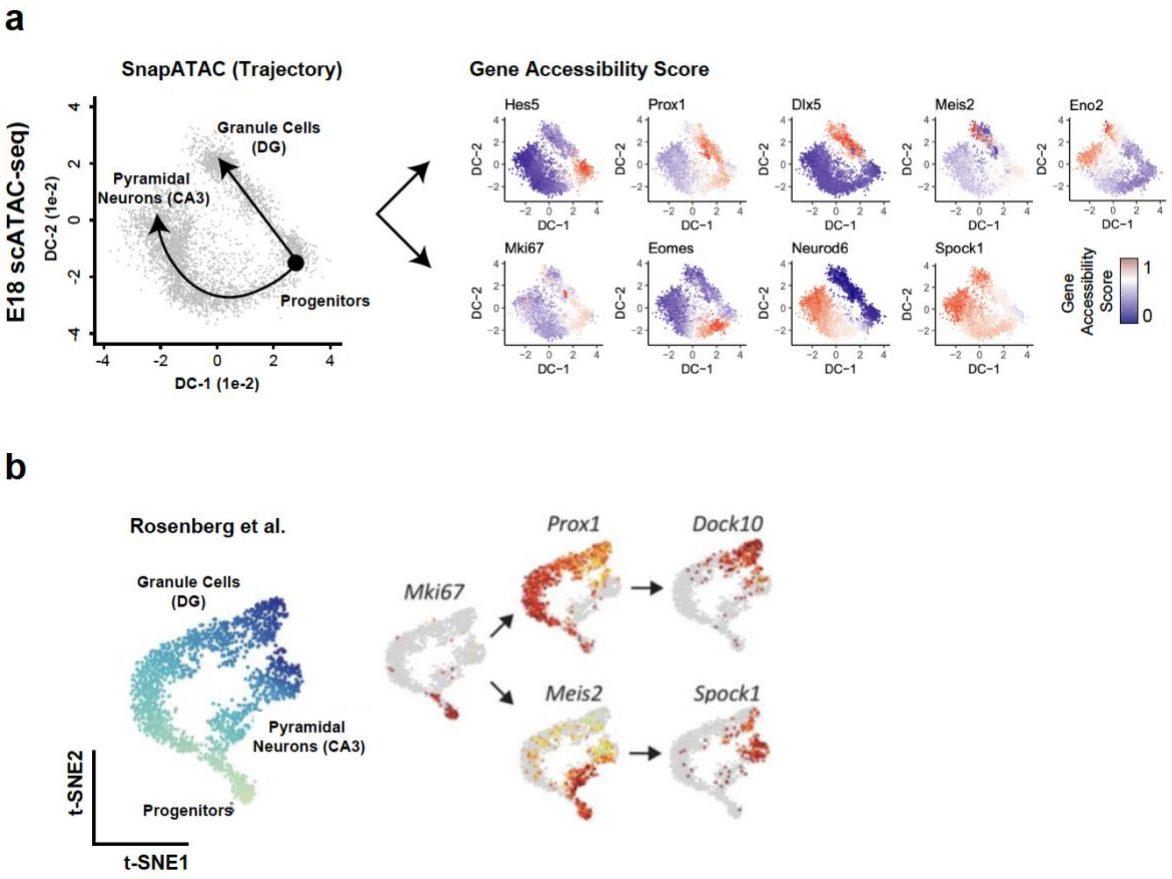
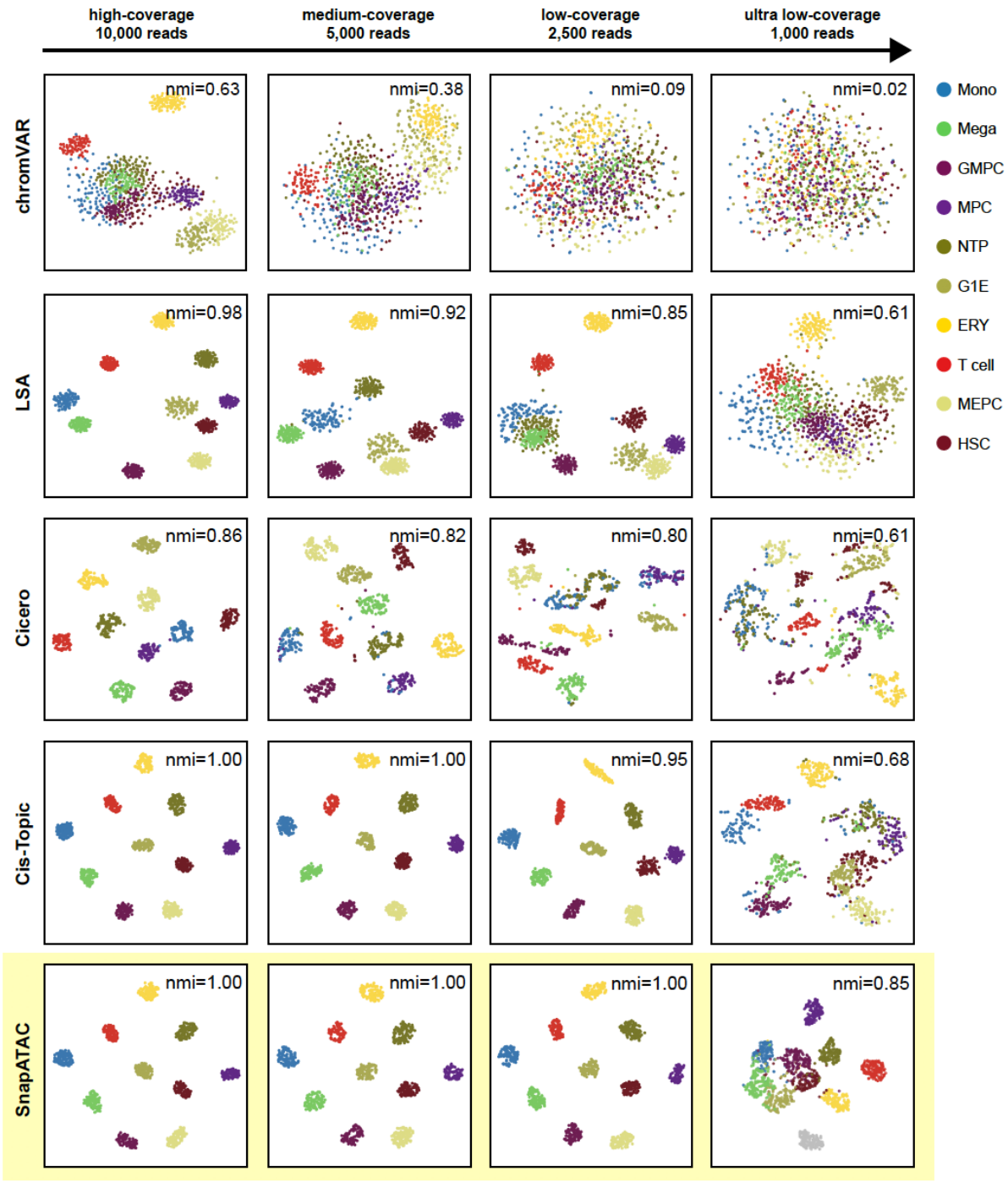


Figure S2.9. SnapATAC constructs cellular trajectories for the developing mouse brain. (**a**) Diffusion component representation of a dataset that contained 4,259 single cell ATAC-seq profiles from the hippocampus and ventricular zone in embryonic mouse brain (E18) revealed differentiation trajectories from progenitor cells to Granule Cells (DG) and Pyramidal Neurons (CA3) (left). Gene accessibility score at canonical differentiation marker genes were projected onto the diffusion components. The lineage was defined using Slingshot. (**b**) T-SNE representation of 1,944 single nucleus gene expression profiles from hippocampus reveals Dentate Gyrus cell lineage, highly similar with result obtained using scATAC-seq in (**a**). Figures were modified and adopted from Rosenberg 2018[29].

Figure S2.10. Evaluation of clustering accuracy of SnapATAC relative to alternative methods on simulated datasets. T-SNE visualization of clustering results on 1,000 simulated cells sampled from 10 bulk ATAC-seq datasets (see **Supplementary Methods** for the simulation) analyzed by five different methods – chromVAR[14], LSA[8], Cicero[17], Cis-Topic[15] and SnapATAC. Clustering results are compared to the original cell type label and the accuracy is estimated using Normalized Mutual Index (nmi). 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.


high-coverage
10,000 reads
medium-coverage
5,000 reads
low-coverage
2,500 reads
ultra low-coverage
1,000 reads
chromVAR
LSA
Cicero
Cis-Topic
SnapATAC
nmi=0.63
nmi=0.38
nmi=0.09
nmi=0.02
nmi=0.98
nmi=0.92
nmi=0.85
nmi=0.61
nmi=0.86
nmi=0.82
nmi=0.80
nmi=0.61
nmi=1.00
nmi=1.00
nmi=0.95
nmi=0.68
nmi=1.00
nmi=1.00
nmi=1.00
nmi=0.85
Mono
Mega
GMPC
MPC
NTP
G1E
ERY
T cell
MEPC
HSC

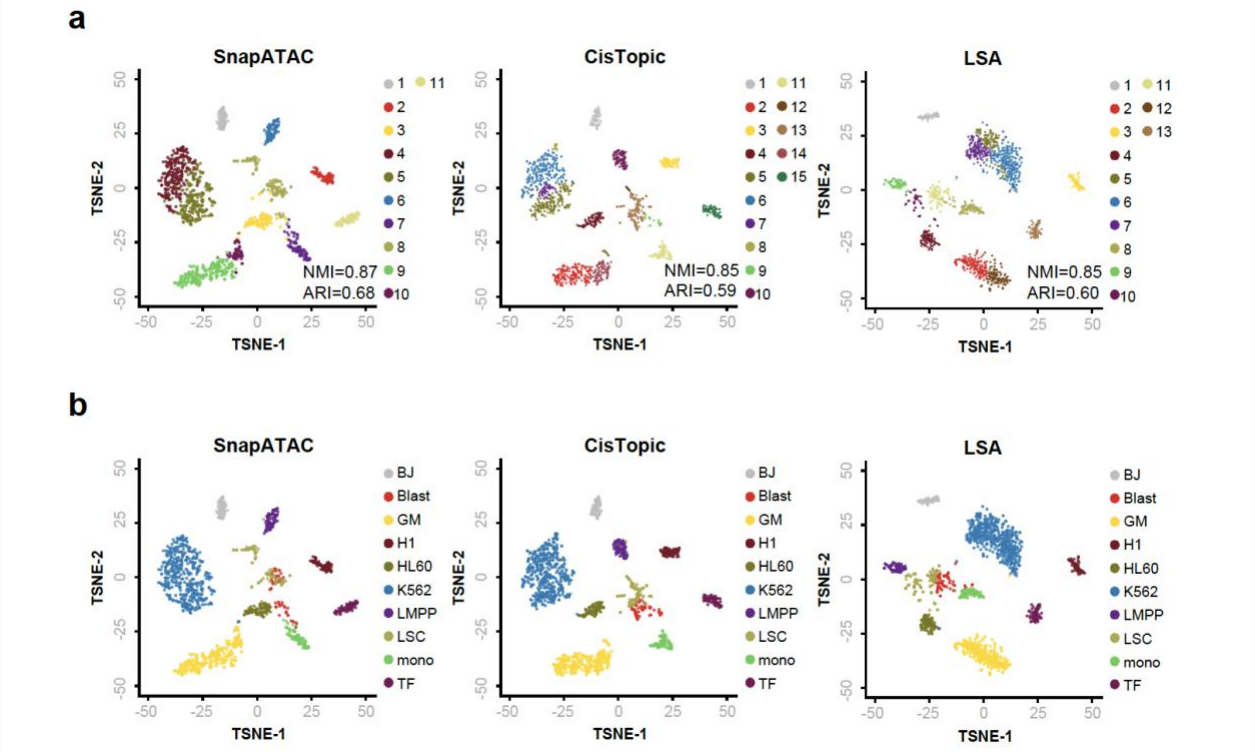
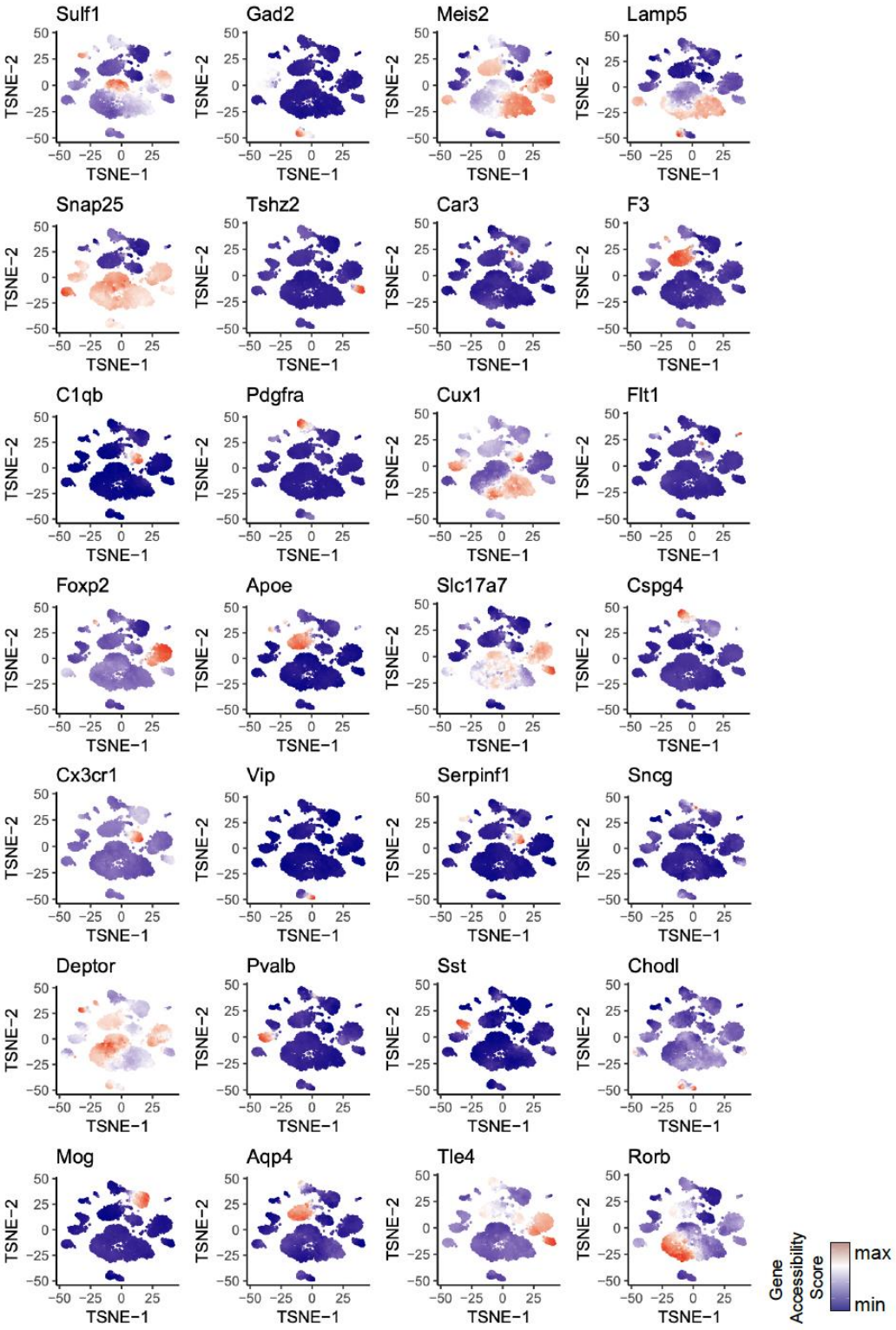


Figure S2.11. Evaluation of clustering accuracy on published single cell ATAC-seq datasets. SnapATAC (left), CisTopic (middle) and LSA (right) clustering performance on single cell ATAC-seq dataset from ten human cell lines generated using Fluidigm C1 platform[14]. (**a**) Clustering results are visualized using t-SNE and cells are colored by cluster labels identified by each of analysis methods. (**b**) T-SNE visualization of the human cells colored by the cell type labels. Clustering accuracy of each method is estimated by comparing the predicted clustering labels to the cell type labels. Blast: acute myeloid leukemia blast cells; LSC: acute myeloid leukemia leukemic stem cells; LMPP: lymphoid-primed multipotent progenitors; Mono: monocyte; HL60: HL-60 promyeloblast cell line; TF1: TF-1 erythroblast cell line; GM: GM12878 lymphoblastoid cell line; BJ: human fibroblast cell line; H1: H1 human embryonic stem cell line.

Figure S2.12. Gene accessibility score of canonical marker genes projected onto t-SNE embedding of mouse secondary motor cortex (MOs-M1) snATAC-seq dataset to guide the cluster annotation. T-SNE is generated using SnapATAC; cell type specific marker genes were defined from previous single cell transcriptomic analysis in the adult mouse brain[34]; gene accessibility score is calculated using SnapATAC (**Supplementary Methods**).


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

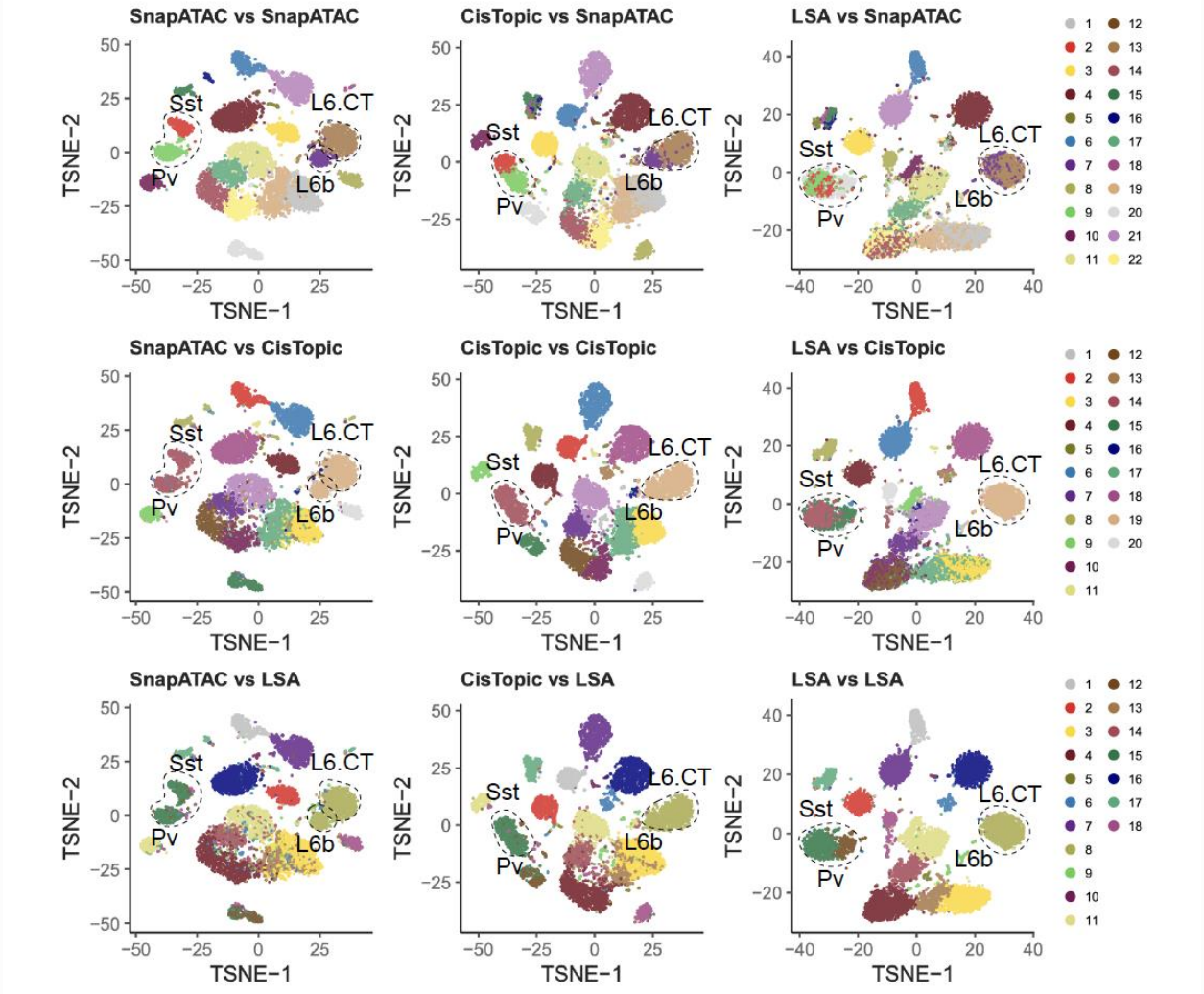
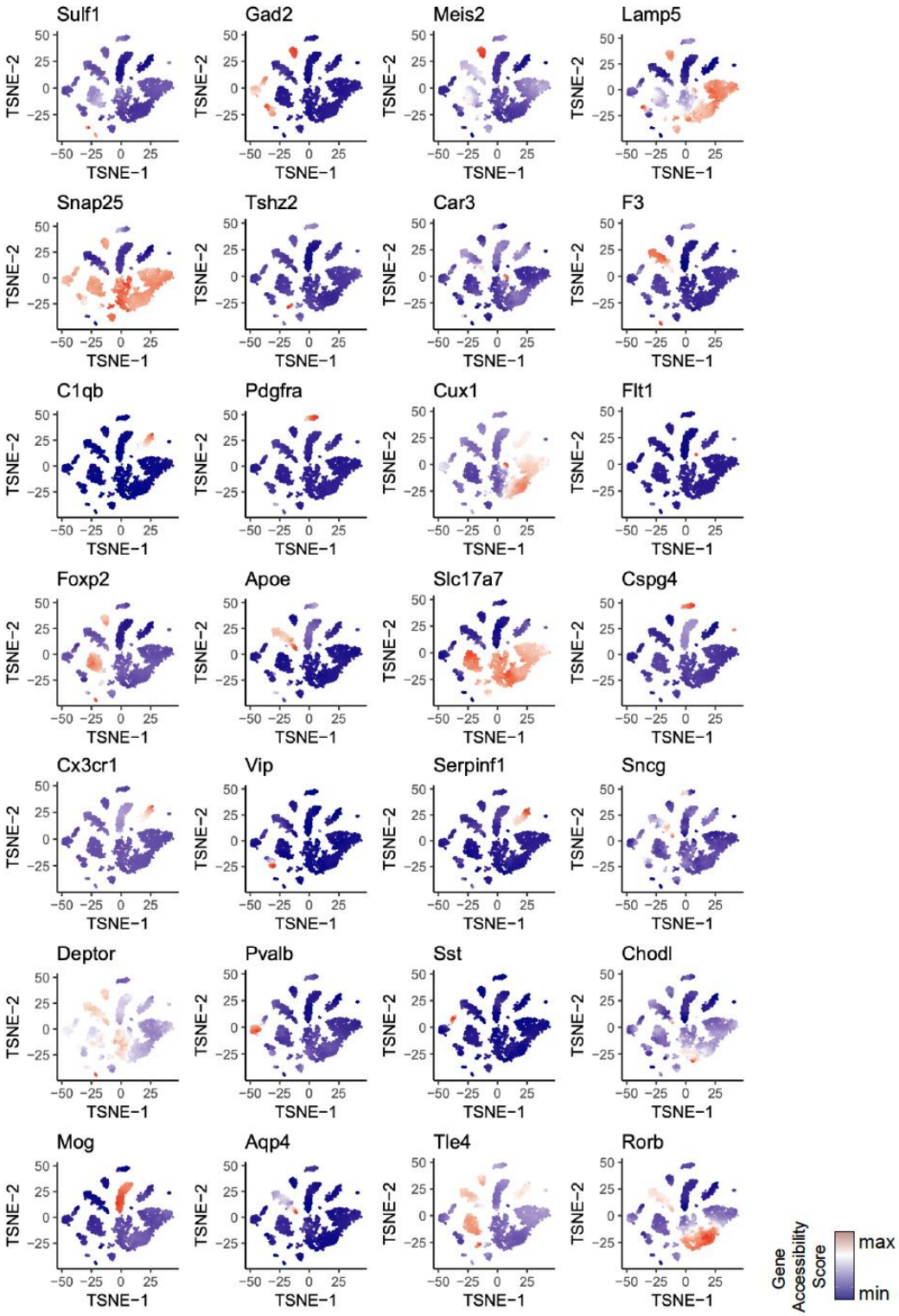


Figure S2.13. Evaluation of clustering sensitivity on in-house mouse secondary motor cortex dataset. Three methods (cisTopic, LSA and SnapATAC) were used to analyze a dataset that contained ~10k single nucleus ATAC-seq profiles from the mouse secondary motor cortex. Pairwise comparison of the clustering results is shown by projecting the cluster label identified using one method onto the t-SNE visualization generated by another method (cluster vs. visualization). Black dash line circles highlight the rare pollutions (Sst, Pv, L6b and L6.CT) that were only identified by SnapATAC.

Figure S2.14. Gene accessibility score of canonical marker genes projected onto t-SNE embedding for a 10X scATAC-seq dataset of the mouse brain to guide the cluster annotation. T-SNE is generated using SnapATAC; cell type specific marker genes is defined from previous single cell transcriptomic analysis[34]; gene accessibility score is calculated using SnapATAC (**Supplementary Methods**).


Sulf1
Gad2
Meis2
Lamp5
Snap25
Tshz2
Car3
F3
C1qb
Pdgfra
Cux1
Flt1
Foxp2
Apoe
Slc17a7
Cspg4
Cx3cr1
Vip
Serpinf1
Sncg
Deptor
Pvalb
Sst
Chodl
Mog
Aqp4
Tle4
Rorb
TSNE-1
TSNE-2
50
25
0
−25
−50
Gene Accessibility Score
max
min

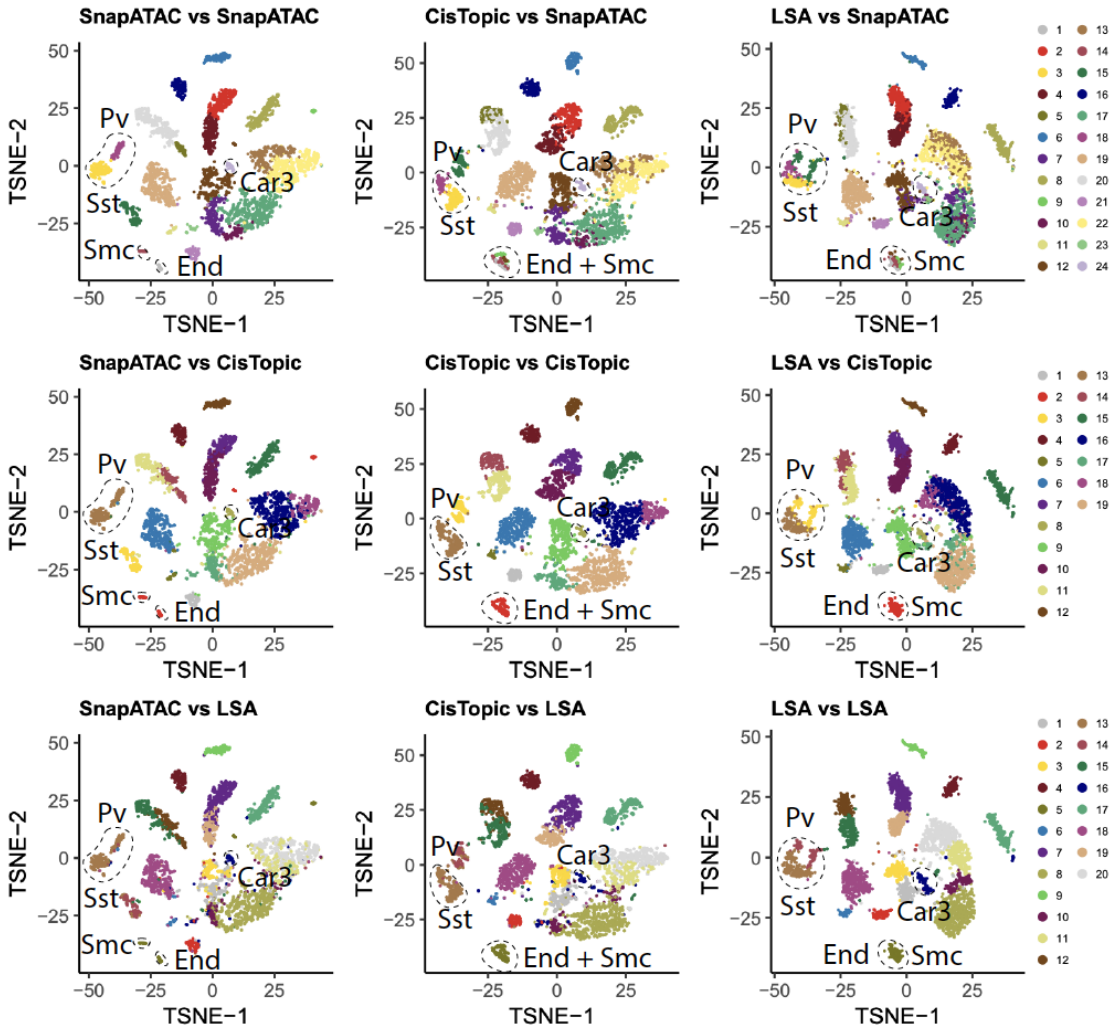


Figure S2.15. Evaluation of clustering sensitivity on a 10X scATAC-seq dataset from the Mouse Brain. Three methods (cisTopic, LSA and SnapATAC) were used to analyze a dataset that contained ~5k single cell ATAC-seq profiles from the adult mouse brain. Pairwise comparison of the clustering results is shown by projecting the cluster label identified using one method onto the t-SNE visualization generated by another method (cluster vs. visualization). Black dash line circles highlight the rare pollutions (Sst, Pv, Car3, End and Smc) that were only identified by SnapATAC.

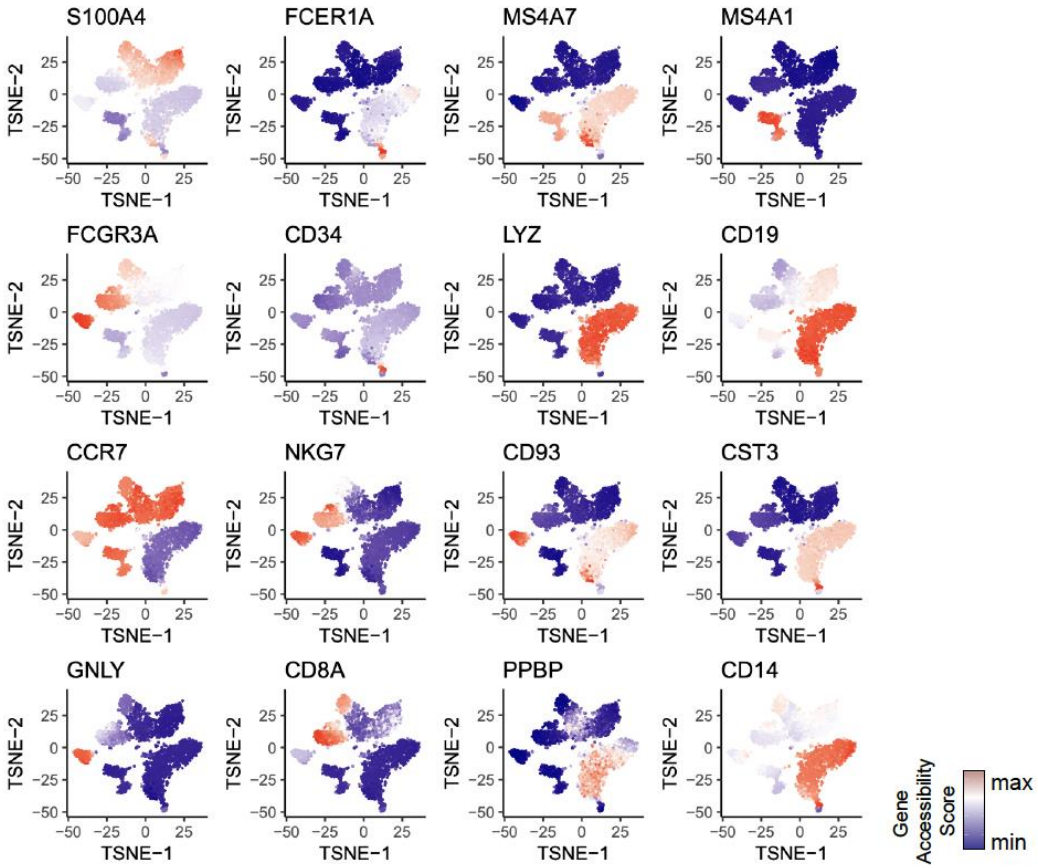


Figure S2.16. Gene accessibility score of canonical marker genes projected onto the t-SNE embedding from 5K PBMC 10X dataset to guide the annotation of the clusters. T-SNE is generated using SnapATAC; cell type specific marker genes are defined from previous single cell transcriptomic analysis; gene accessibility score is calculated using SnapATAC (**Supplementary Methods**).

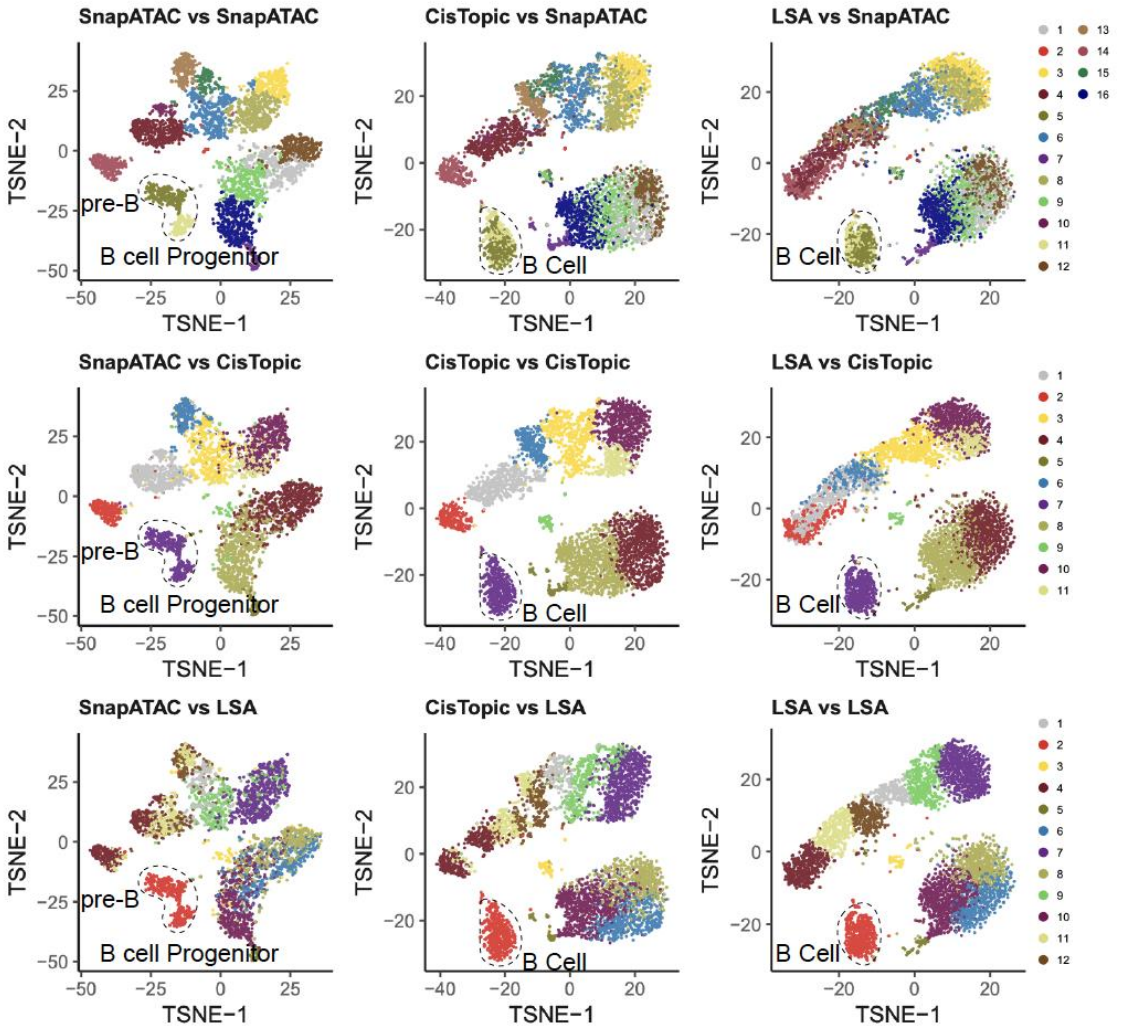


Figure S2.17. Evaluation of clustering sensitivity on a 5K PBMC 10X dataset. Three methods (cisTopic, LSA and SnapATAC) were used to analyze a dataset that contains ~5k single cell ATAC-seq profiles from PBMC. Pairwise comparison of the clustering results is shown by projecting the cluster label identified using one method onto the t-SNE embedding generated by another method (cluster vs. visualization). Dash-line circles highlight the rare pollutions (Pre-B and B cell progenitor) that are only distinguished by SnapATAC.

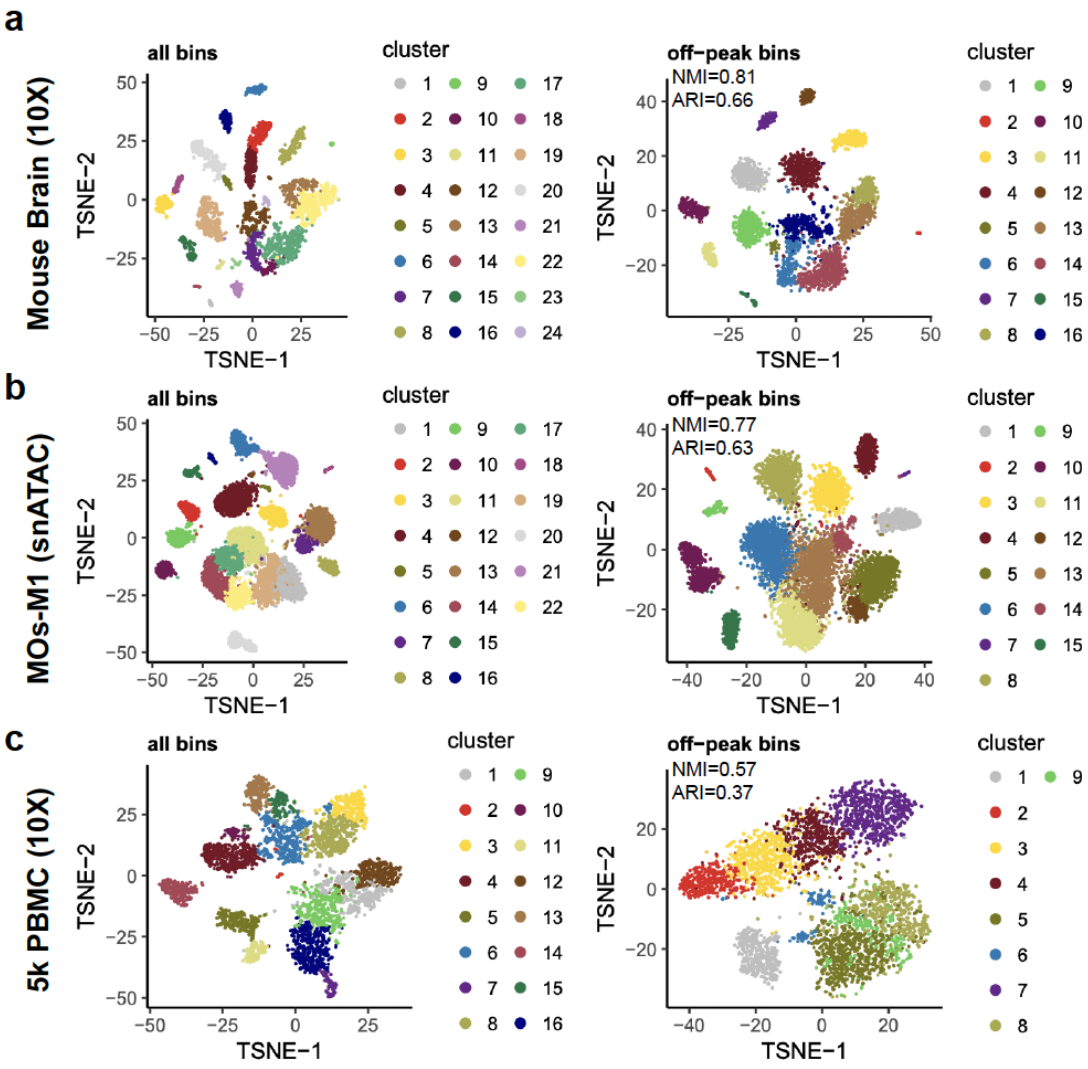


Figure S2.18. Off-peak reads can be used to distinguish different cell types. (**a-c**) SnapATAC clustering result on three benchmarking datasets using all bins versus clustering result only using bins that are not overlapped with peaks.

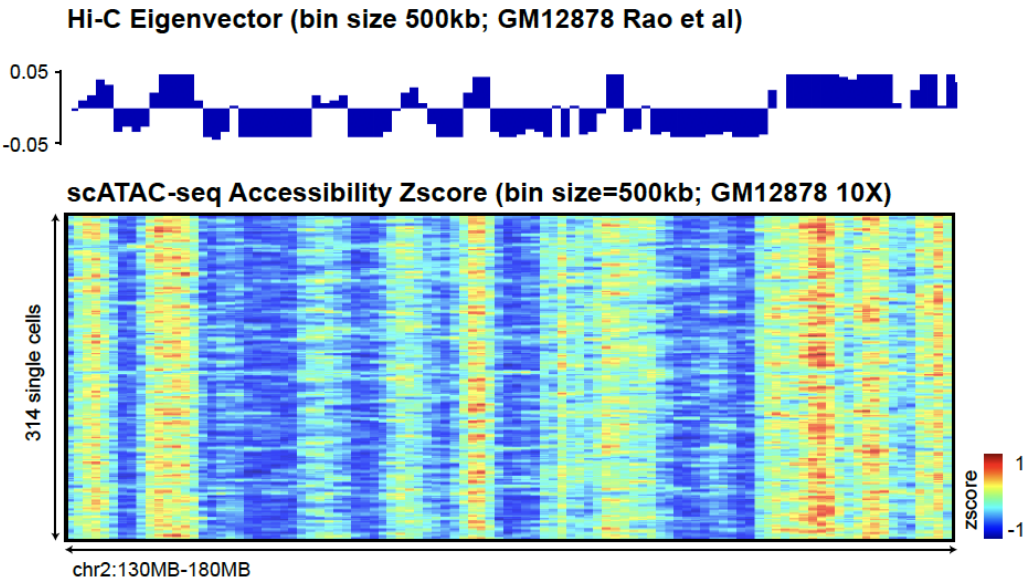


Figure S2.19. Off-peak reads reflect higher-order chromatin structure. At 500kb bin resolution, profile of compartments identified using Hi-C[32] in GM12878 overlaid the density of "off-peak" reads for 314 cells from GM12878 10X scATAC-seq library.

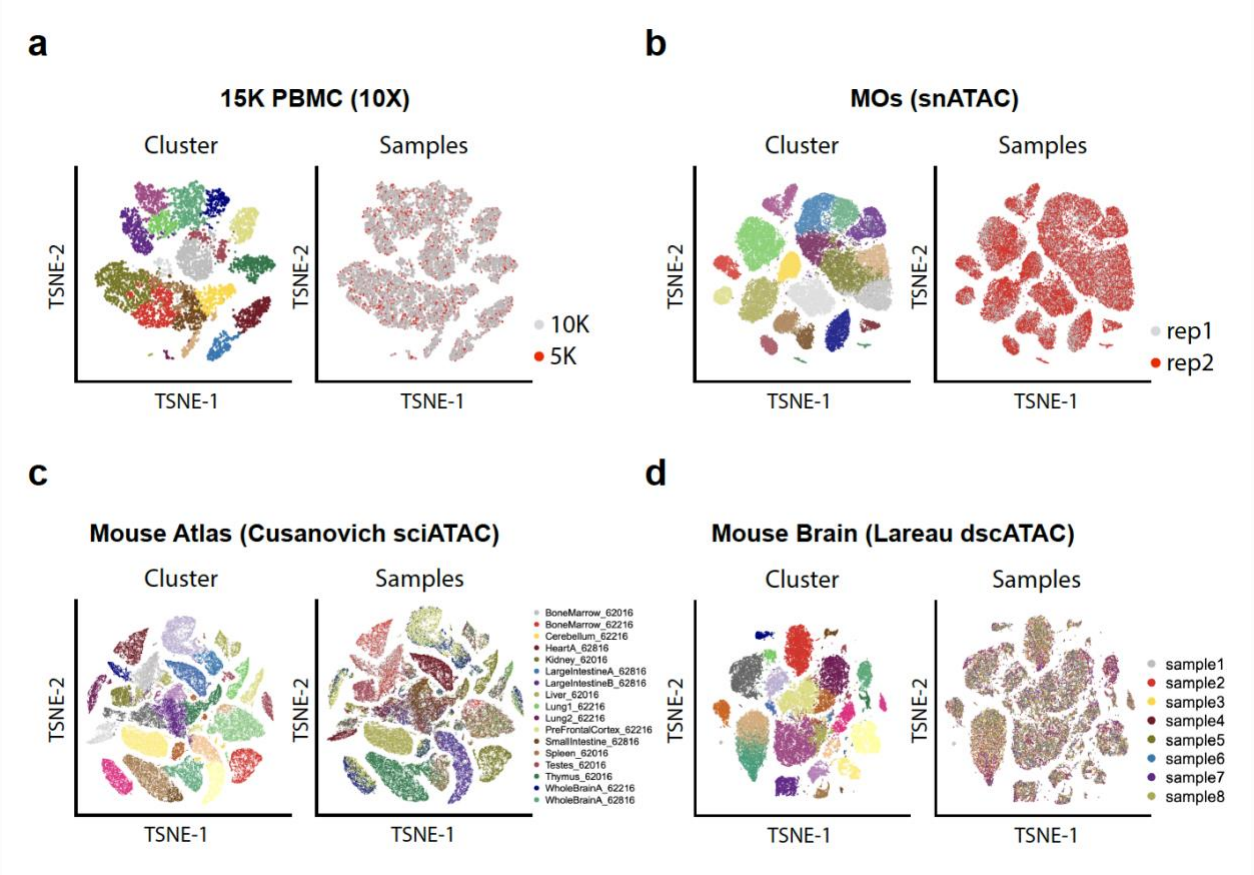


Figure S2.20. SnapATAC is robust to technical variation. Two-dimensional t-SNE visualization of four benchmarking datasets generated using SnapATAC. Cells are color by cluster label (left) and sample label (right). (**a**) 15k PBMC (10X) – a combination of two datasets (PBMC 5k and 10k) publicly available from 10X genomics. (**b**) MOs (snATAC) – an in-house dataset that contains two biological replicates from secondary motor cortex in the adult mouse brain generated using single nucleus ATAC-seq. (**c**) Mouse Atlas (Cusanovich 2018) – a published dataset that contains over 80K cells from 13 different mouse tissues generated using multiplexing single cell ATAC-seq. (**d**) Mouse Brain (Lareau dscATAC) – a published dataset that contains 46,652 cells from 8 samples in the adult mouse brain generated using BioRad droplet-based single cell ATAC-seq.

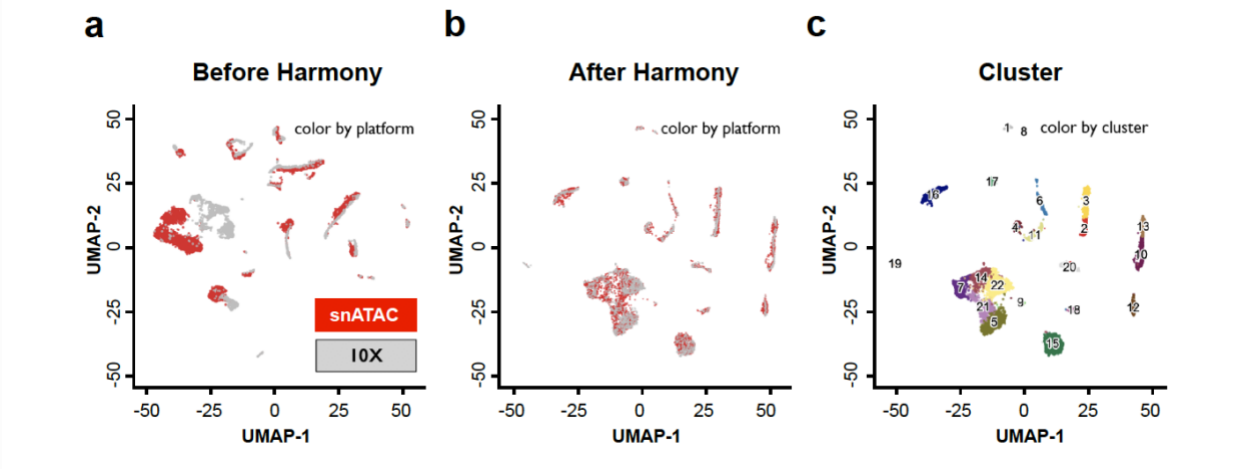
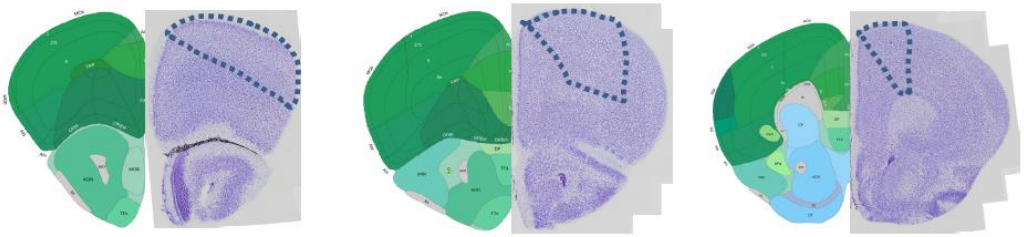


Figure S2.21. SnapATAC eliminates batch effect using Harmony. The joint UMAP visualization of two datasets of mouse brain generated using combinatorial indexing single nucleus ATAC-seq (MOs-M1 snATAC) and droplet-based platform (Mouse Brain 10X) before (**a**) and after (**b**) performing batch effect correction using Harmony.

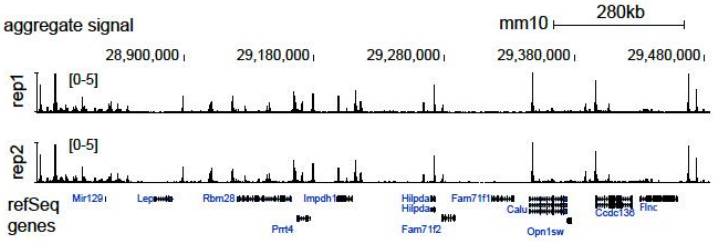
Figure S2.22. Single nucleus ATAC-seq datasets are reproducible between biological replicates. (**a**) Illustration of dissection. Posterior view of three 0.6 mm coronal slices from which the secondary motor cortex (MOs) was dissected. The right side on each image depicts the corresponding view from the Allen Brain Atlas. The left side correspond to the Nissl staining of the posterior side of each slice. The MOs region was manually dissected according to the dashed lines on each slice and following the MOs as depicted in plates 27, 33, and 39 of the Allen Brain Atlas (left side images in figure). Each slice contains two biological replicates named as A1, A2, M1, M2, P1 and P2 (A: Anterior; M: Middle; P: Posterior). In this study, A1, M1 and P1 is combined as replicate 1 and A2, M2 and P2 are combined as replicate 2. (**b**) Genome-browser view of aggregate signal for two biological replicates. (**c**) Pearson correlation of count per million (CPM) at peaks between two replicates. (**d**) Insert size distribution and (**e**) TSS enrichment score for two biological replicates.

a

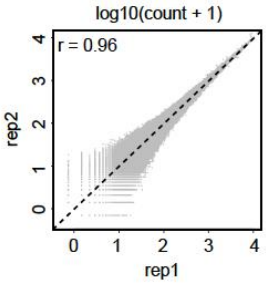


slice 1 (**A**nterior) | slice 2 (**M**iddle) | slice 3 (**P**osterior)

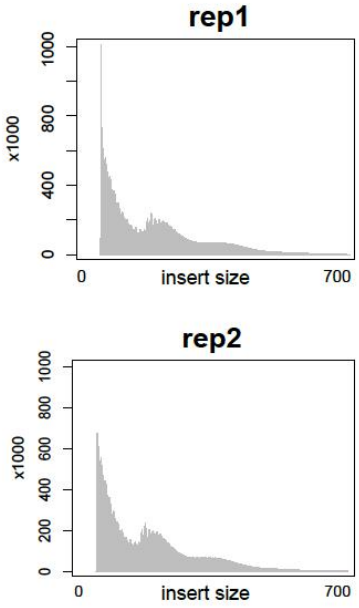
b



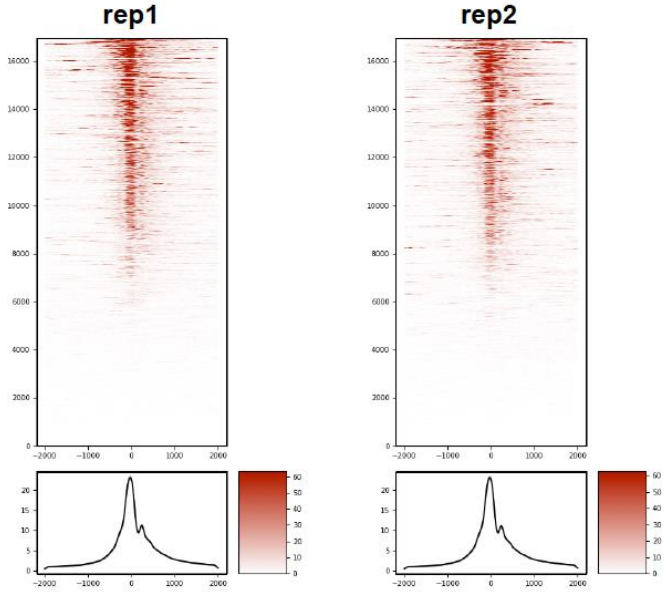
c



d



e



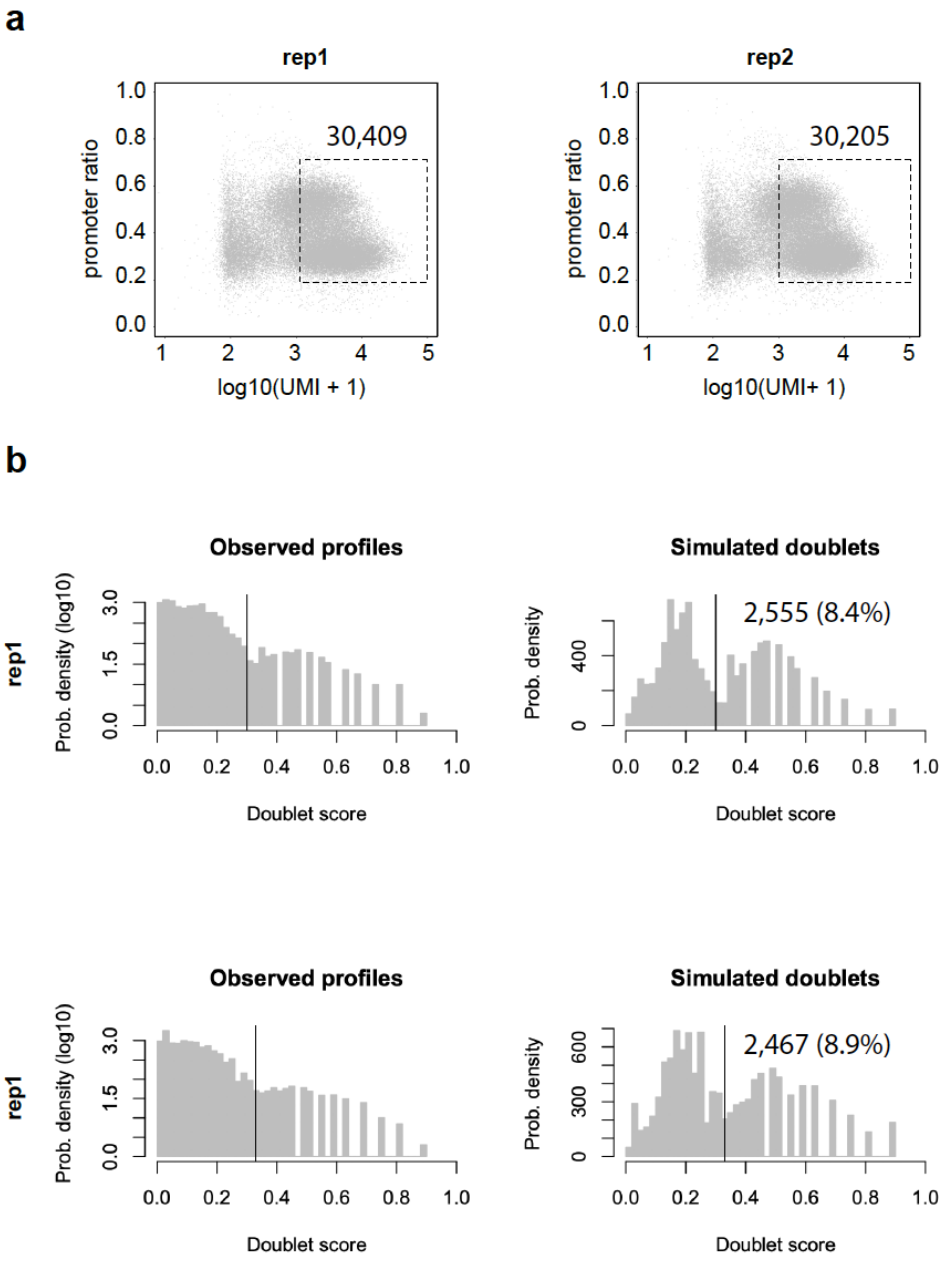


Figure S2.23. Barcode selection of MOs. (**a**) Cells of unique fragments within the range of 1,000-100,000 and fragments in promoter ratio within the range of 0.2-0.7 were selected. This resulted in 30,409 and 30,205 nuclei for two replicates. (**b**) Putative doublets were identified using Scrublets, which predicted 2,555 (8.4%) and 2,467 (8.9%) nuclei to be doublets for each replicate. The predicted doublet ratio is similar to the theoretical calculation of doublet ratio for multiplexing single cell ATAC-seq experiment[5,7].

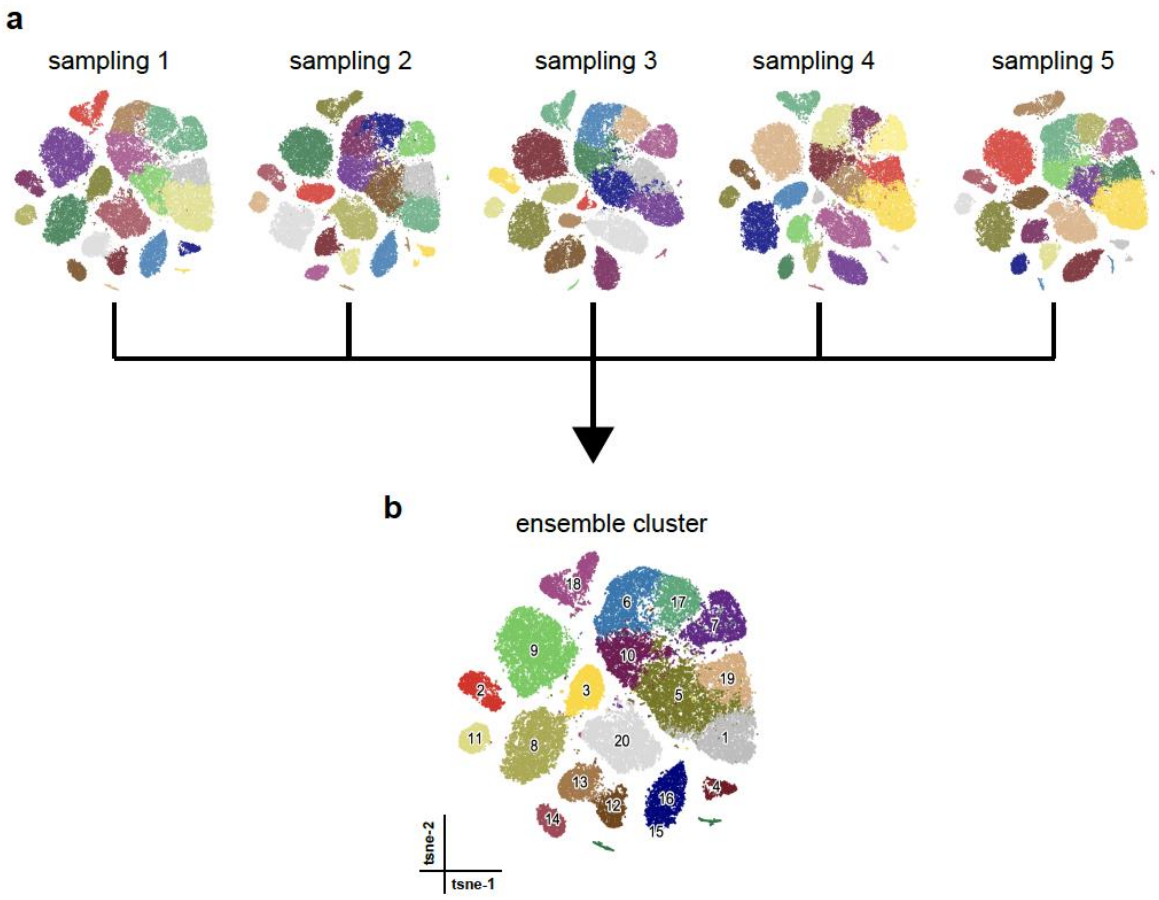


Figure S2.24. Consensus clustering of MOs. (**a**) Five clustering results were generated using SnapATAC with different set of landmarks (10,000). (**b**) These five clustering solutions were combined to create a consensus clustering which identified 20 clusters in MOs (**Supplementary Methods**).

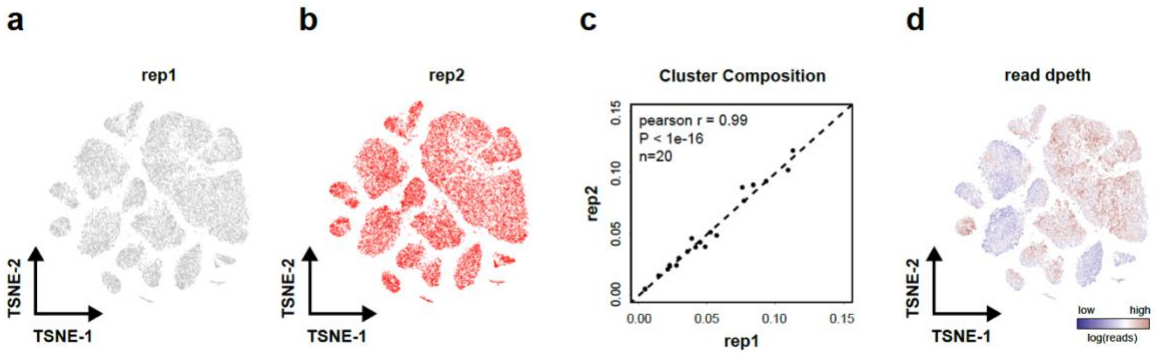


Figure S2.25. MOs clustering result is reproducible between biological replicates. (**a-b**) T-SNE visualization of cells from two biological replicates. (**c**) The cluster composition is highly reproducible between two biological replicates (r=0.99; P-value < 1e-22); (**d**) T-SNE visualization of cells with color scaled by sequencing depth.

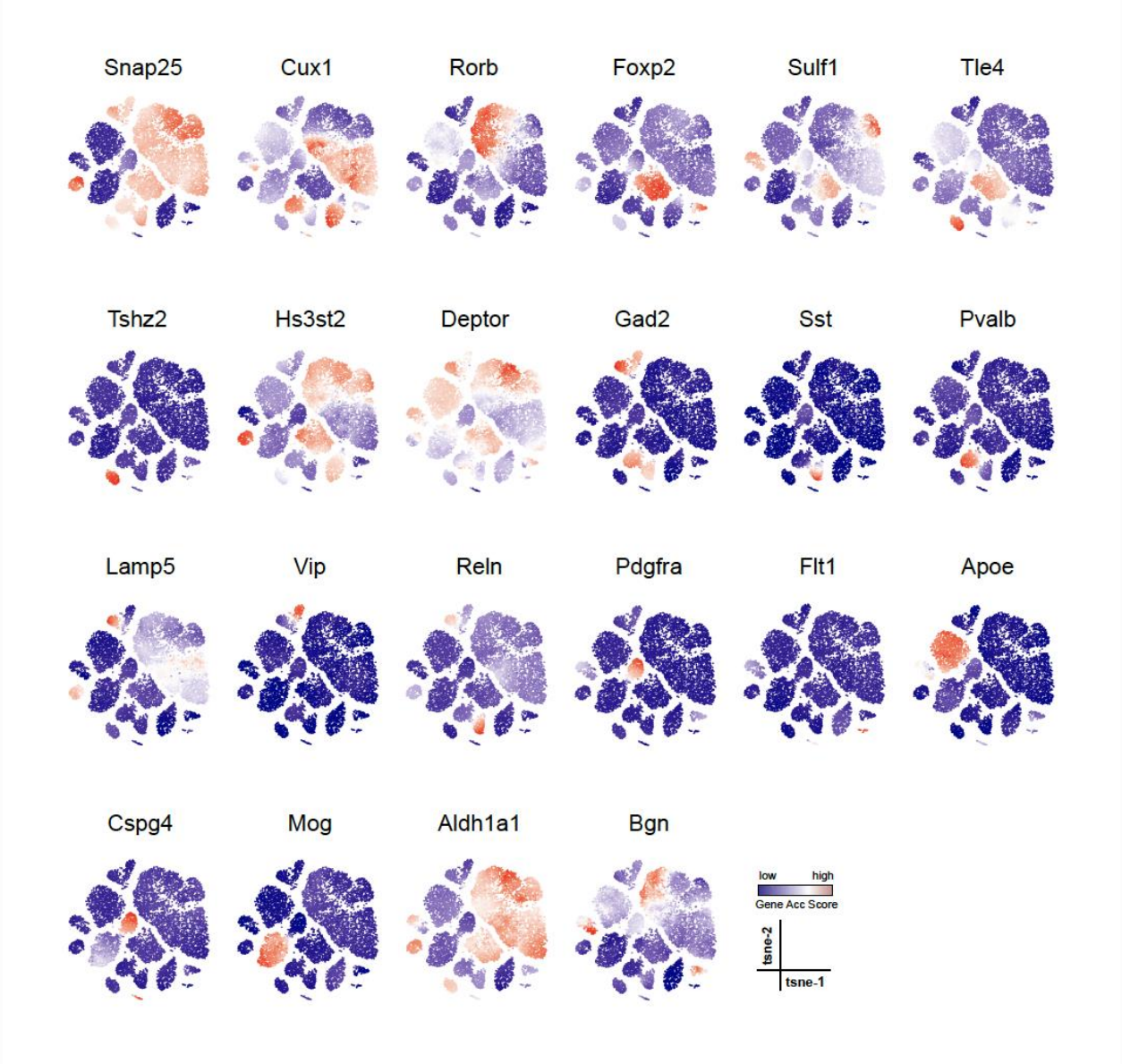


Figure S2.26. Gene accessibility score of canonical marker genes projected onto MOs t-SNE embedding to guide the cluster annotation. T-SNE is generated using SnapATAC for MOs; cell type specific marker genes was defined from previous single cell transcriptomic analysis in adult mouse brain[34]; gene accessibility score is calculated using SnapATAC (**Supplementary Methods**) and projected to the t-SNE embedding.

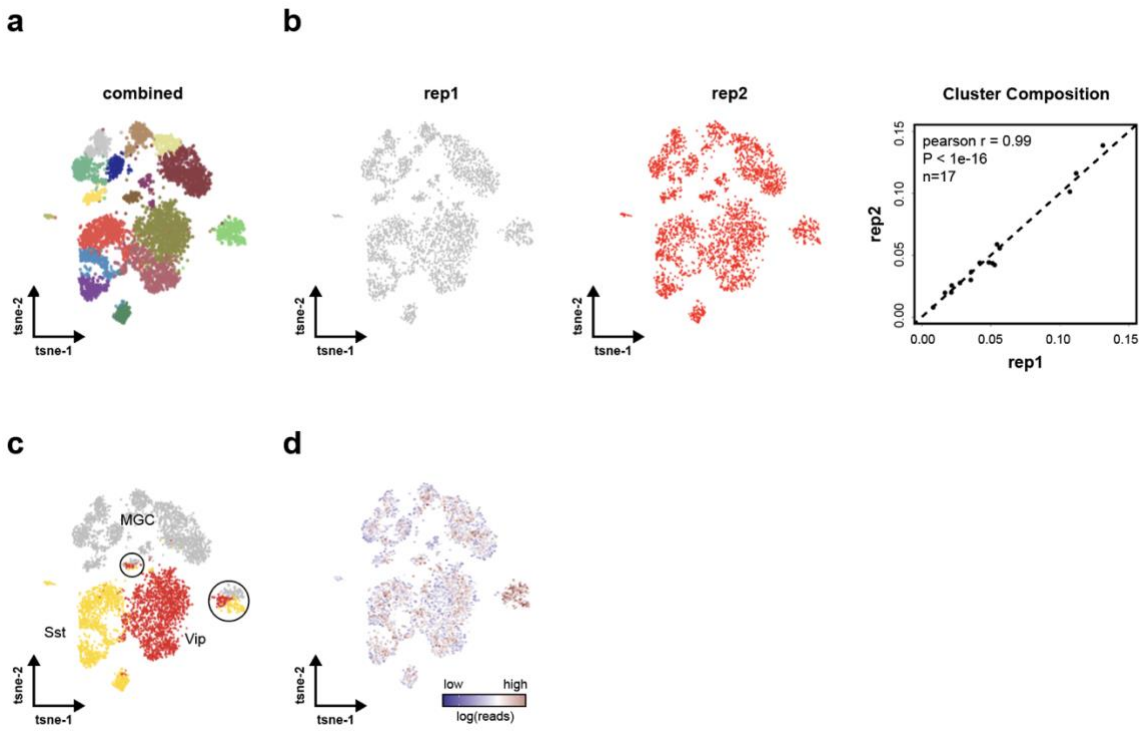
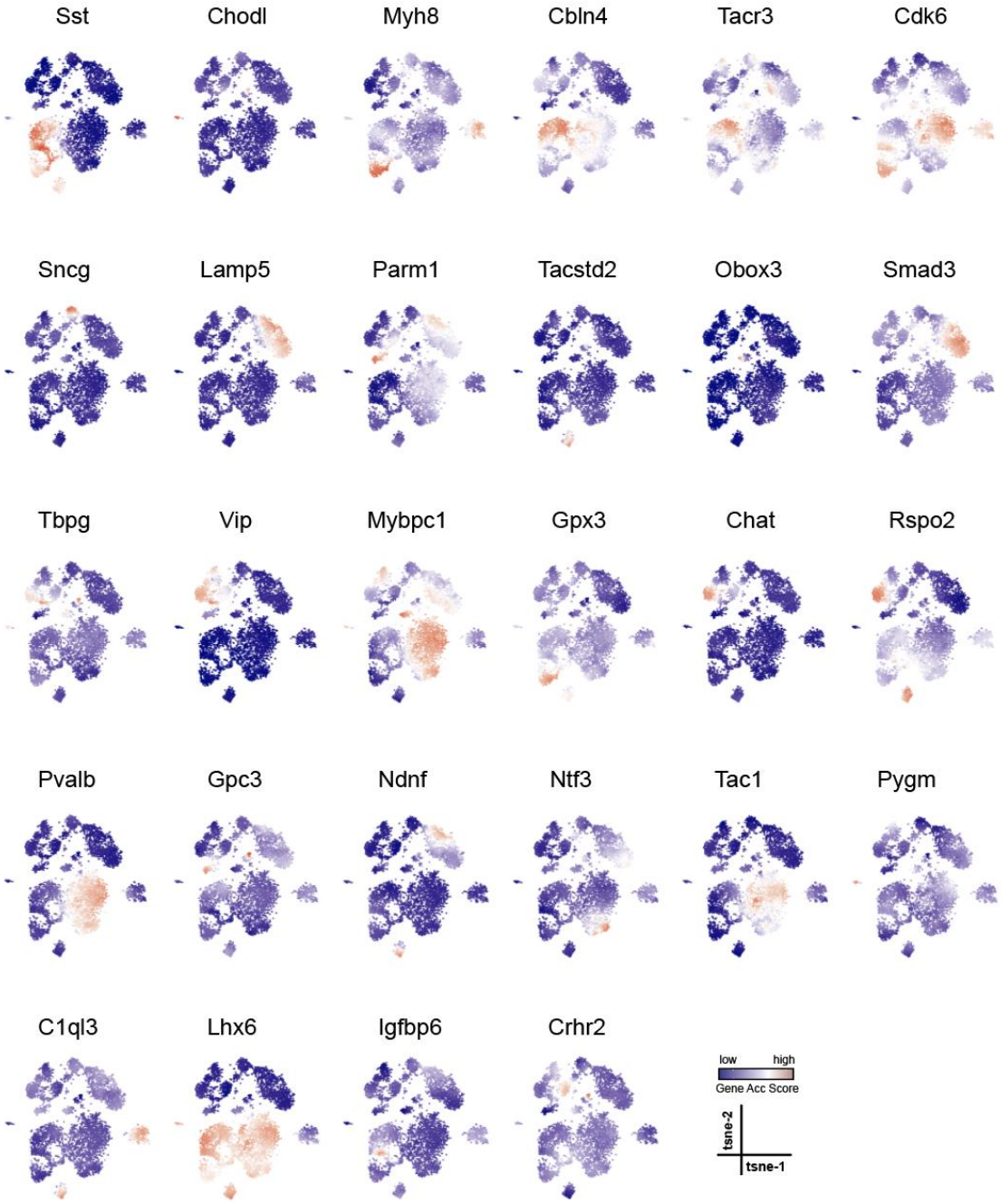


Figure S2.27. Iterative clustering identifies 17 GABAergic neuronal subtypes. (**a**) Sub-clustering of 5,940 GABAergic neurons identified 17 distinct cell clusters. (**b**) Cluster composition was highly reproducible between two biological replicates. (**c**) TSNE visualization of 5,940 GABAergic neurons colored by cell types identified in the initial clustering (shown in **Figure 2.5a**). Black circles mark clusters that are potential doublets, a mixture of multiple cell types. (**d**) TSNE plot of GABAergic neurons colored by sequencing depth.

Figure S2.28. Gene accessibility score of marker genes projected onto t-SNE embedding from GABAergic neurons to guide the cluster annotation. Iterative clustering is performed against GABAergic neurons to identify subtypes. Twenty eight cell type specific marker genes were defined from previous single cell transcriptomic analysis in adult mouse brain[34]; gene accessibility score is calculated using SnapATAC (**Supplementary Methods**).


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

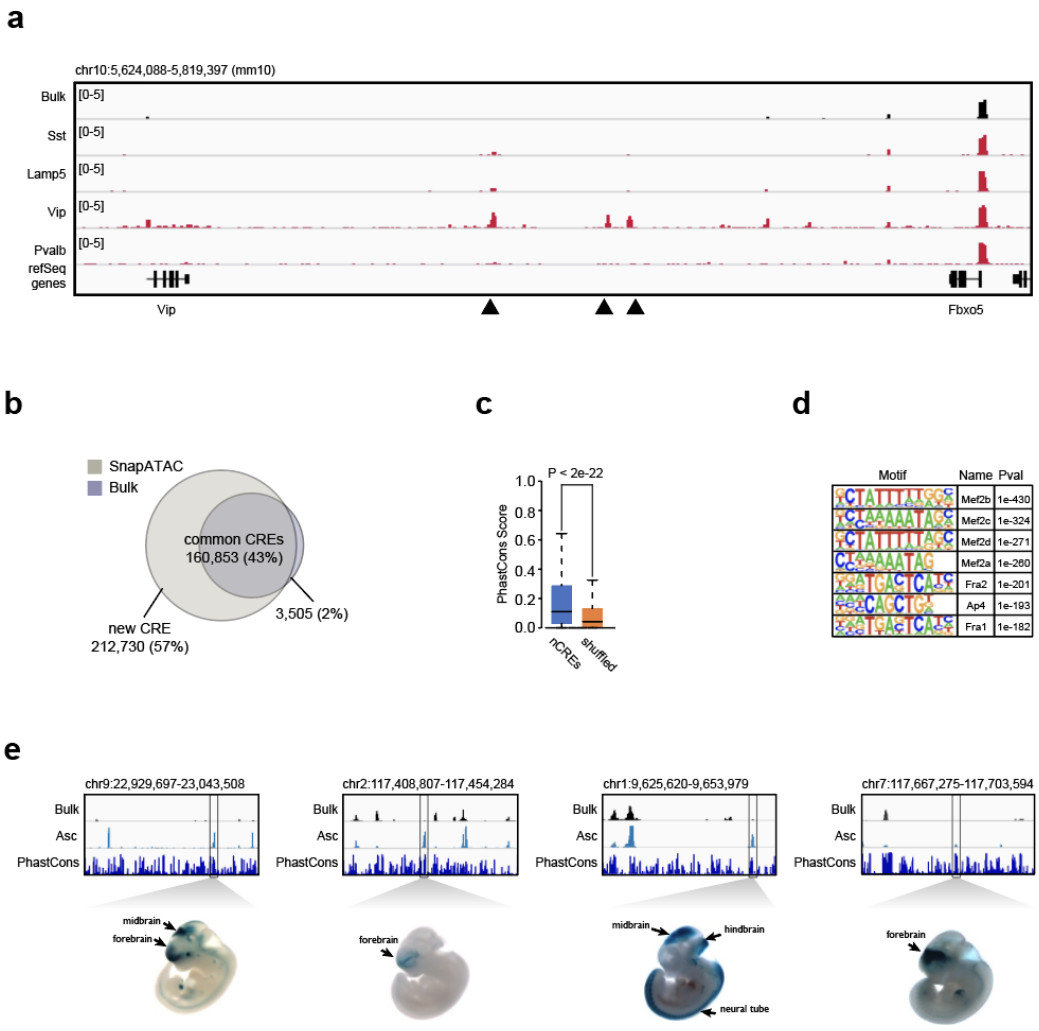


Figure S2.29. SnapATAC uncovers novel candidate cis-regulatory elements in rare cell types. (**a**) Genome browser view of 20Mb region flanking gene *Vip*. Dash line highlight five regulatory elements specific to Vip subtypes that are under-represented in the conventional bulk ATAC-seq signal. (**b**) Over fifty percent of the regulatory elements identified from 20 major cell populations are not detected from bulk ATAC-seq data. (**c**) Sequence conservation comparison between the new elements and randomly chosen genomic regions. (**d**) Top seven motifs enriched in Pv-specific new elements. (**f**) Examples of four new elements that were previously tested positive in transgenic mouse assays and reported in the VISTA database.

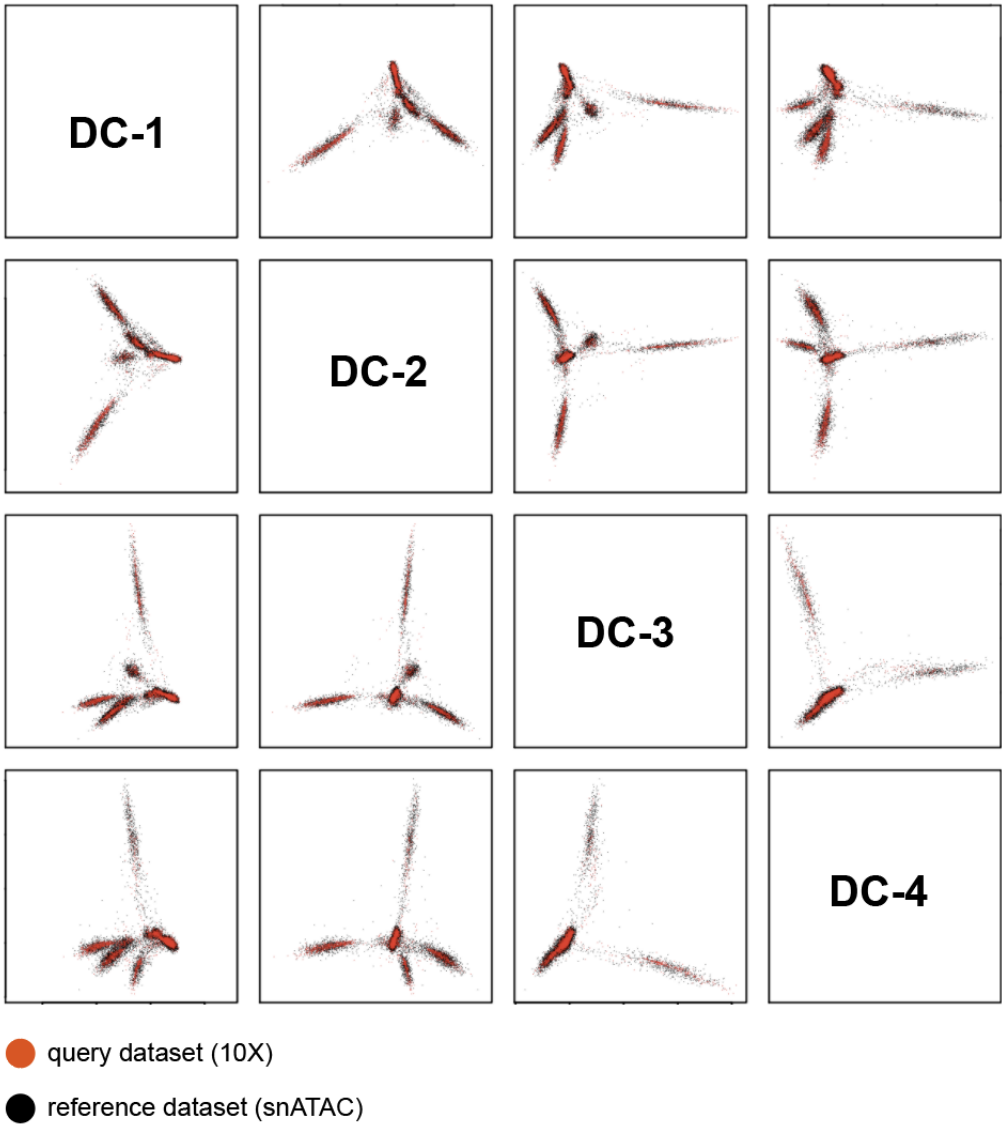


Figure S2.30. Joint diffusion maps embedding for query (Mouse Brain 10X) and reference dataset (MOs snATAC). The query dataset (10X) is projected onto the diffusion component (DC) space precomputed for the reference dataset (snATAC). Batch effect is corrected using Harmony. Pairwise plot of the first four diffusion components (DCs) in which cells are colored by dataset - red for query cells (Mouse Brain 10X) and black for reference cells (MOs snATAC).

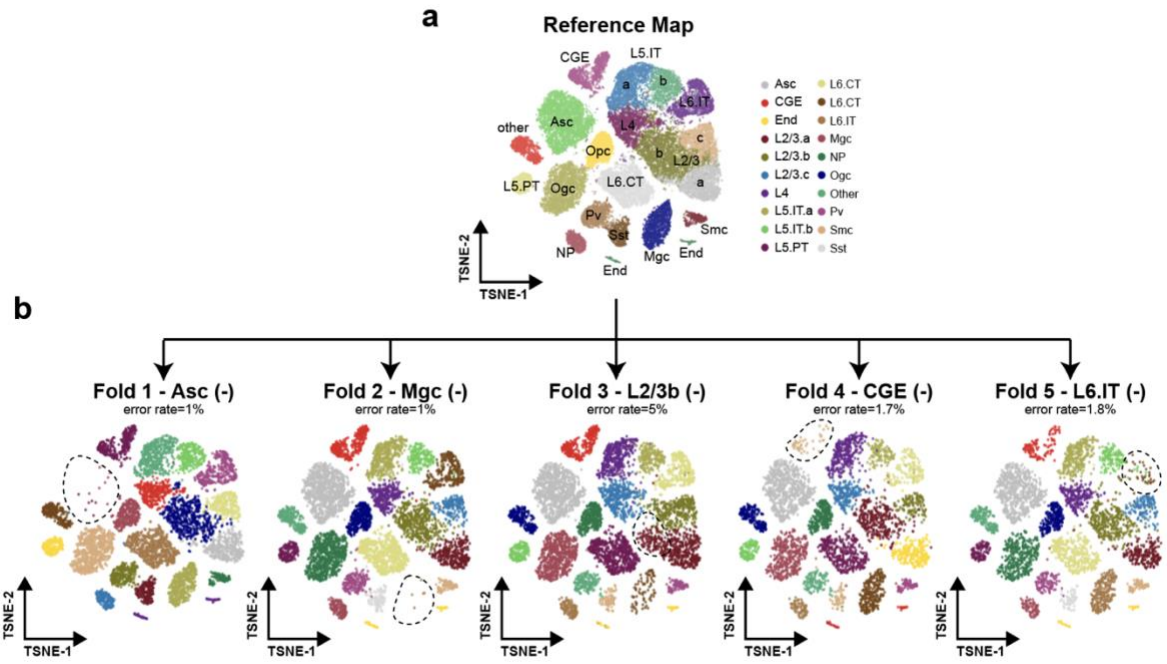


Figure S2.31. SnapATAC is robust for supervised annotation of datasets containing cell types missing in the reference atlas. (**a**) Two-dimensional t-SNE visualization of the reference dataset MOs (snATAC). (**b**) A five-fold cross validation is performed to this reference dataset. For each fold, we introduce perturbation to the 80% training dataset by randomly dropping one cell type (Asc, Mgc, L2/3b, CGE and L6.IT). We then predict on the 20% test dataset using the model learned from the perturbed training dataset. The prediction accuracy for each fold is shown in (**b**) and cell type removed from the training dataset are highlighted by the dash-line circles.

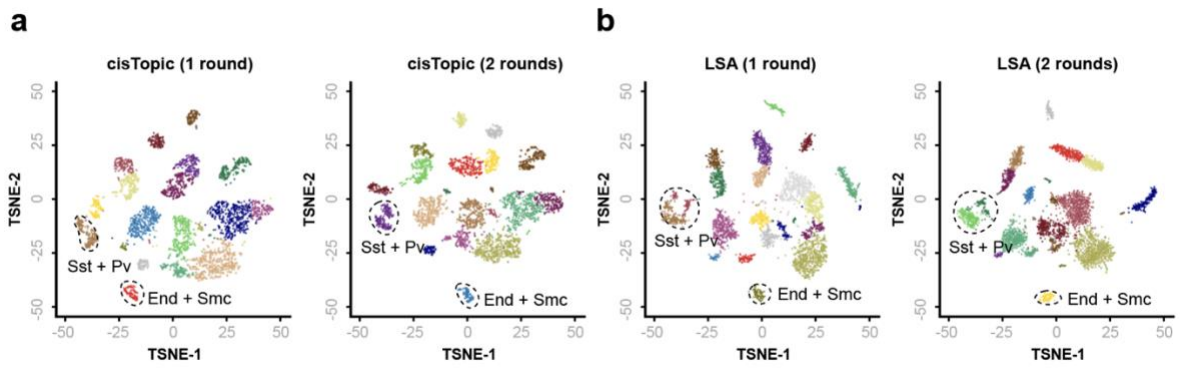


Figure S2.32. Iterative clustering does not substantially improve the clustering sensitivity. One approach[8] aims to overcome the bias introduced by population-level peak annotation by involving iterative clustering, with the first round defining the "crude" clusters in complex tissues followed by identifying peaks in these clusters, which are then used in subsequent round(s) of clustering. To test if this method can improve the sensitivity of identifying rare cell, we apply it to a 10X scATAC-seq dataset from mouse brain using both LSA and cisTopic. We first identify the major types and define peaks in each of clusters of more than 100 cells. We then merge these peaks to create a master peak reference and create a new cell-by-peak matrix for clustering. Iterative clustering result (2 rounds) is compared to 1-round clustering for both cisTopic (**a**) and LSA (**b**). Dash line circles highlight rare populations identified by SnapATAC as shown in **Figure S2.15**.

2.10 References

1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

2. Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120 (2012).

3. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).

4. Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S. & Crawford, G. E. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell 132, 311–322 (2008).

5. Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C. & Shendure, J. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science 348, 910–914 (2015).

6. Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R. M., Aghamirzaie, D., Marco-Ferreres, R., Pliner, H. A., Christiansen, L., Qiu, X., Steemers, F. J., Trapnell, C., Shendure, J. & Furlong, E. E. M. The cis-regulatory dynamics of embryonic development at single-cell resolution. Nature 555, 538–542 (2018).

7. Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D. U., Zhang, Y., Sos, B. C., Afzal, V., Dickel, D. E., Kuan, S., Visel, A., Pennacchio, L. A., Zhang, K. & Ren, B. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nat. Neurosci. 21, 432 (2018).

8. Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., Lee, C., Regalado, S. G., Read, D. F., Steemers, F. J., Disteche, C. M., Trapnell, C. & Shendure, J. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174, 1309-1324.e18 (2018).

9. Lake, B. B., Chen, S., Sos, B. C., Fan, J., Kaeser, G. E., Yung, Y. C., Duong, T. E., Gao, D., Chun, J., Kharchenko, P. V. & Zhang, K. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat. Biotechnol. 36, 70–80 (2018).

10. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y. & Greenleaf, W. J. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature 523, 486–490 (2015).

11. Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., Farmer, A., Fordyce,

P., Linnarsson, S. & Greenleaf, W. High-throughput chromatin accessibility profiling at single-cell resolution. Nat. Commun. 9, 3647 (2018).

12. Lareau, C. A., Duarte, F. M., Chew, J. G., Kartha, V. K., Burkett, Z. D., Kohlway, A. S., Pokholok, D., Aryee, M. J., Steemers, F. J., Lebofsky, R. & Buenrostro, J. D. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat. Biotechnol. (2019). doi:10.1038/s41587-019-0147-6

13. Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., Olsen, B. N., Mumbach, M. R., Pierce, S. E., Corces, M. R., Shah, P., Bell, J. C., Jhutty, D., Nemec, C. M., Wang, J., Wang, L., Yin, Y., Giresi, P. G., Chang, A. L. S., Zheng, G. X. Y., Greenleaf, W. J. & Chang, H. Y. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. bioRxiv (2019). doi:10.1101/610550

14. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods 14, 975–978 (2017).

15. Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J. & Aerts, S. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat. Methods 16, 397–400 (2019).

16. de Boer, C. G. & Regev, A. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. BMC Bioinformatics 19, (2018).

17. Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., Adey, A. C., Steemers, F. J., Shendure, J. & Trapnell, C. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Mol. Cell 71, 858-871.e8 (2018).

18. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902.e21 (2019).

19. Coifman, R. R. & Lafon, S. Diffusion maps. Appl. Comput. Harmon. Anal. 21, 5–30 (2006).

20. Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. & Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proc. Natl. Acad. Sci. 102, 7426–7431 (2005).

21. Kumar, S., Mohri, M. & Talwalkar, A. Ensemble Nystrom Method. 9

22. Li, M., Kwok, J. T. & Lu, B.-L. Making Large-Scale Nyström Approximation Possible. 12

23. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. 8, 281-291.e9 (2019).

24. Korsunsky, I., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R. & Raychaudhuri, S. Fast, sensitive, and accurate integration of single cell data with Harmony. bioRxiv (2018). doi:10.1101/461954

25. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell 38, 576–589 (2010).

26. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. & Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. 28, 495–501 (2010).

27. Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293 (2009).

28. Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C. & Knight, J. C. Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science 343, 1246949–1246949 (2014).

29. Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B. & Seelig, G. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science 360, 176–182 (2018).

30. Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E. & Dudoit, S. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics 19, (2018).

31. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V. V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331–336 (2015).

32. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014).

33. Huang, Z. J. & Paul, A. The diversity of GABAergic neurons and neural communication elements. Nat. Rev. Neurosci. (2019). doi:10.1038/s41583-019-0195-4

34. Tasic, B., Menon, V., Nguyen, T. N., Kim, T. K., Jarsky, T., Yao, Z., Levi, B., Gray, L.

T., Sorensen, S. A., Dolbeare, T., Bertagnolli, D., Goldy, J., Shapovalova, N., Parry, S., Lee, C., Smith, K., Bernard, A., Madisen, L., Sunkin, S. M., Hawrylycz, M., Koch, C. & Zeng, H. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat. Neurosci. 19, 335–346 (2016).

35. Mayer, C., Hafemeister, C., Bandler, R. C., Machold, R., Batista Brito, R., Jaglin, X., Allaway, K., Butler, A., Fishell, G. & Satija, R. Developmental diversification of cortical inhibitory interneurons. Nature 555, 457–462 (2018).

36. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. 35, D88–D92 (2007).

37. Gorkin, D., Barozzi, I., Zhang, Y., Lee, A. Y., Lee, B., Zhao, Y., Wildberg, A., Ding, B., Zhang, B., Wang, M., Strattan, J. S., Davidson, J. M., Qiu, Y., Afzal, V., Akiyama, J. A., Plajzer-Frick, I., Pickle, C. S., Kato, M., Garvin, T. H., Pham, Q. T., Harrington, A. N., Mannion, B. J., Lee, E. A., Fukuda-Yuzawa, Y., He, Y., Preissl, S., Chee, S., Williams, B. A., Trout, D., Amrhein, H., Yang, H., Cherry, J. M., Shen, Y., Ecker, J. R., Wang, W., Dickel, D. E., Visel, A., Pennacchio, L. A. & Ren, B. Systematic mapping of chromatin state landscapes during mouse development. bioRxiv 166652 (2017). doi:10.1101/166652

38. Phillips, J. E. & Corces, V. G. CTCF: Master Weaver of the Genome. Cell 137, 1194–1211 (2009).

39. Kageyama, R., Ishibashi, M., Takebayashi, K. & Tomita, K. bHLH Transcription factors and mammalian neuronal differentiation. Int. J. Biochem. Cell Biol. 29, 1389–1399 (1997).

40. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008 (2008).

41. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. bioRxiv (2019). doi:10.1101/576827

CHAPTER 3: MAPPING OF LONG-RANGE CHROMATIN INTERACTIONS BY PROXIMITY LIGATION-BASED CHIP-SEQ

3.1 Abstract

Formation of long-range chromatin loops is a crucial step in transcriptional activation of target genes by distal enhancers. Mapping such structural features can help define target genes for enhancers and annotate non-coding sequence variants linked to human diseases. Here we present PLAC-seq, a cost-efficient method to map chromatin conformation. PLAC-seq improves nearly 10-fold improvement on the detection efficiency, reduces over 100-fold input requirement and lowers at least 10-fold cost compared to prior technique in detection of long-range chromatin interactions in mammalian cells.

3.2 Introduction

Formation of long-range chromatin loops is a crucial step in transcriptional activation of target genes by distal enhancers[1]. Mapping such structural features can help define target genes for enhancers and annotate non-coding sequence variants linked to human diseases[1–3]. Study of the higher-order chromatin organization has been facilitated by the development of chromosome conformation capture (3C)-based technologies[4,5]. Among the commonly used high-throughput 3C approaches are Hi-C[6] and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)[7]. Global analysis of long-range chromatin interactions using Hi-C has been achieved at kilobase resolution but requires billions of sequencing reads[8]. High-resolution analysis of long-range chromatin interactions at selected genomic regions can be attained cost-effectively through either ChIA-PET[7,9] or targeted capture and sequencing of Hi-C libraries[10]. ChIA-PET has been used to identify long-range interactions at promoters and enhancers at high resolution in various cell types and species[11]. However, this procedure requires hundreds of million cells as starting materials, likely because chromatin immunoprecipitation and proximity ligation are performed after chromatin shearing, which potentially leads to great disruption of protein/DNA complexes. To reduce the amount of input materials and improve the sensitivity and robustness of the assay, we developed Proximity Ligation-Assisted ChIP-seq (PLAC-seq), in which proximity ligation is conducted in nuclei prior to chromatin shearing and immunoprecipitation (**Figure 3.1a**; **Figure S3.1a**). We demonstrated that by switching the order of proximity ligation and chromatin shearing steps, PLAC-seq greatly improves the efficiency and accuracy over ChIA-PET[7,9] in detection of long-range chromatin interactions in mammalian cells.

3.3 Results

We performed PLAC-seq in mouse embryonic stem (ES) cells using antibodies against RNA Polymerase II (Pol II), H3K4me3 and H3K27ac to determine long-range chromatin interactions at promoters and enhancers in the genome. As shown in **Figure 3.1b**, PLAC-seq yielded libraries with higher number of unique read pairs compared with ChIA-PET. As expected, the sequencing reads were strongly enriched at the factor-binding sites detected by ChIP-seq analysis in the mouse ES cells[12] (**Supplementary Methods**; **Figure S3.1b-d**; **S3.1f-h**). Additionally, the PLAC-seq experiments generated long-range chromatin contacts that were highly reproducible between biological replicates (Pearson correlation > 0.90; **Supplementary Methods**; **Figure S3.1e**). To identify long-range chromatin interactions, we used 'FitHiC'[13] to analyze the combined datasets from two biological replicates (**Supplementary Methods**). A total of 72 074, 273 145, and 155 545 chromatin loops (FDR < 0.01) were identified from the Pol II, H3K4me3, and H3K27ac PLAC-seq experiments, respectively. We found that PLAC-seq could be performed with much fewer cells than ChIA-PET. Even with 0.5 million (M) cells, a majority of strong long-range interactions could be detected (**Figure 3.1c**; **Figure S3.1i** and **Supplementary Methods**).

Several lines of evidence support the superior performance of PLAC-seq over ChIA-PET. First, PLAC-seq was nearly 100 times more cost-effective than ChIA-PET in generating long-range intra-chromosomal read pairs, which are typically used to infer chromatin loops. Using 20-fold fewer cells (5 M vs 100 M), Pol II PLAC-seq produced 10 times more reads (175 M vs 16 M) with lower PCR duplication rate (30% vs 44%) than a

previously published Pol II ChIA-PET experiment[14]. In addition, PLAC-seq generated more long-range intra-chromosomal pairs (67% vs 9%) and fewer inter-chromosomal pairs (11% vs 48%) (**Figure 3.1b**). Second, PLAC-seq uncovered chromatin loops in the mouse ES cells with much higher sensitivity and specificity than ChIA-PET. Additionally, PLAC-seq chromatin interactions were typically supported by 24 unique read pairs (medium) compared to 3 PETs supporting ChIA-PET interactions[14] (**Figure 3.1d**). Pol II PLAC-seq analysis identified 57% of Pol II ChIA-PET interactions (FDR < 0.05 and PET count >= 3, 10 kb to 3Mb) and a lot of additional interactions (**Figure 3.1e**). PLAC-seq covered more regulatory elements, such as promoters and distal DNase I hypersensitive sites (DHSs), than ChIA-PET (**Supplementary Methods**; **Figure S3.1j**). As a reference, we performed *in situ* Hi-C with the mouse ES cell line and collected nearly 1.2 billion paired-end sequencing reads, from which we identified 68 781 long-range chromatin interactions (FDR < 0.01) using FitHiC[13]. Compared with chromatin interactions identified by *in situ* Hi-C, PLAC-seq is 8 times more sensitive than ChIA-PET and also more accurate (**Figure 3.1f**). Third, we performed 4C-seq analysis of four randomly selected genomic regions (**Supplementary Methods**). Although both ChIA-PET and PLAC-seq identified many common chromatin interactions (**Figure 3.1g**; **Supplementary Methods**; **Figure S3.2b,c**), PLAC-seq uncovered seven additional strong interactions (marked 2, 4 and 5 in **Figure 3.1g**, and 1-4 in **Supplementary Methods**, **Figure S3.2a-c**) detected by 4C-seq. Taken together, the results above support the superior sensitivity and specificity of PLAC-seq over ChIA-PET.

We also developed a new computational algorithm to identify chromatin interactions at high resolution from PLAC-seq data. We used the binomial test (**Supplementary Methods**) to determine the enrichment of read pairs for an interaction due to chromatin immunoprecipitation using *in situ* Hi-C analysis result as an estimation of background interaction frequency (**Figure 3.1h**). We termed this type of interactions as 'PLACE' (PLAC-Enriched) interactions. A total of 28 822 and 19 429 significant H3K4me3 and H3K27ac PLACE interactions (FDR < 0.05) in the mouse ES cells were identified, respectively. These corresponded to different sets of chromatin interactions, with 26% of H3K27ac PLACE interactions overlapping with 19% of H3K4me3 PLACE interactions (**Figure 3.1i**). A majority of H3K27ac PLACE interactions were enhancer-associated (74%) while H3K4me3 PLACE interactions were generally promoter-associated (78%) (**Figure 3.1j**). Genes involved in H3K27ac PLACE interactions had significantly higher expression levels than genes associated with H3K4me3 PLACE interactions (P < 2.2e-16, **Figure 3.1k**), suggesting that H3K27ac PLAC-seq could be used to discover chromatin interactions at active enhancers and H3K4me3 PLAC-seq at active or poised promoters.

In summary, we developed a fast, sensitive and cost-effective method to map long-range chromatin interactions in mammalian cells. Using PLAC-seq, we obtained high-resolution maps of chromatin interactions at enhancers and promoters in the mouse ES cells. The ease of experimental procedure and small amount of input materials required will allow the mapping of long-range chromatin interactions in a broad set of species, cell

types, and experimental settings. A similar method called HiChIP was recently reported by Mumbach *et al*.[15] when our manuscript was under review.

3.4 Acknowledgments

The work is supported by funding from the Ludwig Institute for Cancer Research and NIH (1U54DK107977-01, 2P50 GM085764 and U54 HG006997).

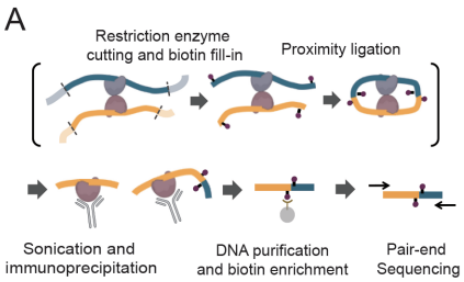
Chapter 3, in full, is a reprint of the material as it appears in Cell Research 2016 “Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq”. Rongxin Fang, Miao Yu, Guoqiang Li, Sora Chee, Tristin Liu, Anthony D Schmitt and Bing Ren. The dissertation author was the primary investigator and author of this paper.

3.5 Author Contributions

This study was conceived and designed by M.Y., R.F. and B.R.; Experiment performed by M.Y. Data analysis performed by R.F. Manuscript written by R.F., M.Y. and B.R. with input from all authors.

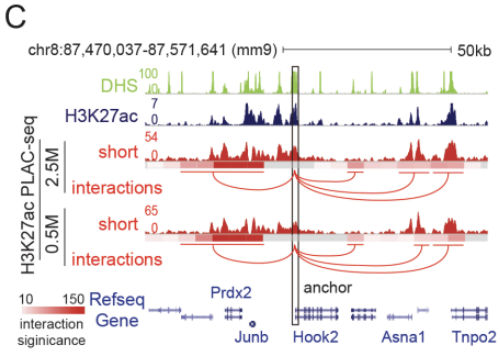
3.6 Figures

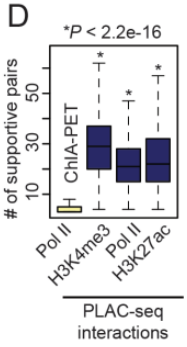
Figure 3.1. PLAC-seq reveals chromatin interactions in mammalian cells at high sensitivity and accuracy. (**a**) Overview of the PLACseq workflow. Formaldehyde-fixed cells were permeabilized and digested with a 4-bp cutter MboI, followed by biotin-tagged nucleotide fill-in and in situ proximity ligation. Nuclei were then lysed and the chromatin was sheared by sonication. The soluble chromatin fraction was then subjected to immunoprecipitation using specific antibodies against a transcription factor or a histone modification. Finally, after reverse-crosslinking the biotin-labeled DNA corresponding to ligation junctions was enriched followed by library preparation and paired-end DNA sequencing. (**b**) Comparison of the sequence outputs between PLAC-seq and ChIA-PET. (**c**) Comparison of short-range signals (short) and long-range chromatin interactions (interactions) identified by H3K27ac PLAC-seq using 2.5 M and 0.5 M cells in the indicated genomic region. Only the interactions with one end overlapping with a selected anchor point (chr8: 87 510 000-87 515 000, black rectangle) were shown. PLAC-seq interactions are marked by red arcs and interaction significance is denoted by –log (FDR). (**d**) Box plots of number of the unique read pairs supporting interactions identified by ChIA-PET and PLAC-seq. (**e**) Venn-diagram comparing the chromatin loops identified in Pol II PLAC-seq and Pol II ChIA-PET experiments. (**f**) Comparison of sensitivity (SE) and accuracy (AC) between PLAC-seq and ChIA-PET interactions using the loops detected by in situ Hi-C as a reference (SE = number of in situ HiC interactions overlapping with PLAC-seq or ChIA-PET interactions / total number of in situ HiC interactions; AC = number of PLAC-seq or ChIA-PET interactions overlapping with in situ HiC interactions / total number of PLAC-seq or ChIA-PET interactions). (**g**) Comparison of chromatin interactions identified by PLAC-seq, ChIA-PET and 4C-seq at the Mreg promoter (the anchor point is marked by a black rectangle, chr1: 72 255 000-72 260 000). PLAC-seq and ChIA-PET interactions were demonstrated by red and blue arcs, respectively; significance of interactions in PLAC-seq is denoted by –log (FDR). (H) Normalized Pol II PLAC-seq signals and PLACE (**Supplementary Methods**) analysis revealed chromatin interactions between Sox2 and its super enhancer at nearly single-element resolution (anchor region, chr3: 34 546 927-34 553 382). (**i**) Overlap between H3K27ac and H3K4me3 PLACE interactions. (**j**) Distribution of promoter-promoter (P-P), promoter-enhancer (P-E), enhancer-enhancer (E-E) and other interactions for H3K27ac and H3K4me3 PLACE interactions. (**k**) Boxplot of expression of different groups of genes. H3K27ac PLACE interactions are associated with genes with significantly higher expression than other genes ($P < 2.2e-16$). 2.5 M cells were used for H3K27ac PLAC-seq experiments in d, j and k.

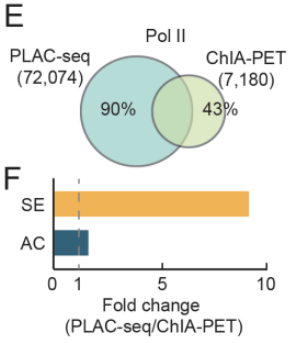


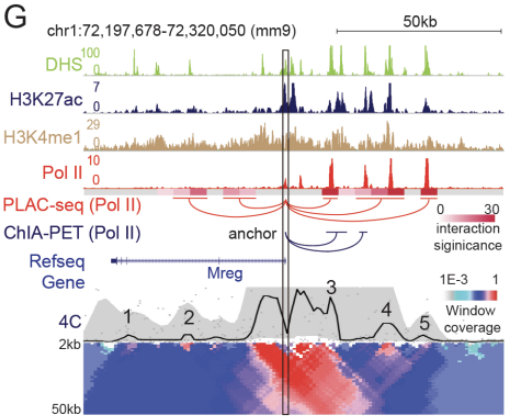
B

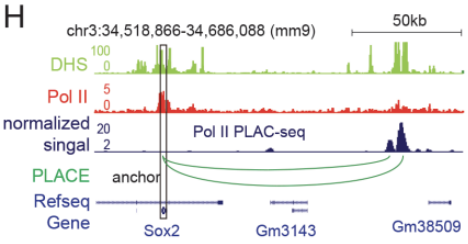
Statistic	Pol II PLAC-seq	Pol II ChIA-PET[14]
Cell number	5M	100M
Total pairs	175M	16M
PCR duplication	30%	44%
cis pairs (%)	89%	52%
long-range *cis* pairs	67% (>10kb)	8.6% (>8kb)
Processing time	3-4 Days	6 Days

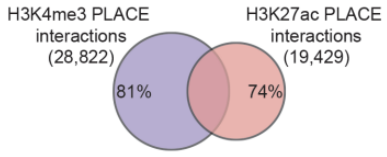






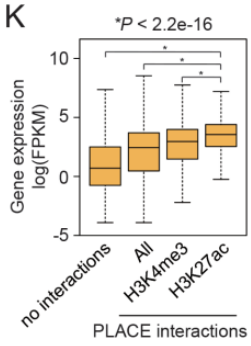






J

PLACE interactions	Categoty of PLACE interactions			
	P-P	P-E	E-E	Other
H3K4me3	33%	45%	6%	16%
H3K27ac	10%	37%	37%	16%



3.7 Supplementary Methods

Cell culture and fixation. The F1 Mus musculus castaneus × S129/SvJae mouse ESC line (F123 line) was a gift from Dr. Rudolf Jaenisch and was previously described[1]. F123 cells were cultured as described previously[2]. Cells were passaged once on 0.1% gelatin-coated feeder-free plates before fixation.

To fix the cells, cells were harvested after accutase treatment and suspended in medium without Knockout Serum Replacement at a concentration of 1x106 cells per 1ml. Methanolfree formaldehyde solution was added to the final concentration of 1% (v/v) and rotated at room temperature for 15 min. The reaction was quenched by addition of 2.5 M glycine solution to the final concentration of 0.2 M with rotation at room temperature for 5 min. Cells were pelleted by centrifugation at 3,000 rpm for 5 min at 4 o C and washed with cold PBS once. The washed cells were pelleted again by centrifugation, snap-frozen in liquid nitrogen and stored at -80 oC.

PLAC-seq. PLAC-seq is comprised of three procedures: in situ proximity ligation, chromatin immunoprecipitation or ChIP, biotin pull-down followed by library construction and sequencing. The in situ proximity ligation and biotin pull-down procedures were similar to previously published in situ Hi-C protocol[3] with minor modifications as described below: 1. In situ proximity ligation. 0.5 to 5 million of crosslinked F123 cells were thawed on ice, lysed in cold lysis buffer (10 mM Tris, pH 8.0, 10 mM NaCl, 0.2% IGEPAL CA-630 with proteinase inhibitor) for 15 min, followed by a washing with lysis buffer once. Cells were then resuspended in 50 µl 0.5% of SDS and incubated at 62 oC for 10 min.

Permeabilization was quenched by adding 25 µl 10% Triton X-100 and 145 µl water, and incubation at 37 oC for 15 min. After addition of NEBuffer 2 to 1x and 100 units of MboI, the digestion was performed for 2 h 37 oC in a thermomixer, shaking at 1,000 rpm. Following inactivation of MboI at 62 oC for 20 min, biotin fill-in reaction was performed for 1.5 h 37 oC in a thermomixer after adding 15 nmol of dCTP, dGTP, dTTP, biotin-14-dATP (Thermo Fisher Scientific) each and 40 unit of Klenow. Proximity ligation was then performed at room temperature with slow rotation in a total volume of 1.2 ml containing 1xT4 ligase buffer, 0.1 mg/ml BSA, 1% Triton X-100 and 4000 unit of T4 ligase (NEB). 2. Chromatin immunoprecipitation (ChIP). After proximity ligation, the nuclei were spun down at 2,500 g for 5 min and the supernatant was discarded. The nuclei were then resuspended in 130 µl RIPA buffer (10 mM Tris, pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate) with proteinase inhibitors. The nuclei were lysed on ice for 10 min and then sonicated using Covaris M220 with following setting: power, 75 W; duty factor, 10%; cycle per burst, 200; time, 10 min; temp, 7 oC. After sonication, the samples were cleared by centrifugation at 14,000 rpm for 20 min and supernatant was collected. The clear cell lysate was mixed with Protein G Sepharose beads (GE Healthcare) and then rotated at 4 oC for pre-cleaning. After 3h, supernatant was collected and ~5% of lysate was saved as input control. The rest of the lysate was mixed with 2.5 µg of H3K27Ac (ab4729, Abcam), H3K4me3 (04-745, Millipore) or 5 µg Pol II (ab817, Abcam) specific antibody and rotate at 4 oC overnight. On the next day, 0.5% BSA-blocked Protein G Sepharose beads (prepared one day ahead) were added and rotated for another 3 h at 4 oC. The beads were collected by centrifugation at 2,000 rpm for 1 min and then washed with RIPA buffer three times, high-salt RIPA buffer (10

mM Tris, pH 8.0, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% sodium deoxycholate) twice, LiCl buffer (10 mM Tris, pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% IGEPAL CA-630, 0.1% sodium deoxycholate) once, TE buffer (10 mM Tris, pH 8.0, 0.1 mM EDTA) twice. Washed beads were first treated with 10 µg Rnase A in extraction buffer (10 mM Tris, pH 8.0, 350 mM NaCl, 0.1 mM EDTA, 1% SDS) for 1 h at 37 oC. Then 20 µg proteinase K was added and reverse crosslinking was performed overnight at 65 oC or at least 2 h. The fragmented DNA was purified by Phenol/Chloroform/Isoamyl Alcohol (25:24:1) extraction and then ethanol precipitation.

Biotin pull-down and library construction. The biotin pull-down procedure was performed according to in situ Hi-C protocol with the following modifications: 1) 20 µl of Dynabeads MyOne Streptavidin T1 beads were used per sample instead of 150 µl; 2) To maximize the PLAC-seq library complexity, the minimal number of PCR cycles for library amplification was determined by qPCR.

PLAC-seq sequencing read mapping. We developed a bioinformatics pipeline (https://github.com/r3fang/PLACseq) to map PLAC-seq and in situ Hi-C data. Paired-end sequencing reads were mapped using BWA-MEM[4] to the reference genome (mm9) in single-end mode with default setting for each of the two ends separately. The independently mapped ends were then paired-up and the read pairs were kept if both ends uniquely mapped to the genome (MQAL>10). Inter-chromosomal pairs were discarded. Next, read pairs were further removed if either end was mapped more than 500bp apart away from the closest MboI site. Read pairs were next sorted based on

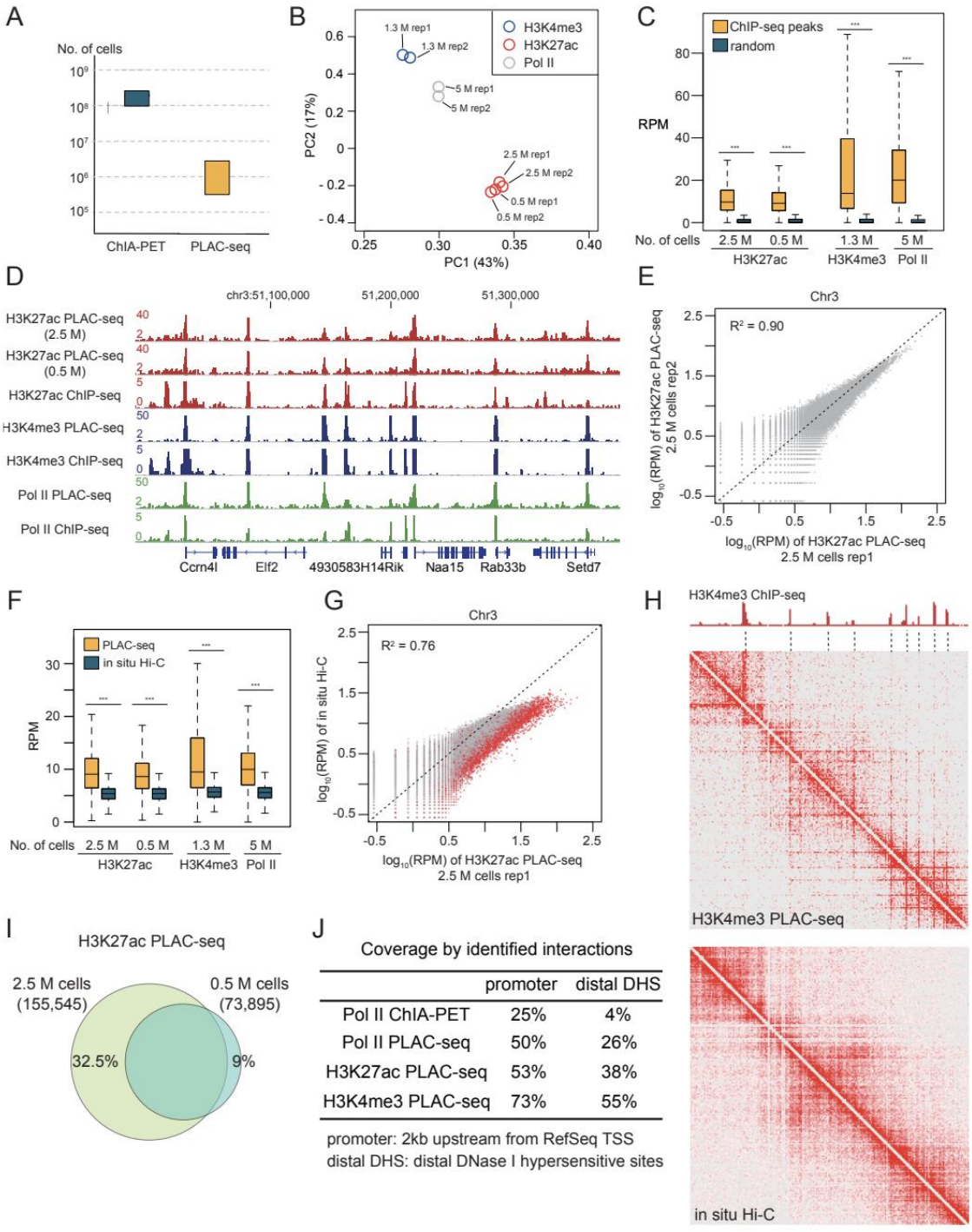
genomic coordinates followed by PCR duplicate removal using MarkDuplicates in Picard tools[5]. Finally, the mapped pairs were partitioned into “long-range” and “short-range” based on the distance between the two ends, with a threshold of larger than 10kb or smaller than 1kb, respectively.

Identification of chromatin loops from PLAC-seq and in situ Hi-C datasets. The algorithm ‘FitHiC’[6] was used to identify long-range interactions (from 10kb to 3MB) in PLAC-seq and in situ Hi-C datasets with 5kb resolution. The P-values were adjusted to FDR using Benjamini and Hochberge approacy[7]. We consider a chromatin interaction significant if the FDR was less than 0.01. In total, we identified 86,629, 290,350, 204,232 and 89,970 significant long-range interactions from Pol II, H3K4me3 and H3K27ac (2.5M) and H3K27ac (0.5M) PLAC-seq, with 83%, 94%, 76% and 82% occupied by corresponding ChIP-seq peaks. We next filtered out interactions that were not occupied by corresponding ChIP-seq peaks. After filtering, there were 72,074, 273,145, 155,545, and 73,895 significant long-range interactions from Pol II, H3K4me3 and H3K27ac (2.5M) and H3K27ac (0.5M) PLAC-seq remaining. Using the same algorithm and FDR cutoff, we also identified 68,781 interactions from our in situ Hi-C data.

Analysis of overlaps between chromatin loops identified in different datasets. We defined that two distinct interactions were overlapped if both ends of each interaction intersect by at least one base pair.

3.8 Supplementary Figures

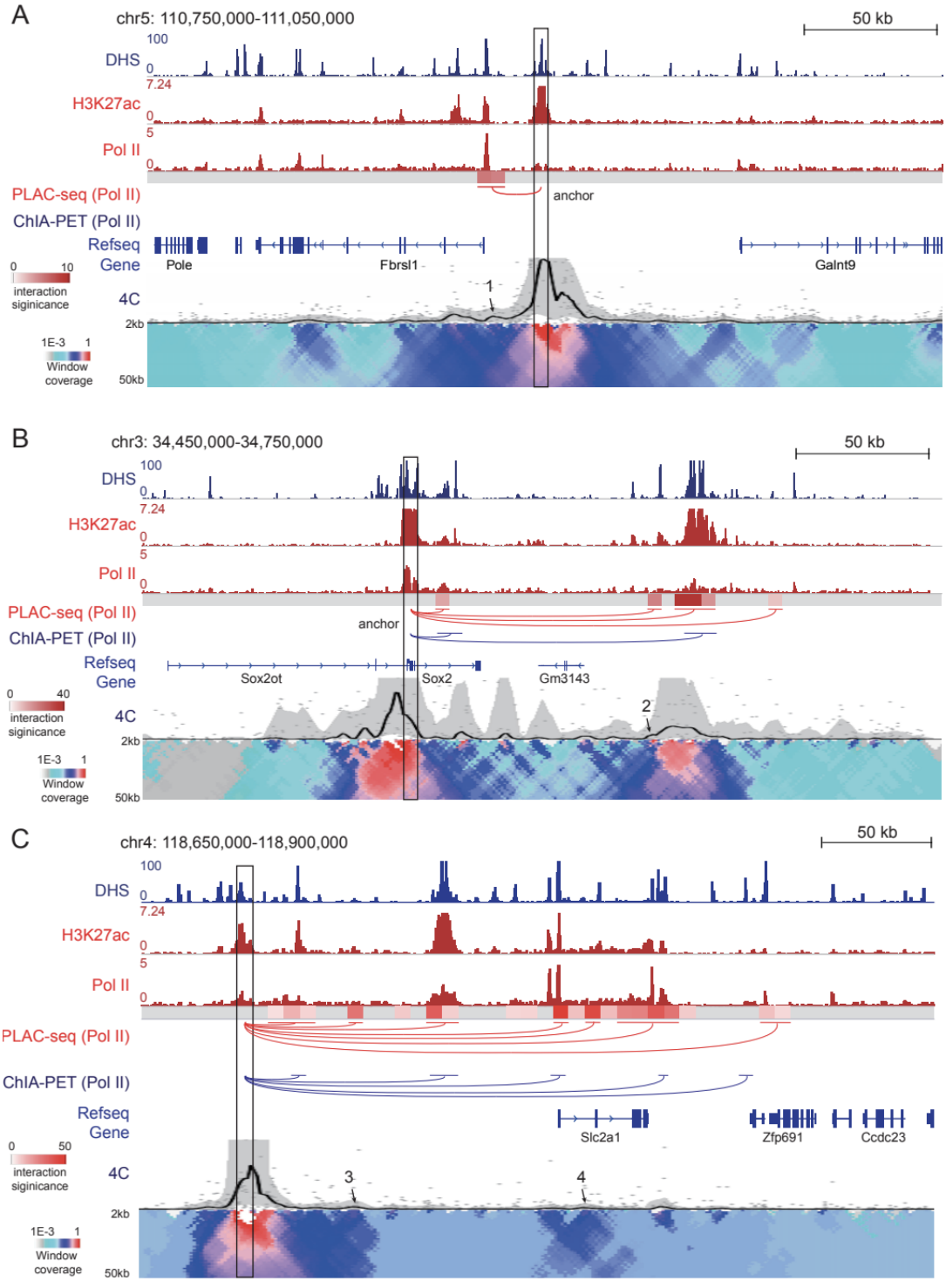
Figure S3.1. Development and validation of PLAC-seq. (**a**) Comparison of input material requirement of PLAC-seq and ChIA-PET. (**b**) Principal component analysis (PCA) of short-range reads in different PLAC-seq experiments highlights the reproducibility between biological replicates. (**c**) Box plots of reads per million (RPM) calculated using PLAC-seq short-range cis pairs (distance < 1kb) suggest that PLAC-seq signals are significantly enriched in ChIP-seq peaks compared to randomly chosen regions (***Wilcoxon tests, P < 2.2e-16). (**d**) The signals of short-range reads (< 1kb) from PLAC-seq were similar to those of ChIP-seq performed on the same set of factors in the mouse ES cells. (**e**) Scatter plots of pair-wise interaction frequency on chromosome 3. PLAC-seq biological replicates were highly reproducible (R2 =0.90). For the other datasets: H3K27ac, 0.5 M cells, between biological replicates, R2 =0.86; H3K4me3, 1.3 M cells, between biological replicates, R2 =0.90; Pol II, 5 M cells, between biological replicates, R2 =0.81. (**f-h**) Long-range cis reads from PLAC-seq were significantly enriched in the ChIP-seq peak regions compared to in situ Hi-C. (F) Box plots of reads per million (RPM) at ChIP-enriched regions for PLAC-seq and in situ Hi-C. Only long-range (>10kb) cis reads were considered (***Wilcoxon tests, P < 2.2e16). (**g**) Scatter plots of pair-wise interaction frequency on chromosome 3 are shown. Interaction intensity is skewed towards PLAC-seq for fragments with H3K27ac ChIP-seq peaks compared to in situ Hi-C (R2 =0.76, Red dots represent fragment pairs with at least one end bound by H3K27ac). For the other datasets: H3K27ac, 0.5 M cells, between replicate 1 and in situ Hi-C, R2 =0.79; H3K4me3, 1.3 M cells, between replicate 1 and in situ Hi-C, R2 =0.72; Pol II, 5 M cells, between replicate 1 and in situ Hi-C, R2 =0.67. (**h**) Examples of enrichment of long-range cis reads in H3K4me3 PLAC-seq compared to in situ Hi-C (visualized by Juicebox). (**i**) Long-range chromatin interactions identified by H3K27ac PLAC-seq were highly reproducible using 2.5 million and 0.5 million cells. (**j**) Comparison of coverage of promoters and distal cis regulatory elements between PLAC-seq and ChIA-PET analyses. H3K27ac PLAC-seq refers to the experiment using 2.5 million cells.



	promoter	distal DHS
Pol II ChIA-PET	25%	4%
Pol II PLAC-seq	50%	26%
H3K27ac PLAC-seq	53%	38%
H3K4me3 PLAC-seq	73%	55%

promoter: 2kb upstream from RefSeq TSS
distal DHS: distal DNase I hypersensitive sites

Figure S3.2. Comparison of chromatin interactions detected by 4C-seq, PLAC-seq, and ChIA-PET at three genomic loci. PLAC-seq and ChIA-PET interactions were demonstrated by red and blue arcs, respectively; significance of interactions in PLAC-seq is –log (FDR). 1-4 mark the 4C interactions identified by Pol II PLAC-seq but not ChIA-PET. Only the interactions with one end overlapping with a selected anchor points (marked by black rectangles) were shown. (**a**) Anchor point, chr5: 110,900,000-110,905,000. No interactions detected by Pol II ChIAPET. (**b**) Anchor point, chr3: 34545000-34,550,000. (**c**) Anchor point, chr4: 118680000- 118685000.


A
chr5: 110,750,000-111,050,000
50 kb
DHS
H3K27ac
Pol II
PLAC-seq (Pol II)
ChIA-PET (Pol II)
Refseq Gene
anchor
Pole
Fbrsl1
Galnt9
interaction siginicance
4C
Window coverage
2kb
50kb
B
chr3: 34,450,000-34,750,000
Sox2ot
Sox2
Gm3143
C
chr4: 118,650,000-118,900,000
Slc2a1
Zfp691
Ccdc23

3.9 References

1. Gorkin, D. U., Leung, D. & Ren, B. The 3D Genome in Transcriptional Regulation and Pluripotency. Cell Stem Cell 14, 762–775 (2014).

2. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature 502, 499–506 (2013).

3. Sexton, T. & Cavalli, G. The Role of Chromosome Domains in Shaping the Functional Genome. Cell 160, 1049–1059 (2015).

4. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics 14, 390–403 (2013).

5. Denker, A. & de Laat, W. The second decade of 3C technologies: detailed insights into nuclear organization. Genes & Development 30, 1357–1382 (2016).

6. Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. & Dekker, J. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 326, 289–293 (2009).

7. Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., Chew, E. G. Y., Huang, P. Y. H., Welboren, W.-J., Han, Y., Ooi, H. S., Ariyaratne, P. N., Vega, V. B., Luo, Y., Tan, P. Y., Choy, P. Y., Wansa, K. D. S. A., Zhao, B., Lim, K. S., Leow, S. C., Yow, J. S., Joseph, R., Li, H., Desai, K. V., Thomsen, J. S., Lee, Y. K., Karuturi, R. K. M., Herve, T., Bourque, G., Stunnenberg, H. G., Ruan, X., Cacheux-Rataboul, V., Sung, W.-K., Liu, E. T., Wei, C.-L., Cheung, E. & Ruan, Y. An oestrogen-receptor-α-bound human chromatin interactome. Nature 462, 58–64 (2009).

8. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014).

9. Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-L., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G. & Ruan, Y. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell 163, 1611–1627 (2015).

10. Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs,

A., LeProust, E., Follows, G. A., Fraser, P., Luscombe, N. M. & Osborne, C. S. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature Genetics 47, 598–606 (2015).

11. Li, G., Cai, L., Chang, H., Hong, P., Zhou, Q., Kulakova, E. V., Kolchanov, N. A. & Ruan, Y. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics 15, (2014).

12. Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120 (2012).

13. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research 24, 999–1011 (2014).

14. Zhang, Y., Wong, C.-H., Birnbaum, R. Y., Li, G., Favaro, R., Ngan, C. Y., Lim, J., Tai, E., Poh, H. M., Wong, E., Mulawadi, F. H., Sung, W.-K., Nicolis, S., Ahituv, N., Ruan, Y. & Wei, C.-L. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. Nature 504, 306–310 (2013).

15. Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J. & Chang, H. Y. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature Methods 13, 919–922 (2016).

CHAPTER 4: A TILING-DELETION BASED GENETIC SCREEN FOR CIS-REGULATORY ELEMENT IDENTIFICATION IN MAMMALIAM CELLS

4.1 Abstract

Millions of *cis*-regulatory elements are predicted in the human genome, but direct evidence for their biological function is still scarce. Here we report a high-throughput method, *Cis*-Regulatory Element Scan by Tiling-deletion and sequencing (CREST-seq), for unbiased discovery and functional assessment of *cis* regulatory sequences in the genome. We use it to interrogate the 2Mbp *POU5F1* locus in the human embryonic stem cells and identify 45 *cis*-regulatory elements of *POU5F1*. A majority of these elements display active chromatin marks, DNase hypersensitivity and occupancy by multiple transcription factors, confirming the utility of chromatin signatures in *cis* elements mapping. Notably, 17 of them are previously annotated promoters of functionally unrelated genes, and like typical enhancers, they form extensive spatial contacts with the *POU5F1* promoter. Taken together, these results support the utility of CREST-seq for large-scale *cis* regulatory element discovery and point to commonality of enhancer-like promoters in the human genome.

4.2 Introduction

Millions of candidate *cis*-regulatory elements have been annotated in the human genome based on histone modification, transcriptional factor binding, and DNase I hypersensitivity[1–6]. These putative regulatory sequences harbor a disproportionally large number of sequence variants associated with diverse human traits and diseases, supporting the hypothesis that non-coding sequence variants contribute to common traits and diseases by disrupting transcriptional regulation[7–9]. However, research on the role of these putative functional elements in human development and disease has been hindered by a dearth of direct evidence for their biological function in the native genomic context.

High-throughput CRISPR/Cas9-mediated mutagenesis using single guide RNAs (sgRNAs) has been used to functionally characterize *cis*-regulatory elements in mammalian cells[10–15]. However, current approaches are limited because: (1) Not all sequences are suitable for CRISPR/Cas9-mediated genome editing due to the lack of protospacer adjacent motifs (PAMs) that are required for targeting and DNA cutting by CRISPR/Cas9[16–18]; (2) CRISPR/Cas9 mediated genome editing with individual sgRNAs tends to cause point mutations or short insertions or deletions, necessitating the use of an unrealistically large number of sgRNAs to interrogate the human genome; (3) it has been challenging to distinguish the *cis*- and *trans*-regulatory elements. To overcome these limitations, we developed CREST-seq, short for *Cis*-Regulatory Elements Scan by Tiling-deletion and Sequencing, which enables efficient discovery and functional characterization of *cis*-regulatory elements by introducing massively parallel, kilobase-long deletions to the genome. Below, we provide evidence supporting the utility of

CREST-seq for large-scale *cis*-regulatory element identification in the human embryonic stem cells (hESC). We report the discovery of 45 regulatory sequences of *POU5F1* and a surprisingly large number of enhancer-like promoters.

4.3 Results

CREST-seq identified *cis*-regulatory elements of *POU5F1*. In a CREST-seq experiment, a large number of overlapping genomic deletions are first introduced to a genomic locus by CRISPR/Cas9-mediated genome editing using paired sgRNAs[16] (**Figure 4.1a**). Cells with lowered expression of the gene of interest (**Figure 4.1b**) are then isolated and the enriched sgRNA pairs determined by high-throughput sequencing. The enriched sgRNA-pair sequences are then used to infer the functional *cis*-regulatory sequences of the gene (**Figure 4.1a**). To demonstrate the utility of CREST-seq, we applied it to the 2Mbp *POU5F1* locus. As a model cell system we used a hESC line in which one *POU5F1* allele was genetically tagged by *eGFP,* allowing transcription level of this allele to be monitored by eGFP expression[19] (**Figure 4.1b**).

We designed a total of 11,570 sgRNA pairs to introduce the same number of genomic deletions (**Figure 4.1a**; **Figure S4.1a**) to the *POU5F1* locus. The average size of each deletion is ~2kb, with an overlap of 1.9kb between two adjacent deletions (**Figure S4.1a**) such that each nucleotide in this locus is covered by ~20 distinct genomic deletions on average. As negative controls, we included 424 sgRNA oligos lacking the PAM sequence necessary for effective dsDNA breaks. As positive controls, we included six sgRNA pairs that target the *eGFP* coding sequence. We constructed a lentiviral library that express these sgRNA pairs (**Figure S4.2a-e**) and transduced it into the hESC line at low multiplicity of infection (MOI = 0.1), which ensures that the majority of cells receives one or no lentiviral particle (**Supplementary Methods**).

To isolate mutant cells with deletion in *POU5F1*'s *cis*-regulatory sequences, we used FACS to sort out cells showing lowered *POU5F1* expression from the *eGFP*-tagged allele but relatively unchanged expression from the non-tagged allele (**Figure 4.1c**). We refer to this *eGFP*-/*POU5F1*+ subpopulation as "Cis" population (**Figure 4.1b**, **c**). As a control, we also collected a sample of cells before FACS sorting (referred to as "Ctrl"). Finally, we collected the *eGFP*+/*POU5F1*+ population (referred to as "High") (**Figure 4.1b**, top; **Figure 4.1c**). Genomic DNA was purified from each cell populations, and the sgRNA pairs present in each subpopulation were then determined by massively parallel sequencing. The experiment was conducted in multiple replicates (**Figure S4.3a**), with the abundance of sgRNA pairs highly reproducible between replicates (Pearson Correlation Coefficients R=0.90 for "Cis", R=0.92 for "Ctrl" and R=0.97 for "High", respectively) (**Figure S4.3b**).

To identify *cis*-regulatory elements of *POU5F1*, we first compared the abundance of sgRNA pairs between the "Cis" population and the "Ctrl" population using a negative binomial test and computed the fold enrichment and *P*-value of each sgRNA pair (**Figure S4.3c**). We found 495 sgRNA pairs to be significantly enriched ($P < 0.05$ and log(fold change) > 1) in the "Cis" samples (**Figure 4.1d**, red dots; **Figure 4.1e** red bars). As expected, all six sgRNA pairs targeting the *eGFP* sequence were highly enriched in the "Cis" population (**Figure 4.1d**, green circles). By contrast, only 2 of the 424 negative control sgRNAs were enriched, corresponding to an empirical FDR smaller than 0.5%. Further supporting the effectiveness of our experimental design, the sgRNA pairs with significant enrichment in the "Cis" population were generally depleted in the "High"

samples (**Figure 4.1d**, right panel). Next, we sought to identify *cis*-regulatory sequences by taking full advantage of the tiling deletion design (**Figure 4.1e**). We began by ranking all sgRNA pairs based on their enrichment levels in the "Cis" population relative to the "Ctrl". We then partitioned the 2MB *POU5F1* locus into 50bp bins, and used Robust Rank Aggregation (RRA)[20] to calculate a score for each bin to indicate whether the ranks of deletions spanning that bin are skewed toward top of the sorted list (**Supplementary Methods**). Altogether, we identified 45 genomic regions with a significant score (**Figure 4.1e**). Using the same criteria, no genomic region was identified as positive in the "High" cell population (**Figure S4.4a**). We named each of the 45 CREST-positive elements (referred to hereafter as "CRE") using its relative genomic distance (kb) to the transcription start site (TSS) of *POU5F1*, with a negative sign denoting upstream of *POU5F1* and a positive for downstream. The 45 CREs include 4 previously identified *POU5F1*-regulatory elements that act in *cis*: its promoter (**Figure S4.4b**), an upstream enhancer[21] (**Figure S4.4b**) and two temporarily phenotypic (TEMP) enhancers[13] (**Figure S4.4c**, DHS_65 and DHS_108). The remaining 41 CREs are novel *POU5F1*-regulatory sequences found in this study.

CREs are enriched with active chromatin marks and dense TF clusters. In order to determine chromatin features of the CREs, we examined the publicly available chromatin accessibility data, transcription factor binding profiles and chromatin modification datasets from the H1 hESC cell line[3,5]. We also generated ATAC-seq[22] and CTCF ChIP-seq with the cell line used in the present study and ensured that the data highly resembles the previous datasets from the same parental cell line[5] (**Figure S4.5a,**

b). As expected, a majority of CREs were associated with biochemical features characteristic of *cis*-regulatory elements, including DNase Hypersensitivity (69%), transcription factor occupancy, active chromatin marks such as H3K27ac (22%), H3K4me3 (31%), and H3K4me1 (22%)[5]. Notably, CREs are also enriched for binding sites of CTCF/RAD21 (29%), which have been linked to DNA looping and topologically associating domain (TAD) boundaries[23,24] (**Figure 4.2a, b**). It has been reported that transcription factor binding in human cells tend to form dense clusters[25–27]. Accordingly, we found that the CREST-positive regions overlap with dense clusters of TF binding sites (16% CREs are bound by essential pluripotency master regulators and 44% by other TFs; **Figure 4.2a-c**) and are bound by more transcription factors on average than DNase hypersensitive sites (DHS) (**Figure 4.2d**, Wilcoxon tests *P*-value<6e-11). In general, CREST-positive regions are significantly associated with active histones modifications and transcription factor binding (**Figure 4.2e**), and depleted for repressive chromatin marks H3K9me3 and H3K27me3[28] (**Figure 4.2e**, and see **Figure S4.5c** for other features), consistent with previous studies highlighting the role of clustered TF binding sites in gene regulation[25,29]. Interestingly, five CREs lack any canonical chromatin signatures associated with active *cis*-regulatory sequences (**Figure 4.2a**, Unmarked region, 11%), suggesting existing of elements without canonical epigenetic signatures, as recently reported[12].

To validate the function of the novel *POU5F1* CREs, we selected 6 for in-depth analysis (**Figure 4.1e**, orange bars). The regions were chosen based on three criteria: 1) they are located at a wide range of genomic distances, from 38kb to 694kb, from *POU5F1*

TSS; 2) they are surrounded by phased SNPs so that allelic analysis of gene expression could be performed; and 3) they represent a wide range of CREST-seq signals, ranking 9th, 13th, 23rd, 24th, and 37th out of 45. Additionally, while five CREs, CRE (-694), CRE (-652), CRE (-571), CRE (-449) and CRE (+38), are marked by canonical chromatin marks (**Figure 4.2a**; **Figure S4.6a**), one CRE, CRE (-521), is unmarked (**Figure 4.2a**; **Figure S4.6a**). As a control, we tested a CREST-negative region (**Figure 4.1**; **Figure S4.6a**). We used the CRISPR/Cas9 genome-editing to introduce mono-allelic deletions of lengths 2-4kb to remove these regions in the hESC line (**Figure S4.6a**). As shown in Figure 4.2F, all cell clones with mono-allelic deletion (green curves) on the P1 allele showed significant reduction in *eGFP* expression (**Figure S4.6b**, t-test *P*-value <2.2e-16, error bars, s.d.). By contrast, clones bearing mono-allelic deletions of the P2 allele showed normal *eGFP* expression (**Figure 4.2f**, magenta curves), indicating that these sequences act in *cis* to regulate *POU5F1* expression. No change in *eGFP* expression was observed in clones containing bi-allelic deletions of the negative control region (**Figure 4.2f**, "Ctrl site", solid and dash blue curves). Notably, deletion of CRE (-521), which lacks any canonical marks of regulatory sequences (**Figure S4.6a**), also led to a decrease in *POU5F1* expression in *cis*. Interestingly, while deletion of five CREs resulted in durable reduction of *POU5F1*, deletion of the CRE (-652) element led to only temporary reduction of *eGFP* expression that was fully recovered by day 50 (**Figure 4.2f**; **Figure S4.6b**), suggesting that it belongs to the type of temporarily phenotypic enhancers (TEMP-enhancer) that we recently reported[13]. Taken together, these results provided strong evidence that CREST-seq can be used to identify *cis*-regulatory sequences of a specific target gene in an unbiased and high-throughput manner.

Promoters acting as distal enhancers. Results from the above CREST-seq experiments showed that 18 gene promoters, including the *POU5F1* promoter, are necessary for optimal *POU5F1* expression in hESC. This is surprising because promoters are traditionally thought to mediate transcription of its immediate downstream sequences. Although recent reports indicated that some lncRNA and mRNA promoters may act as enhancers of their adjacent genes[12,30,31], definitive evidence illustrating a causative role of promoters acting as distal enhancers is still lacking. Identification of CRE(-449), CRE(-571) and CRE(-694) as *cis*-regulatory elements of *POU5F1* suggests that promoters of *PRRC2A*, *MSH5* and *NEU1* genes may act as distal enhancers of *POU5F1* in the hESC (**Figure S4.6a**). To rule out the possibility that promoter-proximal elements in these genes were responsible for *POU5F1* regulation, we deleted 216-285bp core promoter sequences containing the TSS of each gene and carried out allelic expression analysis in the resulting cell clones (**Figure 4.3a**; **Figure S4.7**). To avoid potential off-target effects, we used two sets of sgRNA pairs (Deletion 1 and Deletion 2, **Figure 4.3a**; **Figure S4.7**) for the genome editing, and recovered a total of 37 independent clones carrying mono-allelic deletions for in-depth analysis (**Figure S4.8**). We found that all mutants with the P1 mono-allelic deletion displayed long-lasting reduction in *eGFP* expression (green curves in **Figure 4.3a**, **Figure S4.8a** and **Figure S4.8b**; quantified in **Figure S4.8c**, error bars, s.d.), while in mutant clones with the P2 mono-allelic deletion *eGFP* levels were indistinguishable from WT (magenta curves in **Figure 4.3a**, **Figure S4.8a** and **Figure S4.8b**; see **Figure S4.8c** for quantification, error bars, s.d.). The reduced *eGFP* expression could not be due to loss of the *PRRC2A*, *MSH5* or *NEU1* gene products,

because knockdown of each gene using two sets of siRNA (**Figure 4.3b, c**) and shRNAs (**Figure S4.9a-c**) did not affect the *POU5F1* mRNA or protein levels (**Figure 4.3b, c**; **Figure S4.9d**). Thus, the core promoter sequences of *PRRC2A*, *MSH5* and *NEU1,* but not their gene products, are required for optimal *POU5F1* expression.

To further show whether these gene promoters could function as enhancers in a traditional reporter assay, we constructed reporter plasmids that contain the 360-bp *POU5F1* core promoter sequence driving a luciferase reporter gene, with the core promoter fragments of *PRRC2A*, *MSH5* or *NEU1* inserted downstream of the reporter[32]. We transfected these plasmids into the H1 hESC cells and assayed the luciferase activities 3 days after transfection. All elements exhibited significant enhancer activities compared to the control vector (**Figure S4.9e**).

To rule out the possibility that CRISPR/Cas9-mediated genome editing impacts *POU5F1* expression through locus-wide, non-specific mechanisms, we performed FACS analysis of the CRE deletion mutant clones to monitor levels of both POU5F1-eGFP and HLA-C, located 100kb upstream of *POU5F1* TSS. We found that deletion of a CRE resulted in down-regulation of POU5F1-eGFP expression without affecting levels of HLA-C (**Figure S4.10a, b**). To further exclude the possibility that CRISPR/Cas9 leads to double-strand-DNA-break (DSB)- induced transcriptional silencing in the cells, we examined phosphorylated H2AX (γH2AX, a DNA damage marker) in the mutant clones[33–35]. We found that none of the mutant clones stained positive for γH2AX at the time of the experiments (25 days after transfection) (**Figure S4.10a**) when down-regulation of

POU5F1 was detected. Therefore, identification of multiple promoters serving as distal enhancers of *POU5F1* by CREST-seq was unlikely due to artifacts of the experimental system.

The enhancer-like promoters are spatially close to *POU5F1* TSS. To understand potential mechanisms that allow the 17 CREST-positive promoters, among promoters of ~120 genes in this 2MB locus, to specifically regulate *POU5F1*, we examined the 3D chromatin organization of the locus, reasoning that long-range chromatin interactions may allow these enhancer-like promoters to act as distal *cis*-regulatory sequences. Indeed, analysis of H1 hESC Hi-C data[36] indicate that 14 of the 17 *POU5F1*-regulating promoters display significantly higher levels of chromatin interactions with the *POU5F1* TSS than expected by chance (**Figure 4.4a, b**; Wilcoxon tests *P*-value < 0.01). The enhancer-like promoters are also characterized by other chromatin features that distinguish them from other promoters in the region, such as high levels of POL2 binding, H3K4me3, and H3K27ac (**Figure S4.11a, b**; permutation *P*-value < 0.01). In addition, mRNA transcription from these promoters is significantly higher than other genes in the same region (**Figure S4.11c**; Wilcoxon test, *P*-value < 0.01).

To further characterize the features of enhancer-like promoters, we developed a random forest-based classifier capable of predicting which promoters are *cis*-regulatory sequences of *POU5F1.* As input, we used datasets of transcription factor binding sites (TFBS), histone modification[5] profiles, gene expression profiles, and the long-range chromatin contacts centered at *POU5F1*[36]. The performance of the classifier was

evaluated using leave-one-out cross validation. Strikingly, our model can distinguish *POU5F1*-regulating promoters from control promoters in the 2Mbp screen region with high accuracy (**Figure 4.4c**, AUC = 0.89, error rate = 6.3% and PPV=97.2%). We next determined feature importance by estimating the average decrease in node impurity after permuting each predictor variable, finding that the chromatin interaction frequency is the single most important predictor (**Figure 4.4d** and **Figure S4.12**; "Hi-C" for normalized Hi-C interacting frequency). This result provides strong evidence that the enhancer-like promoters specifically affect *POU5F1* expression through chromatin interactions. This observation promoted us to use spatial proximity alone to make a single-variable random forest model, which also achieves high accurate prediction (AUC=0.93, error rate=9.0%) but lower PPV (74.5%), suggesting the physical proximity is an important predictor for predicting regulatory relationship, but other factors are also crucial.

4.4 Discussion

In summary, we have developed a high-throughput method for functional screening of *cis*-regulatory elements in their native genomic context. We demonstrated the utility of this method by applying it to the 2Mbp *POU5F1* gene locus in human ES cells and validated the results by extensive experiments using allelic gene expression analysis.

Our finding that nearly 40% of the *cis*-regulatory sequences of *POU5F1* correspond to promoters of other genes reveals the commonality and widespread use of promoters as distal enhancers. Previous studies have suggested that promoters and enhancers share common properties in terms of transcription factor binding and ability to produce RNA transcripts[37]. Recently, it was shown that the promoters of lncRNAs and mRNAs could act as enhancers of adjacent genes[38]. The current study adds to the accumulating literature that distal promoters can regulate the expression of a gene other than the immediate downstream gene. Our results further showed that one potential mechanism for promoters to act as enhancers is via long-range chromatin interactions. This is consistent with previous studies showing extensive promoter-promoter interactions in mammalian cells[39–46], and reports that many promoters indeed show enhancer activity in heterologous ectopic luciferase reporter assay[30,47].

CREST-seq is a highly scalable tool for unbiased discovery of *cis*-regulatory sequences in the human genome. Compared to the previous CRISPR/Cas9 screens, which typically require more than 100 gRNAs-expressing oligos to “saturate” a targeted region, CREST-seq achieved 20x coverage for the entire 2Mbp POU5F1 locus with less

than six sgRNAs per kilobase. CREST-seq also outperforms the dCas9-KRAB based CRISPRi screen[15] in which the size of H3K9me3 peaks generated by dCas9-KRAB is less than 850bp[48]. Although the size of positive hits identified by CREST-seq are usually larger than the size of element/motif identified by single sgRNA approach, by generating overlapping deletions in a massively parallel fashion, CREST-seq allows functional interrogation of a large fraction of the genome with high sensitivity and specificity. More importantly, CREST-seq can distinguish *cis*- and *trans*-regulatory sequences by monitoring the allelic expression of a reporter gene, without the knowledge of haplotypes of the genome. Finally, it is feasible to design nested tiling deletions across a whole chromosome or even the genome. Combination of CREST-seq and single sgRNA screen approaches would allow us to achieve both high coverage and high resolution, thereby enabling truly comprehensive discovery of transcriptional regulatory sequences in the human genome.

4.5 Acknowledgements

We thank D. Gorkin and J. Yan for feedback on previous versions of the manuscript. We thank Z. Ye and S. Kuan for technical assistance. This work was supported by the US National Institutes of Health (NIH) (grants U54 HG006997, U01 DK105541, R01HG008135, 1UM1HG009402 and 2P50 GM085764 to B.R.), the Ludwig Institute for Cancer Research (to B.R.) and the Human Frontier Science Program (HFSP) (Long Term Postdoctoral Fellowship to Y.D.).

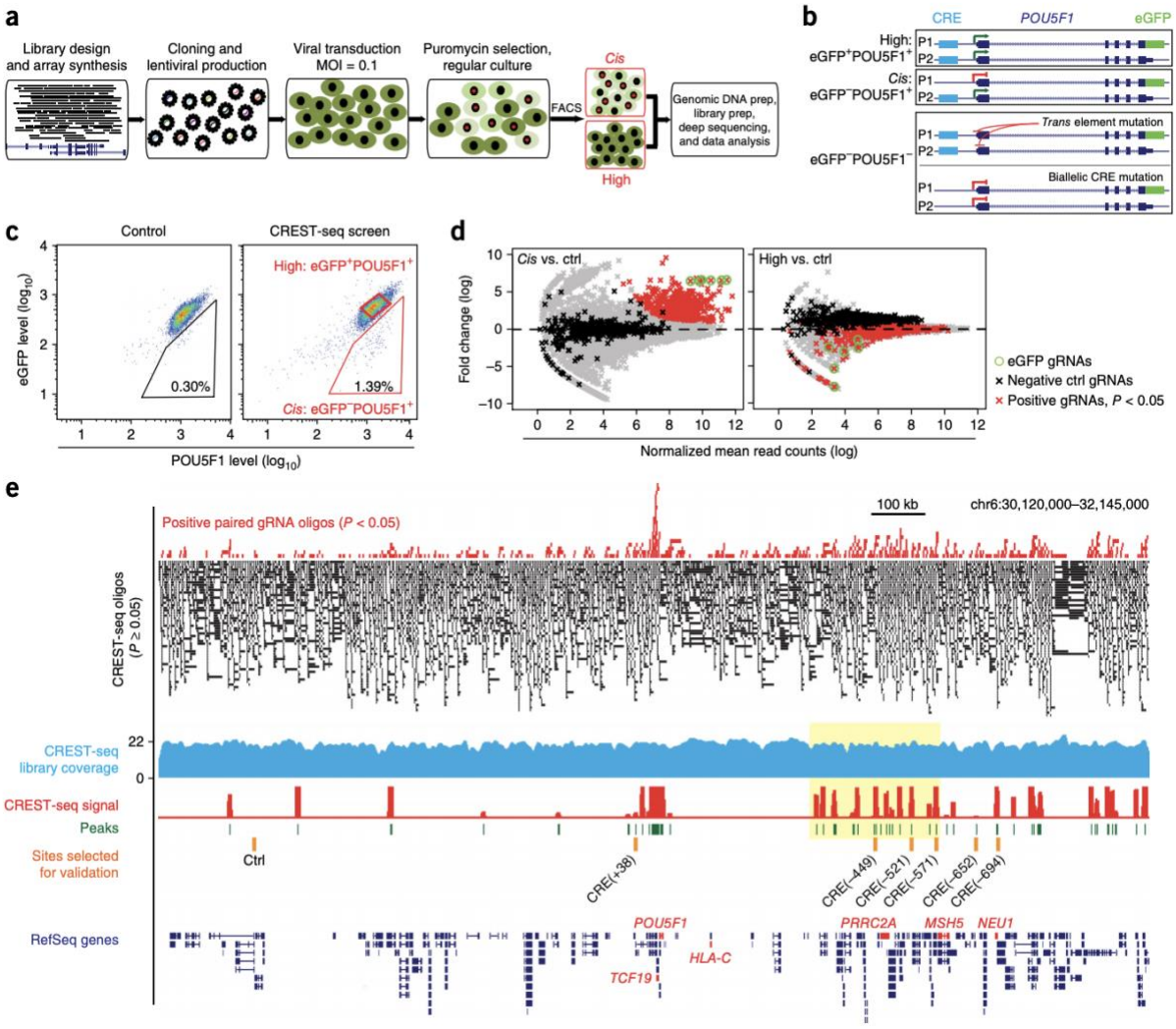
Chapter 4, in full, is a reprint of the material as it appears in Nature Methods 2017 “A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells”. Yarui Diao, Rongxin Fang, Bin Li, Zhipeng Meng, Juntao Yu, Yunjiang Qiu, Kimberly C Lin, Hui Huang, Tristin Liu, Ryan J Marina, Inkyung Jung, Yin Shen, Kun-Liang Guan & Bing Ren. The dissertation author was the primary investigator and author of this paper.

4.6 Author Contributions

Y.D. and B.R. conceived the idea for CREST-seq; R.F., Y.D. and B.L. conducted integrative data analysis with help from Y.Q., H.H. and I.J.; B.L. and Y.D. designed paired sgRNA libraries; Y.D., Z.M., J.Y., K.C.L., T.L., H.H., R.J.M. and Y.S. performed the experiment; Z.M., K.C.L. and K.-L.G. packaged the lentiviral library; and Y.D., R.F., B.L. and B.R. wrote the paper.

4.7 Figures

Figure 4.1. CREST-seq experimental design and application to the POU5F1 locus in hESC. (a) Workflow of CREST-seq. A total of 11,570 oliogs containing dual sgRNA sequences were cloned into a lentiviral library that was in turn transduced into the H1 POU5F1-eGFP cells with MOI=0.1. After Puromycin selection, the cells were stained with antibodies specifically recognizing POU5F1 (PE) or eGFP (APC), respectively. The indicated "Cis" and "High" populations were sorted by FACS, and the integrated sgRNA pairs were amplified by PCR from genomic DNA followed by high-throughput sequencing. (**b**) Schematic illustration of mono-allelic or bi-allelic deletions of *cis*-regulatory elements of *POU5F1*. The eGFP-tagging allele is designated as P1 and the wild-type allele as P2. Mono-allelic disruption of a *POU5F1* CRE on the P1 allele would lead to reduced *eGFP* expression while POU5F1 protein levels remain relatively unchanged (eGFP-/POU5F1+). Bi-allelic disruption of a *POU5F1* CRE would lead to reduction of both eGFP and POU5F1 protein level. **(c)** FACS analysis of H1 POU5F1-eGFP cells transduced with control lentivirus expressing Cas9 but not sgRNA (left) or the CREST-seq lentiviral library (right) 14-day post transduction. **(d)** The read counts of sgRNA from "Cis" (left) and "High" (right) are compared to those from a non-sorted control population (Ctrl). The fold changes represent the ratios between read counts in the "Cis" or "High" populations and the "Ctrl" population, with the significance of enrichment calculated by a negative binomial test. Green circles denote eGFP targeting sgRNA pairs; Red dots correspond to sgRNA pairs enriched in the "Cis" population with *P*-value < 0.05 and log(fold change) > 1. Black dots denote the negative control sgRNA pairs and grey dots for the rest of pairs. **(e)** Genome browser screenshot showing CREST-seq positive sgRNA pairs (*P*-value < 0.05, top) and CREST-seq negative sgRNA pairs (*P*-value>0.05, black bars); genomic coverage of the CREST-seq library (blue track); the computed CREST-seq signals (see Methods), and the genomic regions identified as *cis*-regulatory sequences of *POU5F1* (peaks, green), along with the CRE sites selected for further in-depth validation (orange bars). Yellow box highlighted a region enriched for CREs with a close-up view in **Figure 4.2b**.


a
Library design and array synthesis
Cloning and lentiviral production
Viral transduction MOI = 0.1
Puromycin selection, regular culture
FACS
Cis
High
Genomic DNA prep, library prep, deep sequencing, and data analysis
b
CRE
POU5F1
eGFP
High: eGFP+POU5F1+
Cis: eGFP−POU5F1+
eGFP−POU5F1−
P1
P2
Trans element mutation
Biallelic CRE mutation
c
Control
CREST-seq screen
High: eGFP+POU5F1+
Cis: eGFP−POU5F1+
0.30%
1.39%
eGFP level (log10)
POU5F1 level (log10)
d
Cis vs. ctrl
High vs. ctrl
Fold change (log)
Normalized mean read counts (log)
eGFP gRNAs
Negative ctrl gRNAs
Positive gRNAs, P < 0.05
e
100 kb
chr6:30,120,000–32,145,000
Positive paired gRNA oligos (P < 0.05)
CREST-seq oligos (P ≥ 0.05)
CREST-seq library coverage
CREST-seq signal
Peaks
Sites selected for validation
Ctrl
CRE(+38)
CRE(−449)
CRE(−521)
CRE(−571)
CRE(−652)
CRE(−694)
RefSeq genes
POU5F1
TCF19
HLA-C
PRRC2A
MSH5
NEU1

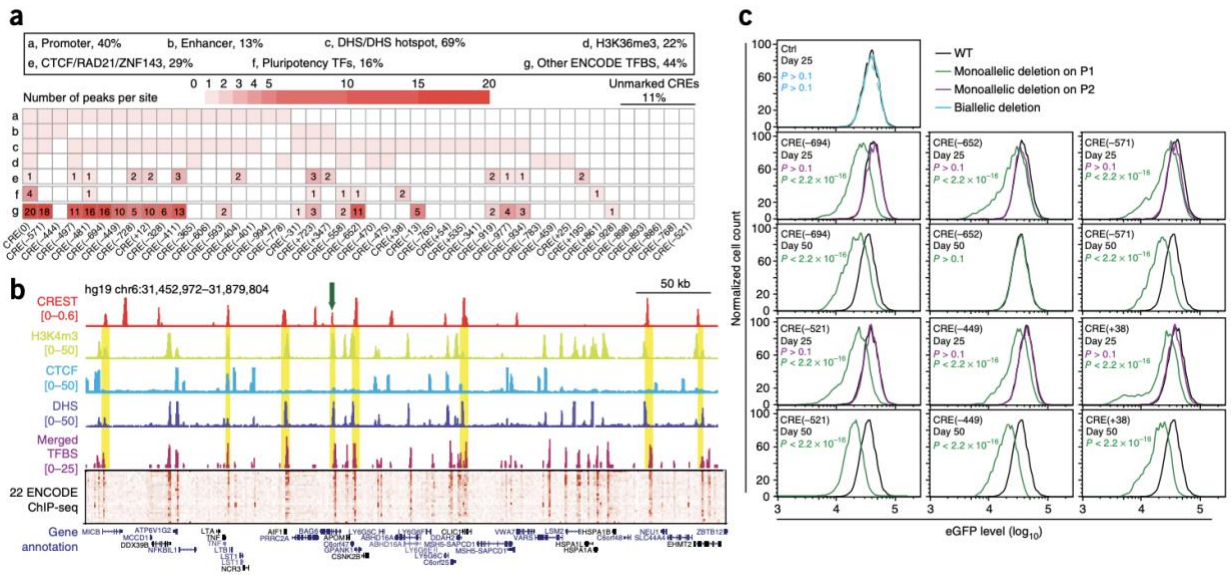


Figure 4.2. CREs tend to be associated with canonical active chromatin markers of cis-regulatory elements and dense TF clusters. (a) A matrix showing the chromatin features and transcription factor binding at the 45 CREs. "Pluripotency TFs" denotes POU5F1, SOX2, NANOG, and PRDM14. **(b)** A close-up view of genome browser snapshot of the yellow highlighted region in **Figure 4.1e** with tracks corresponding to chromatin modifications, DHS, merged TFBS ChIP-seq peaks and a heatmap of normalized ChIP-seq signals for 22 transcription factors in hESCs. The height of merged TFBS bars indicates the number of bound TF. Yellow bars highlighted regions where CREs overlap with active chromatin marks and TFBS clusters. The green arrow points to the CREs in **(c). (c)** A close-up view of a 5kb CRE occupied by a cluster of TFs. **(d)** A box plot shows that transcription factor binding sites more frequently cluster at CREs than at typical *cis*-regulatory elements represented by DHS. (Wilcoxon test *P*-value < 6e-11). **(e)** A bar chart shows the degree of enrichment of each chromatin feature in the CREs. To calculate the "Enrichment Test Score", we first calculated the fraction of CREST-seq peaks that intersected with sites associated with each feature as a ratio between the observed over expected. An average ratio is calculated from 1,000 random permutations of the CREs. The enrichment test score is defined as the percentage that observed ratio is greater than expected. (*χ^2 *P*-value < 0.01). **(f)** Six CREs and one CREST-seq negative site (Ctrl) were selected (orange bars in **Figure 4.1e**) for individual validation. Mutant clones were generated harboring bi-allelic deletion (Ctrl, blue curves), mono-allelic deletion on the P1 allele (green curves), or mono-allelic deletion on the P2 allele (magenta curves) at the indicated genomic loci. P1 is the eGFP-containing allele and P2 is the non-eGFP allele. FACS analysis was performed for all the mutant clones and wide-type cells (WT: black curves) at day 25 and day 50 after CRISPR/Cas9 transfection. The FACS data was quantified with FlowJo and P-value is calculated with two-sample t-test. "*p*" in Green and magenta letter "*p*" represent the *P*-values for mono-allelic mutants harboring P1-specific or P2-specific deletion, respectively.

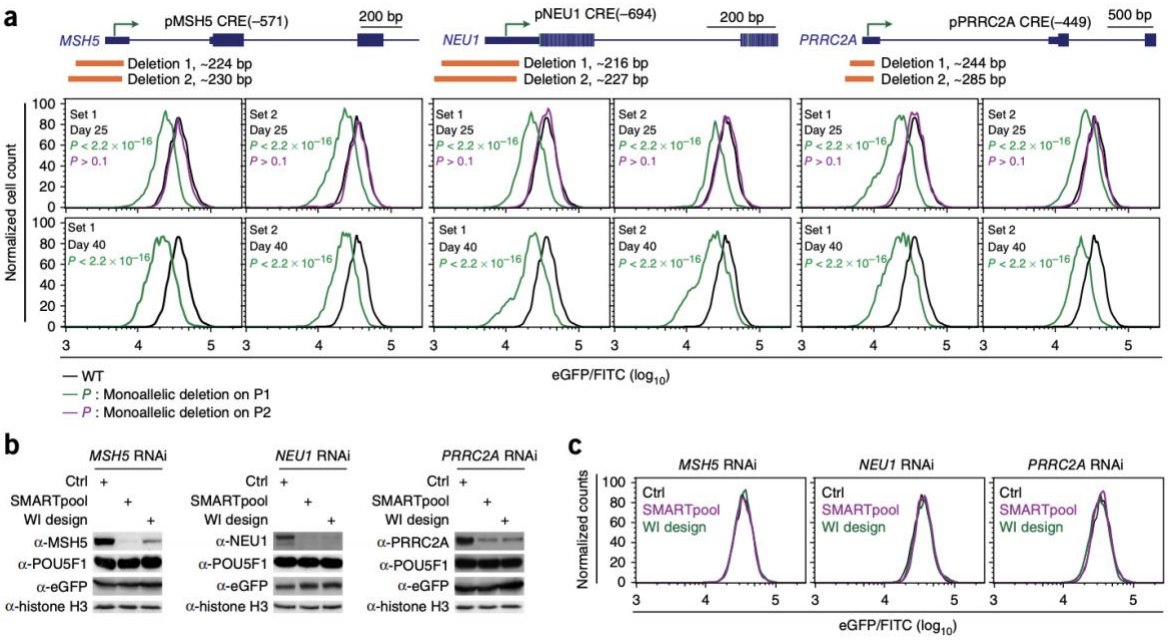


Figure 4.3. The core promoter regions of MSH5, NEU1, and PRRC2A are required for optimal POU5F1 expression in hESC. (**a**) The core promoter regions of MSH5, NEU1, and PRRC2A were deleted by two sets of distinct sgRNAs (orange bars, Deletion 1 and 2). Mutant cell clones harboring mono-allelic deletions on the P1 allele (green curves), or P2 allele (magenta curves) were identified after genotyping and sequencing of the phased SNPs. FACS analysis was performed for all the mutant clones and wild-type cells (WT: black curves) at day 25 and day 40 after transfection. The FACS data is quantified with FlowJo. P-value is computed using two-sample t-test. (**b, c**) The H1 POU5F1-eGFP cells were transfected with either control scrambled siRNA or siRNAs targeting the gene as indicated. Each gene is targeted by two sets of siRNAs (SMARTpool and WI design) with different sequences. The cells were analyzed 48 hours after transfection. (**b**) Whole cell extract was collected and subjected to western blot analysis with indicated antibodies. (**c**) An aliquot of cells was dissociated into single cells for FACS analysis. Black, magenta, and green curves represent the data from cells treated with Scrambled siRNA (Ctrl), SMARTpool siRNA and WI (http://sirna.wi.mit.edu/) designed siRNA, respectively.

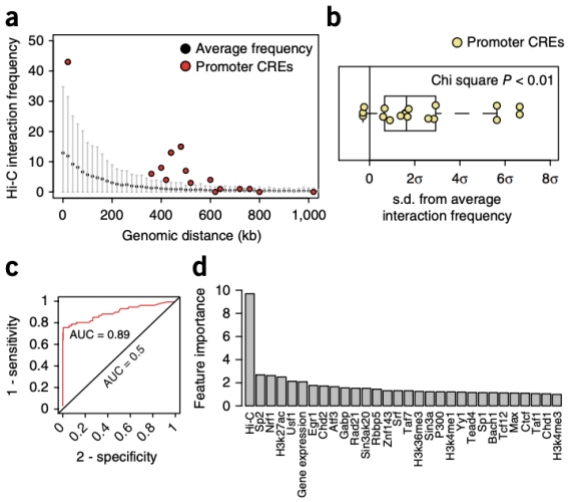


Figure 4.4. Analysis of chromatin interactions between the enhancer-like promoters and POU5F1 promoter in hESC. (**a**) A dot plot shows the distribution of pairwise Hi-C contact frequencies within the 2Mbp locus, and between the *POU5F1* TSS and the 17 *POU5F1*-regulating promoters (red dots, promoter-CREs). The black dots and the gray bar represent the average and standard deviation of Hi-C read counts at a given genomic distance, respectively. (**b**) A boxplot shows the number of standard deviations of the Hi-C read counts between *POU5F1* TSS and the promoter-CREs (yellow dots) compared to the expected (0, black line) (χ^2 *P*-value < 0.01). (**c**) ROC curve shows that *POU5F1*-regulating promoters can be separated from the other promoters in the 2Mbp region with a high accuracy (AUC=0.89) using a random forest model built from binding sites of 52 TFs, seven histone modifications profiles, gene expression profile and maps of long-range chromatin interactions (see **Supplementary Methods** for more details). (**d**) A bar chart shows the relative importance of each feature to the Random Forest classifier in predicting enhancer-like promoters.

4.8 Supplementary Methods

CREST-seq protocol. A detailed protocol of CREST-seq has been deposited here[49].

Cell culture. The POU5F1-eGFP H1 hESC line was purchased from WiCell (Log number: DL-02) and described previously[19]. The cells were cultured on Matrigel-coated (Corning, Cat #354277) plates and maintained in TeSR-E8 media (STEMCELL Technologies, Cat#05940), and passaged by Accutase (STEMCELL Technologies, Cat#A1517001) with 10uM ROCK inhibitor Y-27632 (STEMCELL Technologies, Cat# 72302) supplement. The cells have been tested by WiCell Research Institute and UCSD human Stem Cell Core facility to confirm no mycoplasma contamination.

Design of sgRNA pairs for CREST-Seq. CREST-seq library design is available online (http://crest-seq.ucsd.edu/web/) and includes the following steps: 1) all 20-bp potential sgRNA sequences followed by PAM motif 'NGG' within the 2-MB screened region were first identified; 2) Bowtie[50] was used to map these 20-bp sgRNA sequences to the reference genome (hg19) with following parameter '-t -a -f -m 1000 --tryhard -v 3' which outputs alignments up to 1000 candidates with less than 4 mismatches; 3) In order to prevent off-target binding, a sgRNA sequence was filtered out if it a) perfectly maps to another region on the genome; or b) has suboptimal alignment with 1 or 2 mismatched bases outside the sgRNA "seed" region, i.e. the 10bp sequence adjacent to PAM motif[51]; or d) has suboptimal alignment with 3 mismatches but all three mismatched bases are 17-bp further to the PAM sequence; 4) the identified sgRNA sites were paired in order to

generate 2kb-deletions evenly across the 2 Mbp-region. Based on the distribution of the filtered sgRNA, a chain of unique single guide RNAs were selected as follows: First, the initial sgRNA was picked, and the next sgRNA was chosen based on a pre-determined distance cutoff (D, for example 100bp) and an odd number of step size (S, for example 15) such that the distance between the target sequences of the two sgRNAs is no less than D; the procedure was repeated until no more unique sgRNA was found. Next, the first sgRNA pair was designed using the 1st sgRNA and the 16th (1+S) sgRNA, then the second pair using 3rd and 18th (3+S), the procedure was repeated to the end of the chain. The distance cutoff D and step S were both adjustable to allow for different deletion sizes and genomic coverage. For example, using D=100, and S=15, the deletion size would be a minimum of 1,500 bp, an average of 2,000 bp in the current design. The average coverage was (1+S)/2, 8 times with S=15, since there were 8 sgRNAs (relatively 1st, 3rd, … 15th) crossover to 8 guide RNAs on other side (relatively 16th, 18th, … 30th) for any region in the middle. Three different sets of deletion/steps were used: 100/15, 200/13, 500/13. An unique guide RNA was not used if it has been used in previous selection. After a pair of dual CRISPR guide RNAs, namely {a, b}, were selected, we used the following template to link two guide RNAs:

TGTGGAAAGGACGAAACACC{a}GTTTAGAGACG{rnd}CGTCTCACCTT{b}GTTTTAG

AGCTAGAAATAGCAAGTT, note that if a guide RNA start with A, C, or T, a G was added in front. The ${rnd} was selected from all combinations of 9-bp nucleotide sequence excluding either number of GC less than 4 or more than 6, or include any subsequence within: {"AAAA", "CCCC", "TTTT", "GGGG", "GAGACG", or "CGTCTC"}.

Oligo synthesis and library cloning. The CREST-seq oligo library with sequences shown in **Figure S4.2a** was amplified with the following primers:

Forward primer: CTTGTGGAAAGGACGAAAC

Reverse primer: TTTTAACTTGCTATTTCTAGCTCTAAAAC

The PCR product was size selected and gel-purified with NucleoSpin Gel and PCR Clean-Up Kit (Clontech, Cat# 740609), and then inserted into BsmbI digested lentiCRISPRv2 plasmid by Gilbson Assembly (Addgene plasmid #52961). The end product was electro-transformed into 5-alpha Electrocompetent E. coli (NEB, Cat#C2989K) and grown on Agar plates. About 20 million independent bacterial colonies were collected and the plasmids were extracted with QIAGEN Plasmid Giga Kit (Cat#12191). The resulting plasmid DNA was linearized by BsmbI digestion, gel purified and ligated with a DNA fragment (see complete IDT gBlocks sequence) containing tracRNA(E/F) and the mouse U6 promoter (mU6). The ligates was electro-transformed into 5-alpha Electrocompetent E. coli and plated on Agar plates. About 20 million bacterial colonies were collected and purified with EndoFree Plasmid Giga Kit (QIAGEN, Cat#12391)

Lentiviral library production. The CREST-seq lentiviral library was prepared as previously described[52] with minor modifications. Briefly, 5ug of lentiCRISPR plasmid library was co-transfected with 4 ug PsPAX2 and 1 ug pMD2.G (Addgene #12260 and #12259) into a 10-cm dish of HEK293T cells in DMEM (Life Technologies) containing 10% FBS (Life Technologies) by PolyJet transfection reagents (Signagen, Cat# SL100688). Growth medium was replaced 6 hours after transfection. The supernatant of cell culture media was harvested at 24 hours and 48 hours after transfection and filtered

by Millex-HV 0.45 μm PVDF filters (Millipore, Cat# SLHV033RS). The viruses were further concentrated with 100, 000 NMWL Amicon Ultra-15 Centrifugal Filter Units (Amicon, Cat#UFC910008).

For viral titration, 0.5 million hESC POU5F1-eGFP cells were seeded per well on 6-well plate. 12 hours later, different amount (1ul, 2ul, 4ul, 8ul) of concentrated viral-containing media were added to the cell culture media to infect the hESC following the same protocol described in the lentiviral screening section. The same amount of non-infected cells was seeded and not treated with puromycin as the control. 24 hours post-infection, the viral infected cells were treated with 500ng/ml Puromycin (Life Technologies, Cat#A1113802) for another 72 hours. We counted the number of Puromycin resistant cells and the control cells to calculate the ration of infected cells, and then viral titer. In the screening, about 10 million POU5F1-eGFP hESCs were used in each independent screening replicate and infected with viral particles at low MOI (0.1) to make sure each infected cell gets one viral particle.

Lentiviral transduction and FACS. Briefly, the screening was performed following previous protocol described earlier[13] with minor modifications. In each independent screen, about 10 million cells per 12-well plates were spin infected with CREST-seq lentiviral library at MOI=0.1. 24 hours post infection, the cells were dissociated with Accutase, and plated into l5cm culture dish coated with Matrigel (4 million cells per dish). The cells were treated with E8 media containing 250ng/ml Puromycin for 7 days, followed by another 7-day culture without Puromycin treatment. For CREST-seq

screen FACS sort, the cells were dissociated and co-immunostained with PE-POU5F1 antibody and APC-eGFP antibody. The eGFP-/POU5F1+, eGFP+/POU5F1+, and non-sorted control cells were collected by FACS sort for further analysis.

Sequencing library construction. Genomic DNA was extracted from the eGFP-/POU5F1+, eGFP+/POU5F1+ or the non-sorted control cells populations. The sgRNAs inserts were then amplified from genomic DNA PCR using the following primers:

Forward: AATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCG

Reverse: GGACTGTGGGCGATGTGCGCTCTG

The PCR products were gel purified and subjected to the 2nd PCR reaction to add Illumina TruSeq adaptor sequence with the following primers:

Forward:

AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTctTGTGGAAAGGACGAAAC

Reverse (N indicate the index sequence):

CAAGCAGAAGACGGCATACGAGANNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTTTTAACTTGCTATTTCTAGCTCTAAAAC

Sequencing and processing of CREST-seq libraries. CREST-seq libraries were sequenced using HiSeq 4000 in pair-ended mode with 100bp read length. A sgRNA pair {a, b} was considered valid if it matched the initial sgRNA design and met the following criteria: (1) a subsequence of the read1 matched GGACGAAACACCG, followed by 19 or 20 nucleotides (namely, {a'}), and GTTTAAGAGCTATGCTG, (2) a

subsequence of read2 matched AAAC, followed by 19 or 20 nucleotides (namely, {b'}), and followed by CAA; (3) {a} exactly matched {a'} if length of {a'} was 20, or {a} exactly matched G+{a'} if length of {a'} was 19; (4) {b} exactly matched reverse complementary of {b'} if length of {b'} was 20, or {b} exactly matched G+reverse complementary {b'} if length of {b'} was 19. Those sgRNA pairs with total read count less than 30 among all samples were filtered out. In the end, we kept 10,159 sgRNA pairs for further analysis.

Peak calling in CREST-seq data. For each sgRNA pair, the MAGeCK algorithm [20] was used to estimate the statistical significance (using Negative Binomial test) of enrichment in the cell population relative to the control population. Next, sgRNAs pairs were ranked by $log(NB\ P-value) \times sign(log(exp/control))$ in an increasing order. Third, we partitioned the 2-MB screened region into a set of non-overlapping 50-bp bins $B = (b_1, \ldots, b_n)$, and a bin was considered positive if many of the sgRNA pairs spanning it rank near the top of the sorted list. A Robust Rank Aggregation (RRA) algorithm[53] was then used to identify the positive bins. Specifically, let $R_i = (r_{i1}, \ldots, r_{ik})$, be the vector of ranks of sgRNA pairs that span bin b_i, we normalized R_i into percentiles $U_i = (u_{i1}, \ldots, u_{ik})$ where $u_{ij} = r_{ij}/M$ (M is the total number of sgRNA pairs). The goal was to identify the bins whose normalized rank vector U_i is strongly skewed toward zero. Under null hypothesis where the normalized ranks follow a uniform distribution between 0 and 1, the j-th smallest value among (u_{i1},...,u_{ik}) is an order statistics $\rho(u_{ij})$ which can be calculated by a beta distribution $Beta(j, k+1-j)$. We defined the final score for the rank vector U_i as the minimum of negative score:

$$\rho(U_i) = \min_{j=1 \to k} \rho(u_{ij})$$

$\rho(U_i)$ score was converted to *P*-value by permutation test as proposed by Li et al[20] and finally P-value was finally adjusted to FDR. A bin was considered as significant if its FDR was smaller than a given threshold.

Calculation of Enrichment Test Score. We downloaded DNase Hypersensitive Sites (DHSs) and peaks of ChIP-seq datasets from H1 hESC from ENCODE data portal[5]. Enhancers were predicted using RFECS[54], and promoter coordinates were based on RefSeq gene annotation. The observed overlap ratio o_i of feature i was computed as the fraction of CREST-seq peaks that overlapped with this feature. We then randomly shuffled CREST-seq peaks in this region using 'shuffleBed'[55], and the expected overlap rate e_i was counted as the fraction of shuffled peaks that overlapped with feature i. Fold enrichment was computed as o_i/e_i. We repeated this process 1000 times for each feature and defined the enrichment test score as the fraction of tests where the fold enrichment was greater than 1. The significance of enrichment was derived using the χ^2 test.

Analysis of chromatin signatures of *POU5F1*-regulating promoters. We randomly shuffled CREST-seq peaks in the 2Mbp *POU5F1* region using 'shuffleBed'[55] and only kept those permutations with 18 peaks overlapping promoter regions. The expected overlap rate for each shuffle was counted as the fraction of permutations that contain active promoter signature (Pol2/H3k4m3/H3k27ac). We repeated this process 1000 times and calculated permutation *P*-value as the percentage of tests in which the overlap rate is above 0.78.

Classification of *POU5F1*-regulating promoters by Random Forest. We downloaded RefSeq annotated promoters (2,000bp upstream TSS) from UCSC genome browser within the screened region. Promoters were divided into positive and control groups based on their overlap with CREs. RNA-seq data was downloaded from previously work and gene expression was estimated using software Cufflinks for each transcript. Random forest implemented by R package "randomForest" was applied to classify positive promoters from the negative ones with default parameter setting without further model selection. Prediction performance was evaluated by leave-one-out cross validation. Feature importance was estimated by the average decrease of node purity by permuting each variable.

CRISPR/Cas9-mediated deletion. CRISPR/Cas9 constructs targeting genomic loci indicated on **Figure S4.6a** was made following the protocol described earlier[13]. The designed sgRNAs sequence was cloned into the pX330-U6-Chimeric_BB-CBh-hSpCas9 (Addgene plasmid #42230) vector. After validating the sgRNA sequences by Sanger sequencing, a pair of plasmids targeting 5'- and 3'- boundary of the same element, were mixed at 1:1 ratio and co-transfected with plasmid expressing mCherry into POU5F1-eGFP cells by hESCs Nuclearfector Kits 2 (Lonzo, Cat#VPH-5022) according to the manufacture's instruction. To knockout POU5F1-regulatory core promoters, we used in vitro synthesized CRISPR crRNA and CRISPR tracrRNA (IDT). The Cas9 recombinant protein was purchased from NEB (Cat M0386M) and the Cas9/crRNA/tracRNA was assembled *in vitro* by following a protocol[56]. The RNP complex was electro-transfeced

into POU5F1-eGFP hESC reporter line with Neon Transfection System 10µl kit (ThermoFisher Scientific, Cat#: MPK1096) with the default electrotranfection protocol #9.

After 72 hours post-transfection, the mCherry positive cells were collected by FACS. The mCherry positive single cells were plated into Matrigel-coated plate at low density (about 1000 cells per 10 cm coated petri-dish) and cultured in E8 media supplemented with 10uM ROCK inhibitor. After 10 to 14 days, the surviving sorted single cells formed colonies. Individual colonies were picked and expanded, followed by genotyping and in-depth analysis.

Genotyping of mutant clones. The cells from mutant clones were collected and treated with QuickExtract™ DNA Extraction Solution (Epicentre, Cat# QE0905T), followed by genotyping PCR. Then Topo cloning (Life Technologies, Cat#K2800-20) and Sanger sequencing were conducted to verify the sequences.

FACS analysis. To directly monitor the eGFP expression levels, the wild type or mutant POU5F1-eGFP cells were dissociated with Accutase and subjected to FACS analysis with BD FACSAria II. To examine the levels of HLA-C protein, the cells were stained with PE-conjugated antibody specifically recognizing HLA-C (Millpore, Cat#MABF233). To carry out immunostaining of eGFP, POU5F1, or H2AX, the cells were fixed with 2% PFA for 30 minutes, followed by overnight permeabilization in Methanol at -20°C. The treated cells were stained with the antibodies. PerCP-cy5.5-conjugated mouse anti-H2AX(pS139) was purchased BD Biosciences (Cat#564718); PE-conjugated

anti-human OCT4(OCT3) antibody was from STEMCELL Technologies (Cat# 60093PE.1) and APC-conjugated anti-GFPuv/eGFP antibody is available from R&D Systems (Cat# IC4240A)

Luciferase reporter assays. Luciferase assays were conducted as previously described[57]. Briefly, to test the enhancer activity of CREs with native *POU5F1* promoter, the 360bp *POU5F1* minimal promoter[32] (hg18 Chr 6: 31,246,377-31,246,736) was synthesized as gblock by IDT, and cloned into pGL3-promoter vector to replace the original SV-40 promoter. The core promoter regions of pPRRC2A, pMSH5, pNEU1 and pTFC19 were PCR amplified from H1 hESC genomic DNA and cloned into a modified pGL3-POU5F1 vector (Promega), in which the SV40 promoter has been replaced by a 360bp minimal POU5F1 promoter by In-fusion cloning. After validation by Sanger sequencing, the constructs were co-transfected with pRL-SV40 Renilla reporter vector in H1 hESCs with Fugene HD (Roche) at a 4:1 reagent to DNA ratio. The transfected cells were cultured for an additional 2 days prior to harvest for reporter assay. The Dual-Luciferase Reporter Assay kit (Promega Cat#:E1960) was used according to manufacturer's protocol. The adjusted firefly luciferase activity of each sample was normalized to the average of activities of 3 negative control regions.

RNA interference. The siRNAs were purchased from Dharmacon in the format of ON-TARGETplusSMARTpool-Human targeting MSH5, NEU1 and PRRC2A, respectively. We also designed siRNAs by using WI siRNA selection program. The

siRNAs were transfected into hESC with Human Stem Cell Nucleofector Kit 2 (LONZA) per manufacturer's instruction.

Western blotting. Western blotting was performed by following the protocol described previously[58]. Briefly, whole cell extracts (WCE) were collected and quantified with Pierce™ BCA Protein Assay Kit (Cat#23225). 30μg WCE of each sample was subjected to Western blot analysis with antibodies specifically recognizing NEU1(Thermo Scientific, Cat#PA5-42552), PRRC2A (Abcam, Cat#ab188301), MSH5 (Abcam, Cat#ab130484), Histone-H3(Abcam, Cat#ab1791), POU5F1 (Abcam, Cat#ab19875), and eGFP (Abcam, Cat#ab190584).

ATAC-seq experiment and analysis. ATAC-seq was performed by following the protocol described earlier [22]. Briefly, each library starts with 100k cells which were permeabilized with NPB (0.2% NP-40, 5%BSA, 1Mm DTT in PBS with one complete proteinase inhibitor) at 4 degree for 10min, followed by spin down at 500g for 5min. The resulting nuclei were resuspended in 20ul 1xDMF (33mM Tris-acetate (pH=7.8), 166mM K-Acetate, 10mM Mg-Acetate, 16 % DMF). The chromatin tagmentation was done by adding 0.5ul Tn5 into 10ul solution for 30min at 37 degrees.

We processed our ATAC-seq data in the following steps: 1) ATAC-seq sequencing reads were mapped to hg19 reference genome using Bowtie(*61*) in pair-end mode; 2) poorly mapped, improperly paired and mitochondrial reads were filtered; 3) PCR duplications were further removed using Picards MarkDuplicates

(http://broadinstitute.github.io/picard.); 4) Mapping positions of reads were adjusted accounting for Tn5 insertion; 6) Reads were next shifted for 75bp followed by peak calling using MACS2[59] with following parameters "-q 0.01 --nomodel --shift 175 –B --SPMR --keep-dup all --call-summits"; 7) ATAC-seq signal was normalized into RPKM using deeptools[60] for visualization.

PCA analysis. We first extracted all 478 H1 DHS sites within the screened regions and counted the average RPKM for each site using 122 public DHS datasets and our in-house ATAC-seq dataset. Pair-wise Pearson correlation between the datasets were calculated and used as input for PCA analysis. We found the first two principle components accounted for 80% of the variance and therefore used for 2D visualization as shown in **Figure S4.5b**.

4.9 Supplementary Figures

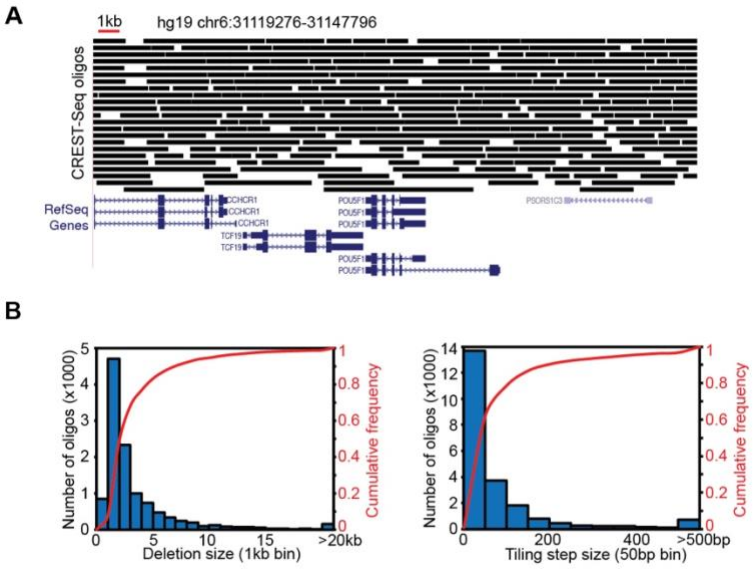


Figure S4.1. Design of sgRNA pairs. (**a**) A genome browser screenshot illustrating the representative tiling design of CREST-seq sgRNA pairs in the POU5F1 locus. Each black bar represents a sequence targeted by a pair of sgRNAs. (**b**) Distribution of the sizes of deletions (top panel) and step sizes of two adjacent deletions (bottom panel).

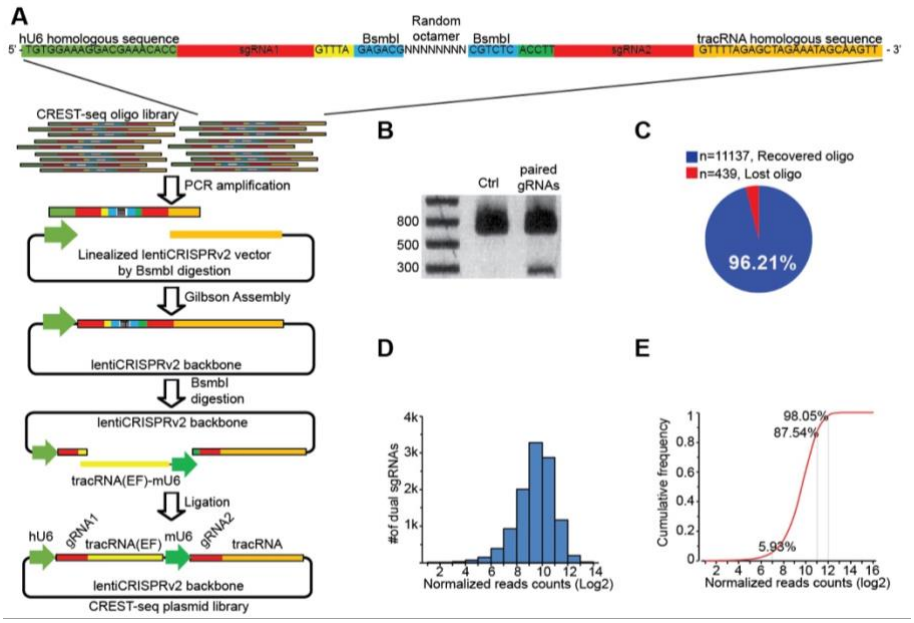


Figure S4.2 CREST-seq library construction and quality control. (**a**) (Top) Schematic illustration of the oligonculeotides containing the pairs of sgRNAs flanked by a common adaptor sequence required for two-step library cloning. (Bottom) Workflow of the two-step plasmid library cloning. The oligo library was synthesized by Custom Array (Seattle, WA), PCR amplified, and cloned into lentiCRISPRv2 backbone via Gibson Assembly. The first step cloning product were then digested by BsmBI and ligated with a DNA fragment containing tracRNA(EF) and mouse U6 promoter (mU6) sequence. tracRNA(EF): tracRNA with extended stem-loops and flipped A/T bases 65. (**b**) Lentiviral particles were packaged as described previously 13 and transduced into H1 hESC via spin infection. 36 hours after viral transduction, the cells were cultured in E8 media containing Puromycin for 72 hours, and in regular media for another 3 days. Genomic DNA was purified for genotyping PCR. The PCR products with smaller sizes indicate the genomic deletion at the target region. (**c-e**) After a two-step cloning procedure and plasmid DNA prep, the dual-sgRNA inserts were amplified from the final CREST-seq plasmid library and subjected to deep sequencing. The paired-end reads were mapped to CREST-seq oligo design file. (**c**) The plasmid library recovered 96.21% of oligos in the CREST-seq library design. (**d, e**) Distribution of CREST-seq oligo read counts (**d**) and cumulative frequency in the plasmid DNA library (**e**).

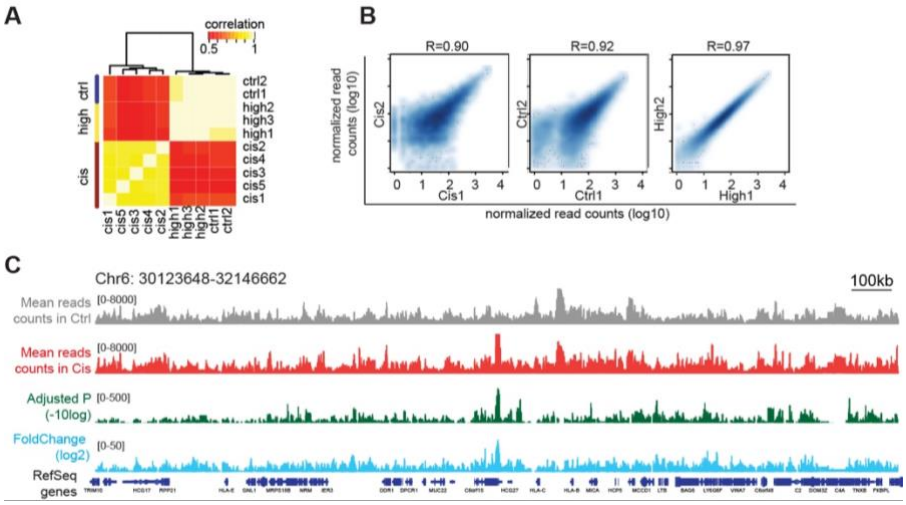


Figure S4.3. Quality control of CREST-seq data from replicates. Genomic DNA isolated from “Cis”, “High” and “Ctrl” cell populations was subjected to PCR amplification and then deep sequencing. (**a**) Unsupervised clustering analysis shows correlation of biological replicates of five “Cis” (cis 1-5), three “high” (high 1-3) and two control (ctrl1, ctrl2) samples. (**b**) Scatter plots show that sgRNA read counts correlate well between replicates. (**c**) Genome browser screenshot showing the gene annotation in the 2Mbp POU5F1 locus (RefSeq genes), mean reads counts in control samples (“Ctrl”) and in Cis samples (“Cis”), -10log (Adjusted P-value) (green tracks) and log2(Fold change) (blue) of sgRNA pairs. We used edgeR to identify significantly enriched oligos (see **Supplementary Methods** for more details).

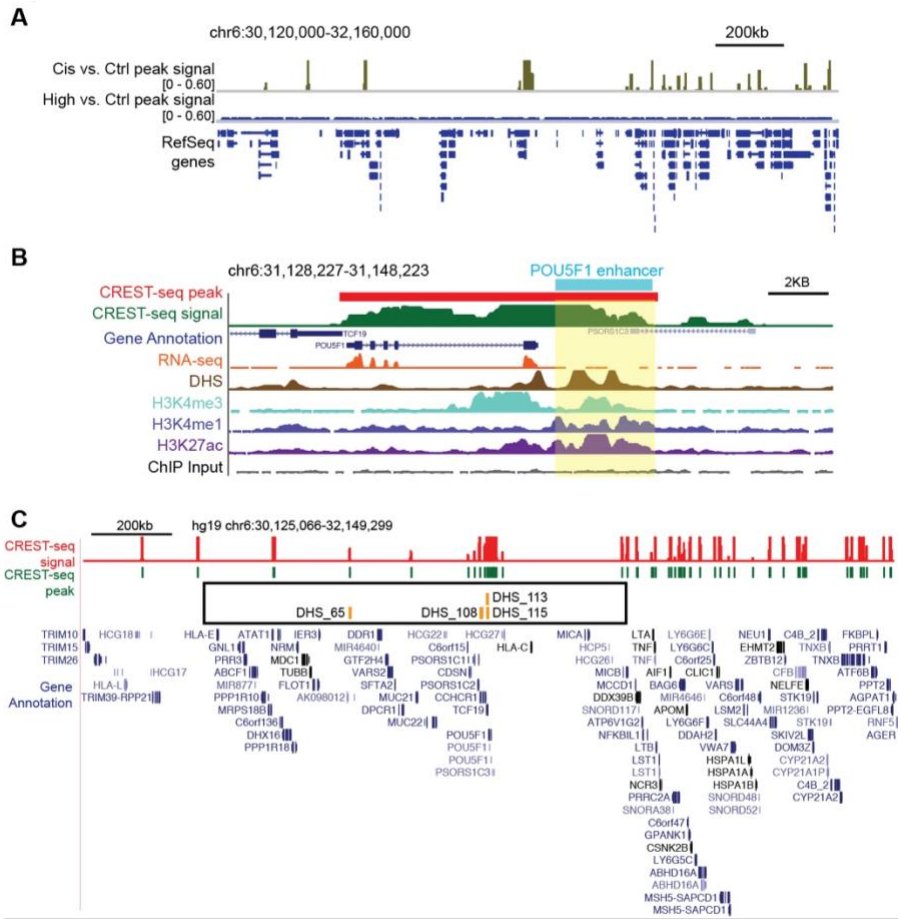


Figure S4.4. CREST-seq identifies the promoter and known enhancers of POU5F1. (**a**) Genome browser screenshot showing the CREST-seq peak predicted from "Cis" sample and "High" sample with the same peak calling method (detailed in Material and Methods). (**b**) Genome browser screenshot showing the CREST-seq peak (top, red bar), CREST-seq signal (dark green track), and the associate features surrounding POU5F1 gene body, promoter and well characterized enhancer (blue bar and the highlighted region by yellow). (**c**) Genome browser screenshot showing the functional sites identified by CREST-seq (red and green tracks on top) compared to previous single sgRNA based screen (orange bars in the middle, DHS_65, DHS_108, DHS_113 and DHS_115). The black box highlighted the 1Mbp POU5F1 locus surveyed in our previous screen.

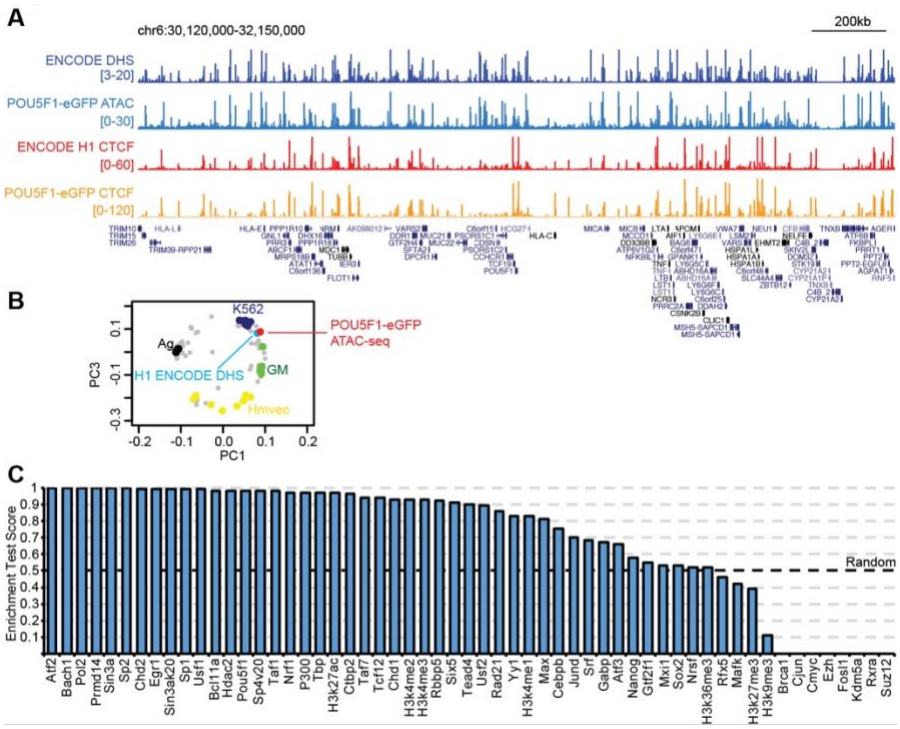


Figure S4.5. Chromatin features enriched on CREs. (**a**) Genome browser snapshot comparing the ENCODE DHS and CTCF-ChIP-seq signal with POU5F1-eGFP reporter line ATAC-seq and CTCF ChIP-seq signal within the 2Mbp tested POU5F1 locus along with gene annotation. (**b**) PCA analysis showing the clustering of 122 public available DHS data sets, including data generated from K562 cell(10x), human lymphoblastoid cell lines (GM, 3x), human fibroblast (Ag, 5x), human dermal microvascular endothelial cells (Hmvec, 8x) and 96 other cell types. ENCODE H1 DHS data and POU5F1-eGFP reporter hESC ATAC-seq data are also included. (**c**) Bar plot shows the enrichment test score for 57 features (49 for TFBS and 8 for histone modifications) at CREs compared to random.

Figure S4.6. Genotype information for the mutant clones with genomic deletion on selected CREs. (**a**) Genomic DNA was isolated from each indicated mutant clones and the genotypes were confirmed by Sanger sequencing of genotyping PCR product after TOPO cloning. The targeted deletion regions are showing on top of each panel. The blue box, green box and red box contain the genotyping for bi-allelic, P1 allele or P2 allele deletion, respectively. P1 is the eGFP containing allele while P2 is the allele with wild-type sequence. The genome browser screenshot shows CREST-seq signal/peak, and other epigenetic features as indicated around each targeted locus including CRE(-694), CRE(-652), CRE(-571), CRE(-521), CRE(-449), CRE(+38) and CREST-seq negative region. (**b**) The eGFP levels on WT cells (WT Ctrl), bia-allelic deletion, P1 allele specific deletion and P2 allele specific deletion mutants was quantified with FlowJo. Both early passage cells (day 25) and long-term cultured cells (day 50) were subjected to FACS analysis. Two-sample t-test was performed to compute the P-value, Error bars, s.d.

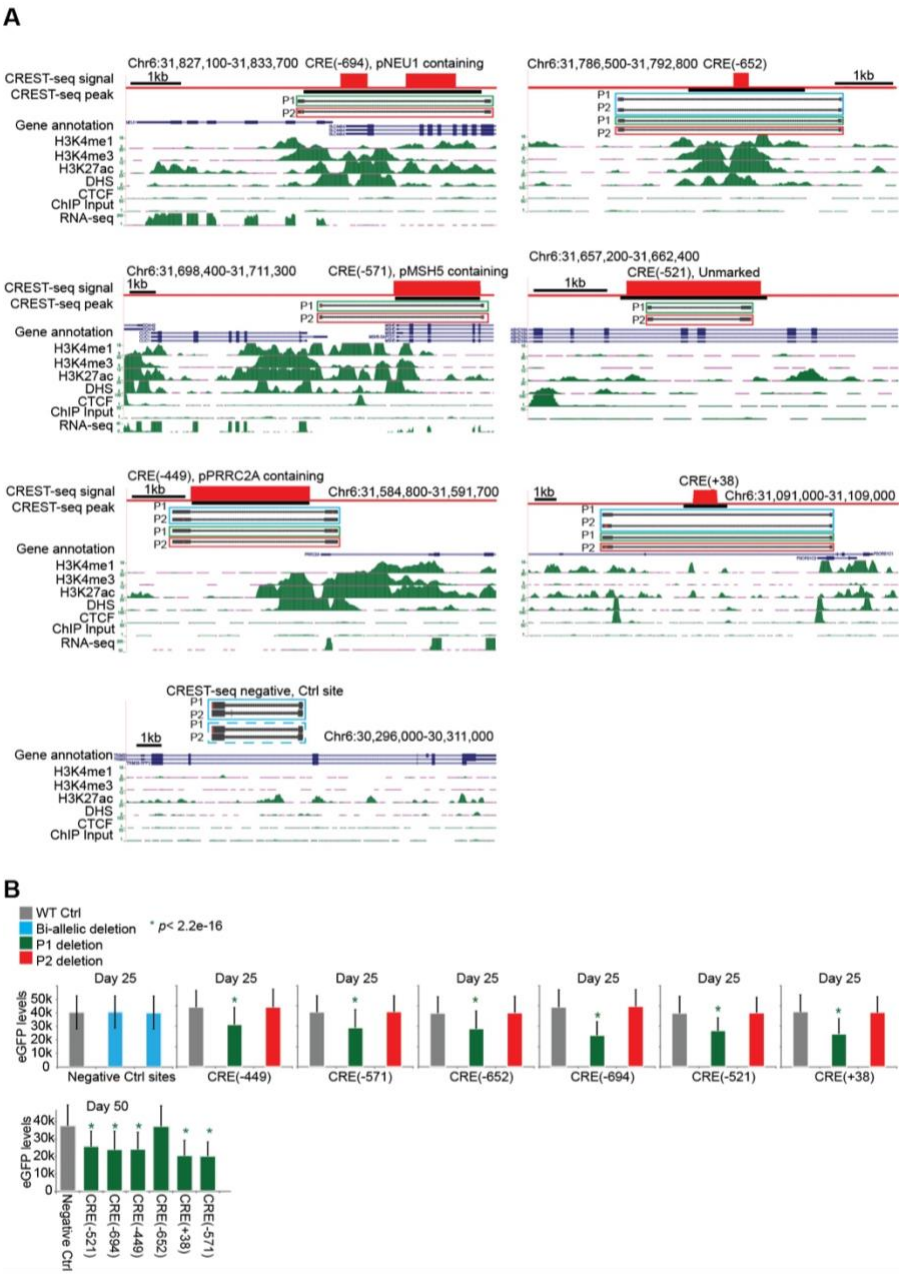

A
Chr6:31,827,100-31,833,700 CRE(-694), pNEU1 containing
Chr6:31,786,500-31,792,800 CRE(-652)
Chr6:31,698,400-31,711,300 CRE(-571), pMSH5 containing
Chr6:31,657,200-31,662,400 CRE(-521), Unmarked
CRE(-449), pPRRC2A containing
Chr6:31,584,800-31,591,700
CRE(+38)
Chr6:31,091,000-31,109,000
CREST-seq negative, Ctrl site
Chr6:30,296,000-30,311,000
CREST-seq signal
CREST-seq peak
Gene annotation
H3K4me1
H3K4me3
H3K27ac
DHS
CTCF
ChIP Input
RNA-seq
1kb
P1
P2
B
WT Ctrl
Bi-allelic deletion
P1 deletion
P2 deletion
* p< 2.2e-16
Day 25
Day 50
eGFP levels
Negative Ctrl sites
CRE(-449)
CRE(-571)
CRE(-652)
CRE(-694)
CRE(-521)
CRE(+38)
Negative Ctrl

Figure S4.7. Genotype information for core promoter mutant clones. The genotype of each mutant clones was determined by genotyping PCR using genomic DNA as template, followed by Sanger sequencing for verification. The blue box, green box and red box highlight the genotyping for bi-allelic, P1 allele or P2 allele deletion, respectively. P1 is the eGFP containing allele while P2 is the allele with wild-type sequence. The genome browser screenshot shows CREST-seq signal/peak, and other epigenetic features as indicated around each targeted locus. From top to bottom: Genotype information of MSH5, NEU1, PRRC2A, and TCF19 core promoter deletion mutants, respectively.

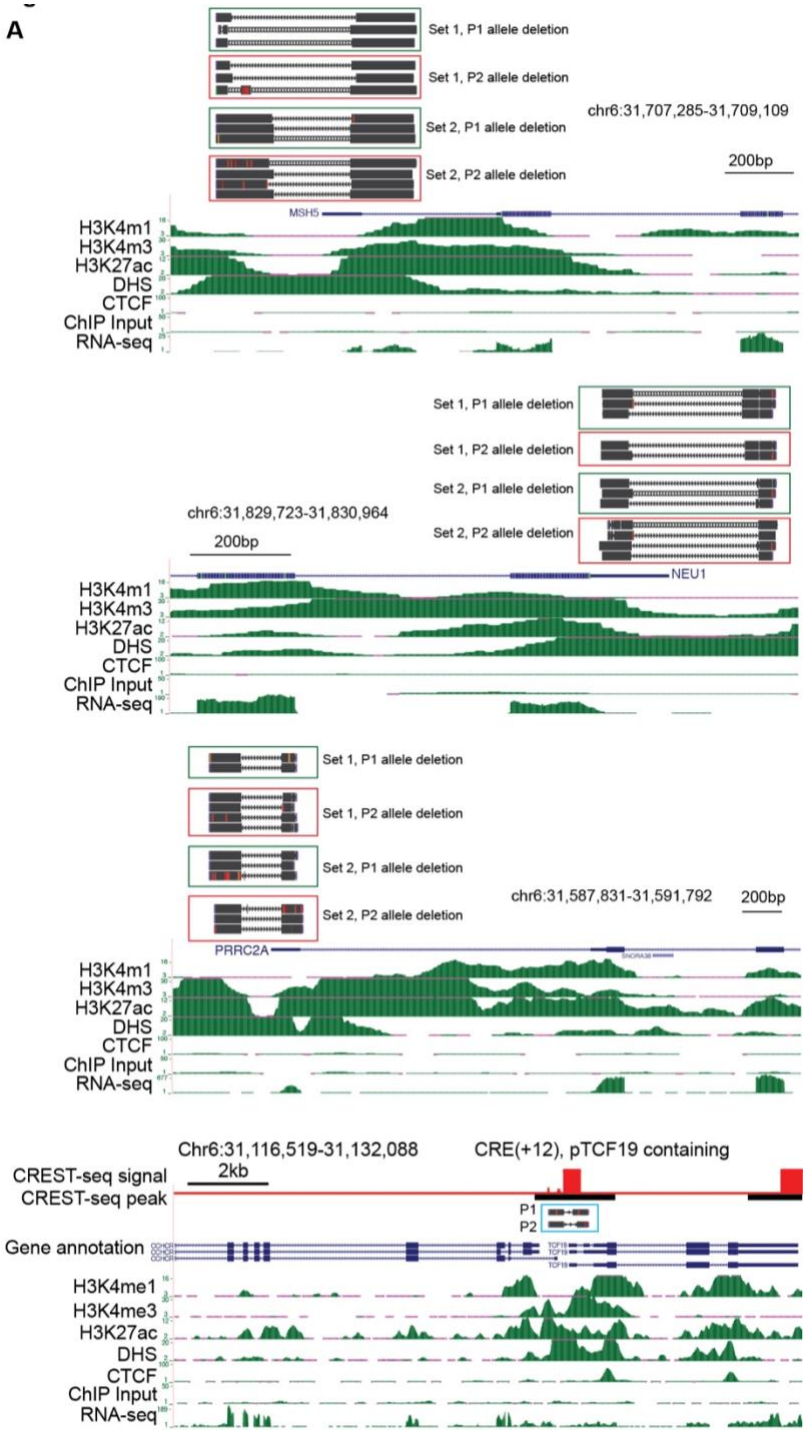
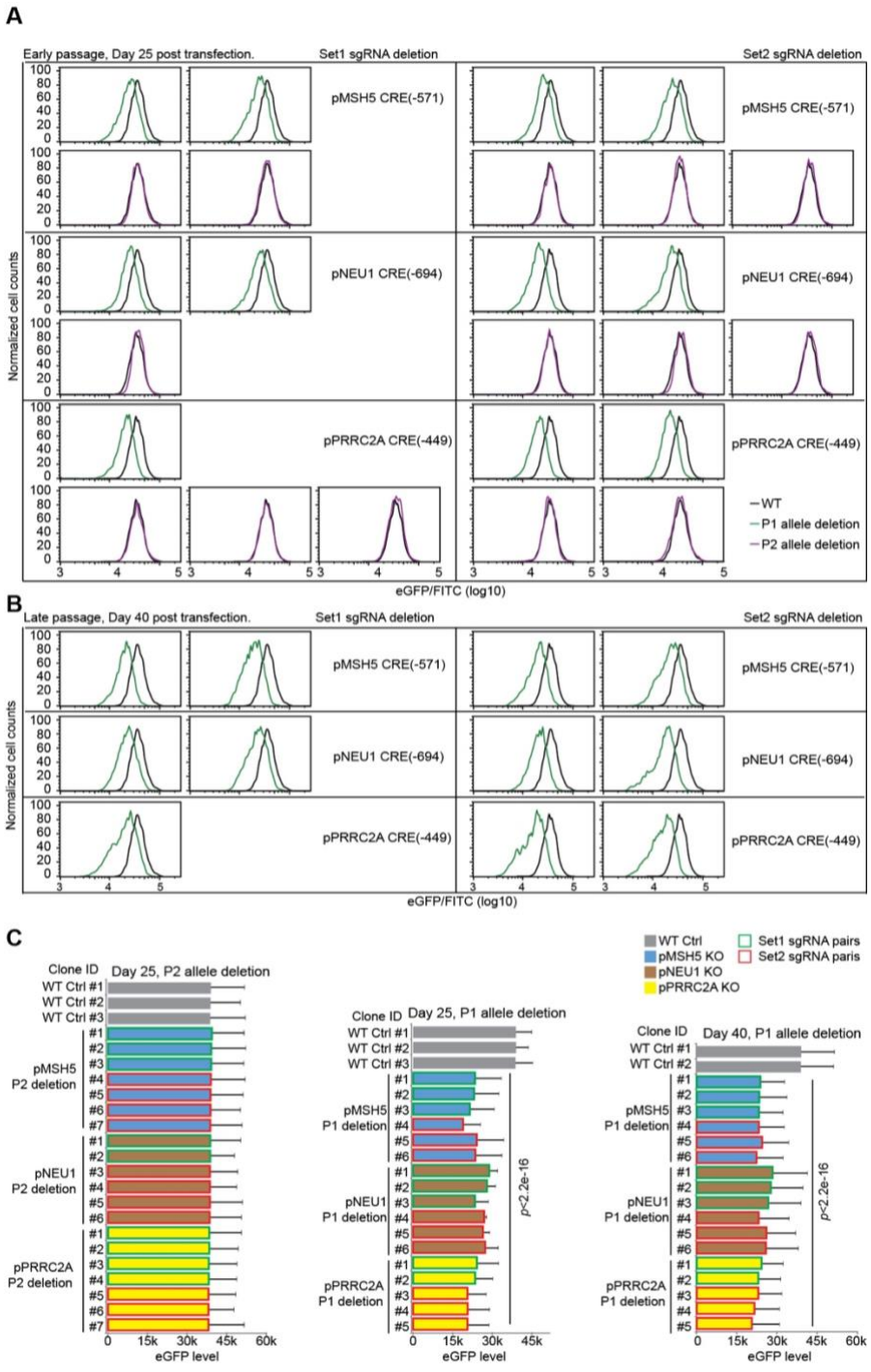

A
Set 1, P1 allele deletion
Set 1, P2 allele deletion
Set 2, P1 allele deletion
Set 2, P2 allele deletion
chr6:31,707,285-31,709,109
200bp
MSH5
H3K4m1
H3K4m3
H3K27ac
DHS
CTCF
ChIP Input
RNA-seq
Set 1, P1 allele deletion
Set 1, P2 allele deletion
Set 2, P1 allele deletion
Set 2, P2 allele deletion
chr6:31,829,723-31,830,964
200bp
NEU1
H3K4m1
H3K4m3
H3K27ac
DHS
CTCF
ChIP Input
RNA-seq
Set 1, P1 allele deletion
Set 1, P2 allele deletion
Set 2, P1 allele deletion
Set 2, P2 allele deletion
chr6:31,587,831-31,591,792
200bp
PRRC2A
H3K4m1
H3K4m3
H3K27ac
DHS
CTCF
ChIP Input
RNA-seq
Chr6:31,116,519-31,132,088
2kb
CRE(+12), pTCF19 containing
CREST-seq signal
CREST-seq peak
P1
P2
Gene annotation
H3K4me1
H3K4me3
H3K27ac
DHS
CTCF
ChIP Input
RNA-seq

Figure S4.8. Characterization and quantification of eGFP levels in multiple core promoter deletion mutant clones. Total of 37 mutant clones were generated in the same way as described in **Figure 4.3a**. In addition to the 12 mutant clones showing in Figure 4.3a, the additional 25 multiple mutant clones were also subjected to FACS analysis at (**a**) day 25 and (**b**) day 40 after CRISPR/Cas9 transfection. (**c**) The FACS data of the mutant clones showing in (**a**), (**b**), and **Figure 4.3a** were analyzed with FlowJo to quantify the eGFP level. P-value was calculated with two-sample t-test. Error bars, s.d.


A
Early passage, Day 25 post transfection.
Set1 sgRNA deletion
Set2 sgRNA deletion
pMSH5 CRE(-571)
pNEU1 CRE(-694)
pPRRC2A CRE(-449)
Normalized cell counts
WT
P1 allele deletion
P2 allele deletion
eGFP/FITC (log10)
B
Late passage, Day 40 post transfection.
C
WT Ctrl
pMSH5 KO
pNEU1 KO
pPRRC2A KO
Set1 sgRNA pairs
Set2 sgRNA paris
Clone ID
Day 25, P2 allele deletion
Day 25, P1 allele deletion
Day 40, P1 allele deletion
eGFP level
p<2.2e-16

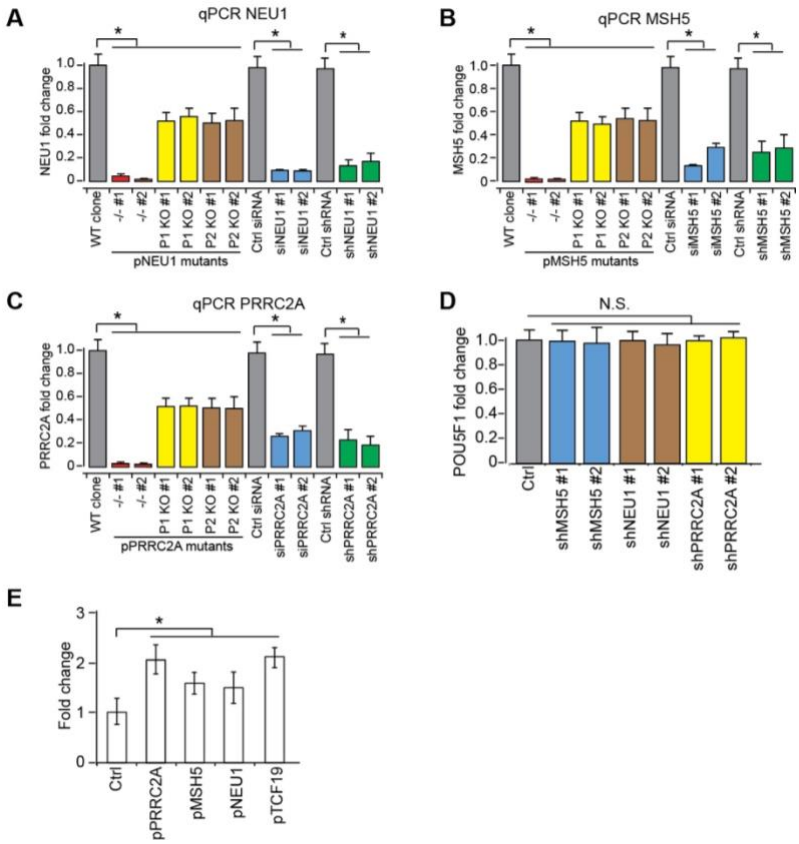


Figure S4.9. Quantification of POU5F1, MSH5, NEU1 and PRRC2A expression in various samples. The H1 POU5F1-eGFP cells were transfected with either control scrambled siRNA or siRNAs targeting each gene as indicated. Each gene is targeted by two sets of siRNAs (siRNA #1 and #2) with different sequence. 48 hours after transfection, the total RNA was collected from the cells for RT-qPCR analysis. We also packaged lentiviral expressing two sets of shRNAs targeting each gene as indicated (shRNA#1 and shRNA#2). 16 days after lentiviral infection and antibiotic selection (1mg/ml puromycin), the cells were collected for RNA purification followed by qPCR analysis. We also selected some mutant clones with core promoter deletion specified as in **Figure S4.9c** for qPCR analysis. (**a-c**) RT-qPCR analysis of NEU1, MSH5 and PRRC2A in the samples treated with siRNA, shRNA expressing lentiviral, or deletion on core promoter sequence as indicated. * P-value < 0.01, N.S. not significant, t-test, error bars, s.d. (**d**) RT-qPCR quantification of POU5F1 mRNA levels in the samples with long-term knockdown of MSH5, NEU1 and PRRC2A. * P-value < 0.01, N.S. not significant, t-test, error bars, s.d. (**e**) Bar chart showing the results from reporter assays testing four different POU5F1-regulatory core promoters. H1 hESC cells were transfected with various luciferase reporter plasmid as indicated. 48 hours post-transfection, cells were lysed and subjected to analysis of luciferase activities. All tested elements are cloned into the downstream of luciferase gene coding sequence in the control reporter (Ctrl) plasmid, which contains the 360bp POU5F1 minimal core promoter sequence to drive reporter gene expression. The reporter activity of each element was compared to the control reporter plasmid containing POU5F1 promoter only. (*t-test: P-value<0.05, error bars, s.d.).

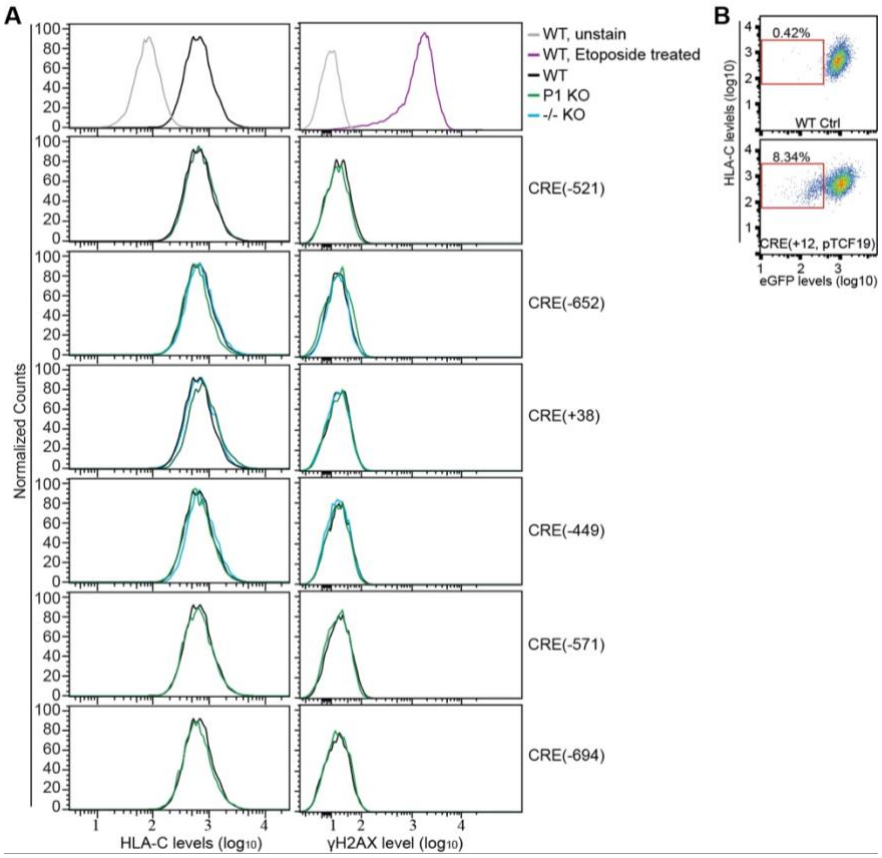


Figure S4.10. The reduced eGFP expression in bi-allelic or P1 allelic specific mutants is not due to DSB induced transcription repression. (**a**)The mutant clones with bi-allelic deletion (blue curves) or P1 allele deletion (green curves) on targeted CRE sites were dissociated into single cells and stained with PE- or PerCP-Cy5.5- conjugated antibodies specifically recognizing HLA-C or γH2AX, respectively. The black curves represent the signal obtained from WT POU5F1-eGFP reporter cells. Grey curves: WT cells without antibody staining; magenta curve: WT cells treated with 250M of Etoposide for 6 hours to induce DNA double strand break (positive control for H2AX staining signal). (**b**) WT POU5F1-eGFP reporter cells (top) and CRE(+12) biallelic (-/-) mutant (bottom, day 25 after CRISPR/Cas9 transfection) were stained with HLA-C antibody, followed by FACS analysis.

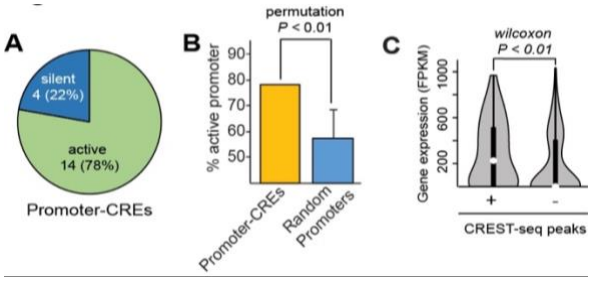


Figure S4.11. Promoter-CREs are associated with active gene expression. (**a**) A Pie-chart shows that 14 promoter-intersected CREST-seq peaks contain active promoter signatures (Pol2/H3K4me3/H3K27ac). (**b**) A Bar chart shows that POU5F1-regulating promoters are enriched for active promoter signatures (Pol2/H3K4me3/H3K27ac) compared to random promoters in the region (permutation P-value < 0.01). To estimate the degree of the enrichment, we randomly shuffled 45 CREST-seq peaks within the 2Mbp region and calculated the ratio of peaks that contain active promoter marks (Pol2/H3K4me3/H3K27ac) as expected active promoter ratio. This is repeated for 1,000 times, allowing definition of permutation P-value as the percentage of observations that active-promoter ratio is above an observed ratio (78%) (see **Supplementary Methods** for more details). (**c**) A Violin plot shows that transcriptional activities of the POU5F1-regulating promoters are higher than other gene promoters in the 2Mbp region (Wilcoxon P-value < 0.01). We used gene expression profiles from ENCODE previously quantified and normalized using ENCODE uniform pipeline.

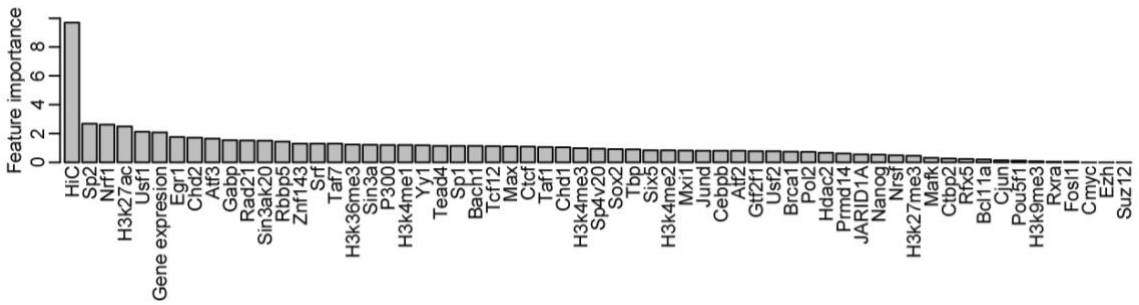


Figure S4.12. List of features that distinguish POU5F1 regulatory promoters from other non-POU5F1-regulatory promoters. Bar plot reveals the relative importance of each feature to the prediction made by random forest model.

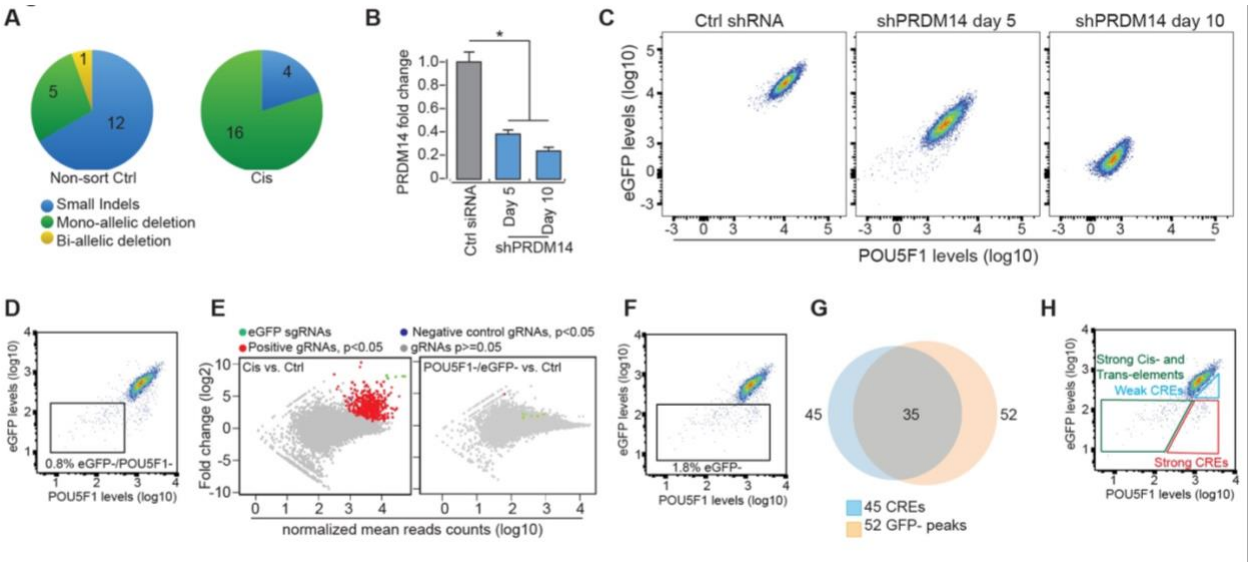


Figure S4.13. Analysis of Cis- and Trans-regulatory elements with dual sgRNA tiling deletion screen. (**a**) 18 and 20 single clones were randomly picked from the non-sorted control population and the eGFP-/POU5F1+ "Cis" population, respectively. Genomic DNA was isolated followed by PCR amplification of paired sgRNA sequence and Sanger sequencing. After confirming the sgRNA sequence, genotyping PCR was performed to check the sgRNA targeting genomic DNA sequence. (**b, c**) POU5F1-eGFP reporter cells were infected with control (Ctrl) lentiviral or shRNA targeting PRDM14 and selected with 1mg/ml puromycin for 3 days. At day 5 and day 10 after infection, (**b**) total RNA was collected and subjected to qPCR analysis to quantify the knockdown effect. * P-value < 0.01, t-test. Error bars, s.d. (**c**) The cells were dissociated and analyzed by FACS. (**d-f**) FACS analysis of H1 POU5F1-eGFP cells transduced with CREST-seq lentiviral library (right) 14 days post transduction. The eGFP-/POU5F1- cells (**d**) and eGFP- cells (**f**) were collected for further studies. (**e**) The counts of sgRNA reads from eGFP-/POU5F1+ cells (left, Cis) and eGFP-/POU5F1- (right) are compared to those from a non-sorted control population (Ctrl). The fold changes represent the ratios between the "Cis" or "eGFP-/POU5F1-" sample compared to "Ctrl" sample, with the enrichment significance calculated by negative binomial test using edgeR package. Green dots denote eGFP targeting gRNA pairs; Red dots correspond to positive oligos enriched in the testing population with P-value < 0.05 and log2 (fold change) > 1; blue dots indicate negative control oligos which are enriched with P-value < 0.05 and log2 (fold change) > 1 in the testing samples compared to Ctrl. Grey dots for the rest of sgRNAs. (**g**) The eGFP- cells were collected, processed and analyzed in the same way as Cis samples. With same peak calling pipeline and cutoff, we identified 45 CREs (blue) and 52 GFP- peaks (orange), with 35 sites overlapped. (**h**) FACS data showing that 45 CREs contains cis-regulatory elements with strong (red) and weak (blue) effect on POU5F1/eGFP expression while the 52 GFP- sites cover strong cis- and strong trans- elements.

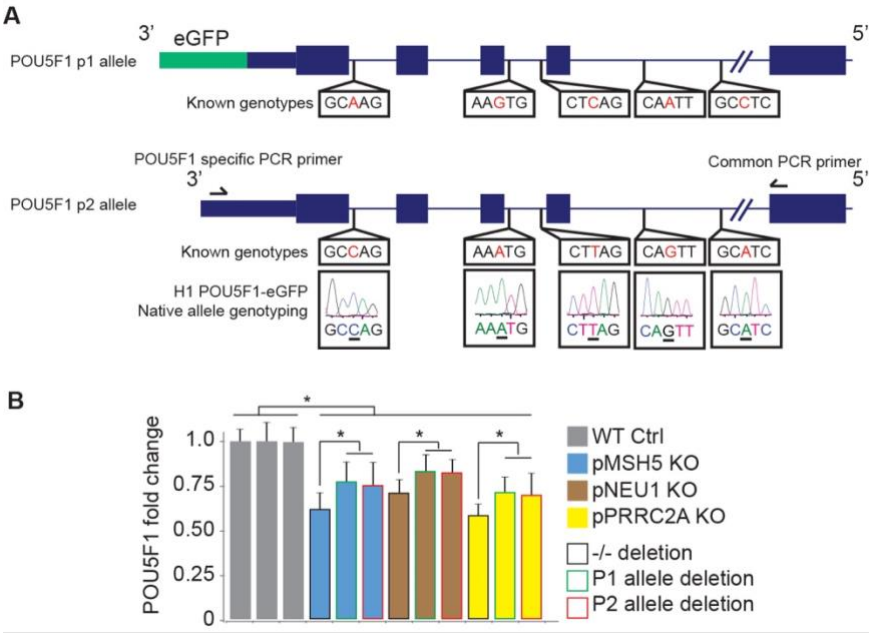


Figure S4.14. The eGFP levels correlate with P1 allele specific POU5F1 expression. (**a**) Schemtic of phasing eGFP (P1) and non-eGFP (P2) alleles of H1 POU5F1-eGFP line. We performed PCR from genomic DNA in the 3' UTR between primer pairs (indicated by black arrows) that would be broken by the inserted transgene, so the only allele that can be amplified is the native one. We then infer what the SNPs on the nontargeted allele are to deduce whether P1 or P2 is the targeted vs. non-targeted allele. (**b**) Total RNA was purified from WT and promoter-CRE mutant clones followed by qPCR analysis to quantify POU5F1 mRNA levels. * t-test, P-value<0.01, Error bars, s.d..

4.10 References

1. Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., Mu, X. J., Khurana, E., Rozowsky, J., Alexander, R., Min, R., Alves, P., Abyzov, A., Addleman, N., Bhardwaj, N., Boyle, A. P., Cayting, P., Charos, A., Chen, D. Z., Cheng, Y., Clarke, D., Eastman, C., Euskirchen, G., Frietze, S., Fu, Y., Gertz, J., Grubert, F., Harmanci, A., Jain, P., Kasowski, M., Lacroute, P., Leng, J., Lian, J., Monahan, H., O'Geen, H., Ouyang, Z., Partridge, E. C., Patacsil, D., Pauli, F., Raha, D., Ramirez, L., Reddy, T. E., Reed, B., Shi, M., Slifer, T., Wang, J., Wu, L., Yang, X., Yip, K. Y., Zilberman-Schapira, G., Batzoglou, S., Sidow, A., Farnham, P. J., Myers, R. M., Weissman, S. M. & Snyder, M. Architecture of the human regulatory network derived from ENCODE data. Nature 489, 91–100 (2012).

2. Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V. & Ren, B. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120 (2012).

3. Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., Yang, H., Wang, T., Lee, A. Y., Swanson, S. A., Zhang, J., Zhu, Y., Kim, A., Nery, J. R., Urich, M. A., Kuan, S., Yen, C., Klugman, S., Yu, P., Suknuntha, K., Propson, N. E., Chen, H., Edsall, L. E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W.-Y., Chi, N. C., Antosiewicz-Bourget, J. E., Slukvin, I., Stewart, R., Zhang, M. Q., Wang, W., Thomson, J. A., Ecker, J. R. & Ren, B. Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell 153, 1134–1148 (2013).

4. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. & Kellis, M. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015).

5. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).

6. Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. & Bernstein, B. E. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43–49 (2011).

7. Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. & Stamatoyannopoulos, J. A. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).

8. Farh, K. K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., Shoresh, N., Whitton, H., Ryan, R. J. H., Shishkin, A. A., Hatan, M., Carrasco-Alfonso, M. J., Mayer, D., Luckey, C. J., Patsopoulos, N. A., De Jager, P. L., Kuchroo, V. K., Epstein, C. B., Daly, M. J., Hafler, D. A. & Bernstein, B. E. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 518, 337–343 (2015).

9. Gjoneska, E., Pfenning, A. R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.-H. & Kellis, M. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature 518, 365–369 (2015).

10. Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., Chen, D. D., Schupp, P. G., Vinjamur, D. S., Garcia, S. P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.-C., Zhang, F., Orkin, S. H. & Bauer, D. E. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature 527, 192–197 (2015).

11. Korkmaz, G., Lopes, R., Ugalde, A. P., Nevedomskaya, E., Han, R., Myacheva, K., Zwart, W., Elkon, R. & Agami, R. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. Nature Biotechnology 34, 192–198 (2016).

12. Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J. M., Gifford, D. K. & Sherwood, R. I. High-throughput mapping of regulatory DNA. Nature Biotechnology 34, 167–174 (2016).

13. Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y., Dixon, J., Maliskova, L., Guan, K., Shen, Y. & Ren, B. A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening. Genome Research 26, 397–405 (2016).

14. Sanjana, N. E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., Cheng, C., Regev, A. & Zhang, F. High-resolution interrogation of functional elements in the noncoding genome. Science 353, 1545–1549 (2016).

15. Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S. & Engreitz, J. M. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science 354, 769–773 (2016).

16. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science 337, 816–821 (2012).

17. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155, 733–740 (2009).

18. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature 507, 62–67 (2014).

19. Zwaka, T. P. & Thomson, J. A. Homologous recombination in human embryonic stem cells. Nature Biotechnology 21, 319–321 (2003).

20. Li, W., Xu, H., Xiao, T., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M. & Liu, X. S. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biology 15, (2014).

21. Ware, C. B., Nelson, A. M., Mecham, B., Hesson, J., Zhou, W., Jonlin, E. C., Jimenez-Caliani, A. J., Deng, X., Cavanaugh, C., Cook, S., Tesar, P. J., Okada, J., Margaretha, L., Sperber, H., Choi, M., Blau, C. A., Treuting, P. M., Hawkins, R. D., Cirulli, V. & Ruohola-Baker, H. Derivation of naive human embryonic stem cells. Proceedings of the National Academy of Sciences 111, 4484–4489 (2014).

22. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature Methods 10, 1213–1218 (2013).

23. Ghirlando, R. & Felsenfeld, G. CTCF: making the right connections. Genes & Development 30, 881–891 (2016).

24. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. Molecular Cell 62, 668–680 (2016).

25. Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M. & Taipale, J. Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. Cell 154, 801–813 (2013).

26. MacArthur, S., Li, X.-Y., Li, J., Brown, J. B., Chu, H. C., Zeng, L., Grondona, B. P., Hechmer, A., Simirenko, L., Keränen, S. V., Knowles, D. W., Stapleton, M., Bickel, P., Biggin, M. D. & Eisen, M. B. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of

thousands of genomic regions. Genome Biology 10, R80 (2009).

27. Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M. & Gerstein, M. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biology 13, R48 (2012).

28. Chandra, T., Kirschner, K., Thuret, J.-Y., Pope, B. D., Ryba, T., Newman, S., Ahmed, K., Samarajiwa, S. A., Salama, R., Carroll, T., Stark, R., Janky, R., Narita, M., Xue, L., Chicas, A., Nũnez, S., Janknecht, R., Hayashi-Takanaka, Y., Wilson, M. D., Marshall, A., Odom, D. T., Babu, M. M., Bazett-Jones, D. P., Tavaré, S., Edwards, P. A. W., Lowe, S. W., Kimura, H., Gilbert, D. M. & Narita, M. Independence of Repressive Histone Marks and Chromatin Compaction during Senescent Heterochromatic Layer Formation. Molecular Cell 47, 203–214 (2012).

29. Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A. & Young, R. A. Super-Enhancers in the Control of Cell Identity and Disease. Cell 155, 934–947 (2013).

30. Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.-L., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K. I., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M. J., Cheung, E., Liu, E., Sung, W.-K., Snyder, M. & Ruan, Y. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell 148, 84–98 (2012).

31. Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M. & Lander, E. S. Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature 539, 452–455 (2016).

32. Chia, N.-Y., Chan, Y.-S., Feng, B., Lu, X., Orlov, Y. L., Moreau, D., Kumar, P., Yang, L., Jiang, J., Lau, M.-S., Huss, M., Soh, B.-S., Kraus, P., Li, P., Lufkin, T., Lim, B., Clarke, N. D., Bard, F. & Ng, H.-H. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. Nature 468, 316–320 (2010).

33. Rogakou, E. P., Boon, C., Redon, C. & Bonner, W. M. Megabase Chromatin Domains Involved in DNA Double-Strand Breaks in Vivo. The Journal of Cell Biology 146, 905–916 (1999).

34. Downs, J. A., Lowndes, N. F. & Jackson, S. P. A role for Saccharomyces cerevisiae histone H2A in DNA repair. Nature 408, 1001–1004 (2000).

35. Burma, S., Chen, B. P., Murphy, M., Kurimasa, A. & Chen, D. J. ATM Phosphorylates Histone H2AX in Response to DNA Double-strand Breaks. Journal of Biological Chemistry 276, 42462–42467 (2001).

36. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V.

V., Ecker, J. R., Thomson, J. A. & Ren, B. Chromatin architecture reorganization during stem cell differentiation. Nature 518, 331–336 (2015).

37. Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A. & Lis, J. T. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature Genetics 46, 1311–1320 (2014).

38. Paralkar, V. R., Taborda, C. C., Huang, P., Yao, Y., Kossenkov, A. V., Prasad, R., Luan, J., Davies, J. O. J., Hughes, J. R., Hardison, R. C., Blobel, G. A. & Weiss, M. J. Unlinking an lncRNA from Its Associated cis Element. Molecular Cell 62, 104–110 (2016).

39. Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W. H., Ye, C., Ping, J. L. H., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C. S., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W.-K., Ruan, Y. & Wei, C.-L. CTCF-mediated functional chromatin interactome in pluripotent cells. Nature Genetics 43, 630–638 (2011).

40. DeMare, L. E., Leng, J., Cotney, J., Reilly, S. K., Yin, J., Sarro, R. & Noonan, J. P. The genomic landscape of cohesin-associated chromatin interactions. Genome Research 23, 1224–1234 (2013).

41. Kieffer-Kwon, K.-R., Tang, Z., Mathe, E., Qian, J., Sung, M.-H., Li, G., Resch, W., Baek, S., Pruett, N., Grøntved, L., Vian, L., Nelson, S., Zare, H., Hakim, O., Reyon, D., Yamane, A., Nakahashi, H., Kovalchuk, A. L., Zou, J., Joung, J. K., Sartorelli, V., Wei, C.-L., Ruan, X., Hager, G. L., Ruan, Y. & Casellas, R. Interactome Maps of Mouse Gene Regulatory Domains Reveal Basic Principles of Transcriptional Regulation. Cell 155, 1507–1520 (2013).

42. Ji, X., Dadon, D. B., Powell, B. E., Fan, Z. P., Borges-Rivera, D., Shachar, S., Weintraub, A. S., Hnisz, D., Pegoraro, G., Lee, T. I., Misteli, T., Jaenisch, R. & Young, R. A. 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell 18, 262–275 (2016).

43. Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-L., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G. & Ruan, Y. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell 163, 1611–1627 (2015).

44. Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. A. & Ren, B. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290–294 (2013).

45. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. & Aiden, E. L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159, 1665–1680 (2014).

46. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380 (2012).

47. Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M. & Stark, A. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science 339, 1074–1077 (2013).

48. Thakore, P. I., D'Ippolito, A. M., Song, L., Safi, A., Shivakumar, N. K., Kabadi, A. M., Reddy, T. E., Crawford, G. E. & Gersbach, C. A. Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nature Methods 12, 1143–1149 (2015).

49. Diao, Y., Fang, R., Li, B. & Ren, B. A dual sgRNA mediated tiling-deletion based genetic screen to identify regulatory DNA sequence in mammalian cells. Protocol Exchange (2017). doi:10.1038/protex.2017.037

50. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, R25 (2009).

51. Wu, X., Kriz, A. J. & Sharp, P. A. Target specificity of the CRISPR-Cas9 system. Quantitative Biology 2, 59–70 (2014).

52. Meng, Z., Li, T., Ma, X., Wang, X., Van Ness, C., Gan, Y., Zhou, H., Tang, J., Lou, G., Wang, Y., Wu, J., Yen, Y., Xu, R. & Huang, W. Berbamine Inhibits the Growth of Liver Cancer Cells and Cancer-Initiating Cells by Targeting Ca2+/Calmodulin-Dependent Protein Kinase II. Molecular Cancer Therapeutics 12, 2067–2077 (2013).

53. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 28, 573–580 (2012).

54. Rajagopal, N., Xie, W., Li, Y., Wagner, U., Wang, W., Stamatoyannopoulos, J., Ernst, J., Kellis, M. & Ren, B. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. PLoS Computational Biology 9, e1002968 (2013).

55. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

56. Kelley, M. L., Strezoska, Ž., He, K., Vermeulen, A. & Smith, A. van B. Versatility of chemically synthesized guide RNAs for CRISPR-Cas9 genome editing. Journal of Biotechnology 233, 74–83 (2016).

57. Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E. & Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics 39, 311–318 (2007).

58. Diao, Y., Guo, X., Li, Y., Sun, K., Lu, L., Jiang, L., Fu, X., Zhu, H., Sun, H., Wang, H. & Wu, Z. Pax3/7BP Is a Pax7- and Pax3-Binding Protein that Regulates the Proliferation of Muscle Precursor Cells by an Epigenetic Mechanism. Cell Stem Cell 11, 231–241 (2012).

59. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. & Liu, X. S. Model-based Analysis of ChIP-Seq (MACS). Genome Biology 9, R137 (2008).

60. Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F. & Manke, T. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research 44, W160–W165 (2016).