UNIVERSITY OF CALIFORNIA SAN DIEGO

DNA methylation landscape of the mouse and human brain at single-cell resolution

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

in

Biology

by

Hanqing Liu

Committee in charge:

Professor Joseph R. Ecker, Chair
Professor Edward M. Callaway
Professor Emma K. Farley
Professor Bing Ren
Professor Terrence J Sejnowski

2022

The Dissertation of Hanqing Liu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my family and the people who have supported me throughout my education.

Thank you for giving me the opportunity to take this adventure.

EPIGRAPH

It is essential to understand our brains in some detail if we are to assess correctly our place in this vast and complicated universe we see all around us.

Francis Crick, What Mad Pursuit

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

DEDICATION iv

EPIGRAPH v

TABLE OF CONTENTS vi

LIST OF FIGURES xii

LIST OF TABLES xv

LIST OF SUPPLEMENTAL FILES xvi

ACKNOWLEDGEMENTS xviii

VITA xxiii

ABSTRACT OF THE DISSERTATION xxv

CHAPTER 1 INTRODUCTION 1

1.1 Building A Mouse Brain Cell Atlas with Single-nucleus Methylome Sequencing 2

1.2 Multi-omic Profiling in the Human Brain 3

1.3 Analysis Framework for Single-Cell Methylome and Multi-omic Datasets 5

CHAPTER 2 DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution 8

2.1 Abstract 8

2.2 Single-cell DNA methylome atlas 9

2.3 Consensus epigenomic profiles 11

2.4 Projection specificity of ET-L5 neurons 12

2.5 Regulatory taxonomy of neuronal subtypes 12

2.6 Enhancer–gene Interactions 15

2.7 3D genome structure of hippocampus 16

2.8 mC Gradients in IT neurons 18

2.9 mC gradients in DG granule cells 19

2.10 Cell type and spatial prediction model 21
2.11 Discussion 21
2.12 METHODS 23
2.12.1 Mouse brain tissues 23
2.12.2 Fluorescence-activated nuclei sorting 24
2.12.3 Library preparation and Illumina sequencing 25
2.12.4 The sn-m3C-seq specific steps of library preparation 25
2.12.5 Mouse brain region nomenclature 25
2.12.6 Analysis stages 26
2.12.7 Mapping and feature generation 26
2.12.8 Clustering-related methods 27
2.12.9 Cell-type-specific regulatory elements 37
2.12.10 Figure-specific methods 40
2.12.11 Prediction model description 44
2.13 Data availability 46
2.14 Code availability 46
2.15 Author Contributions 46
2.16 Acknowledgements 47
2.17 Figures 47
2.18 Supplementary Figures 54
2.19 Supplementary Notes 74
2.19.1 Neuronal Subtype Vignettes 74
2.19.2 Estimate subtype CG-DMR false discovery rate 75
2.19.3 Total Impact score summarizes variation of each gene or motif 76
2.19.4 Spatial axis of DG granule cells 77

2.19.5 Artificial Neuron Network trained for predicting non-neuronal cells. 78

CHAPTER 3 SINGLE NUCLEUS MULTI-OMICS IDENTIFIES HUMAN CORTICAL CELL REGULATORY GENOME DIVERSITY 79

3.1 Abstract 79

3.2 Design 79

3.3 Joint analysis of RNA and DNA methylome in cultured human cells 81

3.4 Multi-omic profiling of postmortem human brain tissue with snmCAT-seq 82

3.5 Paired RNA and mC profiling enables cross-validation and quantification of over-/under-splitting for single-cell clusters 84

3.6 Diverse correlation between gene body mCH and gene expression 87

3.7 Multi-omic integration of chromatin conformation, transcriptome, methylome and chromatin accessibility 89

3.8 snmCAT-seq identifies RNA and mC signatures of neuronal subtypes. 91

3.9 DNA methylation signatures of hierarchical transcription factor regulation in neural lineages 92

3.10 Cortical cell regulatory genomes predict developmental and adult cell types associated with neuropsychiatric diseases 96

3.11 Discussion 99

3.12 Limitations of the Study 101

3.13 Methods 102

3.13.1 Cell cultures 102

3.13.2 Human brain tissues 103

3.13.3 Nuclei isolation from cultured cells for snmCAT-seq 104

3.13.4 Nuclei isolation from human brain tissues and GpC methyltransferase treatment for snmCAT-seq 104

3.13.5 Reverse transcription for snmCAT-seq 105

3.13.6 cDNA amplification for snmCAT-seq 106

3.13.7 Digestion of unincorporated DNA oligos for snmCAT-seq 107

3.13.8 Bisulfite conversion and library preparation 107

3.13.9 The mapping pipeline for snmC-seq, snmC-seq2, and snmCAT-seq 108

3.13.10 Methylome feature generation 109

3.13.11 Preprocessing of snmC-seq and snmC-seq2 data for clustering analyses 110

3.13.12 Preprocessing of snmCAT-seq data for clustering analysis 111

3.13.13 General strategies for clustering and manifold learning 113

3.13.14 Identification of open chromatin regions using snmCAT-seq GCY methylation profiles 115

3.13.15 snATAC-seq data generation 116

3.13.16 snATAC-seq data processing 117

3.13.17 Clustering analysis of snATAC-seq data 118

3.13.18 Open chromatin peak calling using snATAC-seq data 119

3.13.19 snRNA-seq data generation 119

3.13.20 snRNA-seq clustering and annotation 119

3.14 FIGURE-SPECIFIC METHODS 121

3.14.1 Cell line dataset analysis (Supplementary Fig. 3.1) 121

3.14.2 snmCAT-seq baseline clustering (Fig. 3.1) 121

3.14.3 Methylome ensemble clustering (Fig. 3.4) 122

3.14.4 Cross-validation of cell clusters (Fig. 3.2) 122

3.14.5 AIC and BIC metrics in the cluster cross-validation analysis (Fig. 3.2) 123

3.14.6 Quantification of over-splitting and under-splitting of cell clusters (Fig. 3.2) 125

3.14.7 Computational data fusion with SingleCellFusion (Fig. 3.2) 128

3.14.8 Evaluation of Computational Data Fusion Methods (Supplementary Fig. 3.3) 131

3.14.9 Metrics for evaluation of data fusion 133

3.14.10 Correlation analysis of RNA expression and gene body DNA methylation (Fig. 3.3) 133

3.14.11 Eta Squared of Genes Across Clusters (Fig. 3.3) 134

3.14.12 H3K27me3 ChIP-Seq data processing (Fig. 3.3) 134

3.14.13 Fusion of DNA methylome and snATAC-Seq data (Fig. 3.4) 134

3.14.14 snmCAT-seq - snRNA-seq integration (Fig. 3.5, excitatory, inhibitory, non-neuron separately) 135

3.14.15 Cell type dendrogram and sub-cluster merge along the lineage (Fig. 3.6) 136

3.14.16 Neural lineage-specific DMR calling and motif enrichment analysis (Fig. 3.6) 136

3.14.17 TF binding preference to methylated motifs (Fig. 3.6) 137

3.14.18 Chromatin accessibility analysis of TF binding motifs (Fig. 3.6F-M and Supplementary Fig. 3.6I-L) 137

3.14.19 Partitioned heritability analysis (Fig. 3.7 and Supplementary Fig. 3.7) 138

3.14.20 Prioritization of trait-associated cell types using RolyPoly (Supplementary Fig. 3.7) 139

3.15 Data and Code Availability 139

3.16 Author Contributions 140

3.17 Acknowledgements 140

3.18 Tables 142

3.19 Figures 146

3.20 Supplementary Figures 158

CHAPTER 4 ALLCools: a comprehensive and scalable single-cell DNA methylation analysis framework 172

4.1 Abstract 172

4.2 ALLCools Analysis Overview 172

4.3 ALLCools Data Structures 174

4.4 Cellular Analysis with ALLCools MCDS 175

4.5 Genomic Analysis with ALLCools RegionDS 178

4.6 Discussion 180

4.7 Methods 181
4.7.1 Implementation of ALLCools 181
4.7.2 Implementation of ALLC, MCDS and RegionDS format 182
4.7.3 Fraction-based clustering using genome 100kb bins 183
4.7.4 LSI-based clustering using genome 5kb bins 184
4.7.5 Multi-omic analysis 184
4.7.7 Enhancer prediction workflow 184
4.7.8 Chromatin conformation analysis workflow 186
4.8 Data and Code Availability 186
4.9 Acknowledgements 186
4.10 Figures 187
4.11 Supplementary Figures 192
REFERENCES 201

LIST OF FIGURES

Figure 2.1 A survey of single-cell DNA methylomes in the mouse brain.....48

Figure 2.2 Epigenomic diversity of neurons.....49

Figure 2.3. Relating genes and regulatory elements to cell subtype taxonomy.....50

Figure 2.4 Gene–enhancer landscapes in neuronal subtypes.....51

Figure 2.5 Brain-wide spatial gradients of DNA methylation.....52

Figure 2.6 A methylome-based predictive model captures both cellular and spatial characteristics of neurons.....53

Supplementary Figure 2.1 Brain dissection regions.....54

Supplementary Figure 2.2 Major Type labelling and basic mapping metrics of snmC-seq2....56

Supplementary Figure 2.3 Cell-type composition of dissection regions.....58

Supplementary Figure 2.4 Supporting details of cellular and spatial diversity of neurons at the subtype level.....60

Supplementary Figure 2.5 Integration with snATAC-seq and epi-retro-seq.....62

Supplementary Figure 2.6 Controlling the FDR of CG-DMRs.....64

Supplementary Figure 2.7 Subtype taxonomy with related genes and motifs.....65

Supplementary Figure 2.8 Gene-Enhancer landscape related.....67

Supplementary Figure 2.9 DNA methylation gradient of IT neurons.....69

Supplementary Figure 2.10 DNA methylation gradient of DG granule cells.....70

Supplementary Figure 2.11 Evaluation of the predictive model.....72

Figure 3.1. snmCAT-seq generates single-nucleus multi-omic profiles of the human brain 146

Figure 3.2. Integrative analysis of RNA and mC features cross-validates neuronal cell clusters 148

Figure 3.3. Single-cell correlation analysis of RNA expression and gene body non-CG methylation.....150

Figure 3.4. Integrated epigenomic atlas of the human frontal cortex.....152

Figure 3.5. snmCAT-seq identifies RNA and mC signatures of neuronal subtypes.....154

Figure 3.6. DMR phylogeny and transcription factor hierarchy in the human cortex.....155

Figure 3.7. Identification of brain cell types involved in neuropsychiatric traits.....157

Supplementary Figure 3.1 snmCAT-seq captures transcriptome and DNA methylation signatures of H1 & HEK293 cells.....158

Supplementary Figure 3.2 snmCAT-seq generates single-nucleus multi-omic profiles from human brain tissues.....160

Supplementary Figure 3.3 Evaluation of cluster quality with paired transcriptome and methylome profiles.....162

Supplementary Figure 3.4 Diverse correlations between gene expression and gene body mCH 164

Supplementary Figure 3.5 snmCAT-seq recapitulates transcriptome and methylome signatures of neuronal subtypes.....166

Supplementary Figure 3.6 TF binding motif enrichment across the human cortical neuronal hierarchy.....168

Supplementary Figure 3.7 Prediction of causal cell types for neuropsychiatric traits using partitioned heritability analysis.....170

Figure 4.1. ALLCools analysis overview.....187

Figure 4.2. ALLCools Data Model.....188

Figure 4.3. Cellular Analysis of ALLCools.....189

Figure 4.4. Enhancer Prediction in mouse HIP dataset.....190

Figure 4.5. Chromatin contact loops link enhancer to gene TSS.....191

Supplementary Figure 4.1 DMR mCG fraction profiles in mouse HIP cell types.....192

Supplementary Figure 4.2 Enhancer-overlapping DMR accessibility profiles in mouse HIP cell types.....193

Supplementary Figure 4.3 REPTILE prediction scores of the enhancer-overlapping DMRs in mouse HIP cell types.....194

Supplementary Figure 4.4 Normalized chromatin contact strength of the enhancer-overlapping DMRs and DMGs in mouse HIP cell types.....195

Supplementary Figure 4.5 Genome browser tracks at the *Celf2* locus.....196

Supplementary Figure 4.6 Genome browser tracks at the *Slc1a3* locus.....197

Supplementary Figure 4.7 Genome browser tracks at the *Bcl11b* locus.....198

Supplementary Figure 4.8 Subset cell type enhancers with chromatin contact loops.....199

LIST OF TABLES

Table 3.1. Genomic profiling methods discussed in chapter 3 142

Table 3.2. Key resources table for chapter 3 143

LIST OF SUPPLEMENTAL FILES

Supplementary Table 2.1.xlsx

Supplementary Table 2.2.xlsx

Supplementary Table 2.3.xlsx

Supplementary Table 2.4.xlsx

Supplementary Table 2.5.xlsx

Supplementary Table 2.6.xlsx

Supplementary Table 2.7.xlsx

Supplementary Table 2.8.xlsx

Supplementary Table 2.9.xlsx

Supplementary Table 2.10.xlsx

Supplementary Table 2.11.xlsx

Supplementary Table 2.12.xlsx

Supplementary Table 2.13.xlsx

Supplementary Table 3.1.xlsx

Supplementary Table 3.2.xlsx

Supplementary Table 3.3.xlsx

Supplementary Table 3.4.xlsx

Supplementary Table 3.5.xlsx

Supplementary Table 3.6.xlsx

Supplementary Table 3.7.xlsx

Supplementary Table 3.8.xlsx

Supplementary Table 3.9.zip

Supplementary Table 3.10.zip

ACKNOWLEDGEMENTS

First, I want to thank my remarkable supervisor, Joseph Ecker, who ranks among the kindest and most intelligent people I've ever met. Joe provided me the opportunity to work on a grand and challenging project from the beginning of my graduate school. His trust made me more dedicated to my research and more confident in what I could achieve in graduate school. Besides, Joe is an exceptionally patient and visionary advisor. He allows me to explore different possibilities freely but can always provide explicit guidance whenever needed. During my graduate study, Joe has shown me how to achieve an audacious goal with high scientific impact through creative thinking, technology innovation, careful planning, and collaborative teamwork.

Next, I am indebted to other committee members, Bing Ren, Edward Callaway, Emma Farley, and Terrence Sejnowski, for their in-depth discussion and brilliant suggestions during my graduate study. With their kind advice and encouragement, I felt more assured to continue the journey of doing science and becoming a scientist.

Besides, I want to thank other faculties that have provided valuable help on my research projects. Margarita Behrens is a close collaborator on all my projects, who helped me understand many neuroanatomical knowledge and technologies from the beginning. Chongyuan Luo, a prior post-doc in the Ecker lab and now has his lab at UCLA, initialized all the single-cell technologies used in my projects. And Chongyuan also gives me great suggestions on research projects and graduate school. Eran Mukamel provides many ideas, advice, and critical thinking on all the computational analyses of my projects. Jesse Dixon helped apply snm3C-seq to my project, with in-depth discussions on how to analyze this informative dataset. Finally, I also want to thank Cornelis Murre, Kimberly Cooper, and

Roberto Malinow for welcoming me to do rotations in their lab and providing fantastic introductions to immunology, developmental biology, and neuroscience to me.

Current and previous members of the Ecker lab have been great colleagues and friends. Yupeng He, a very nice senior graduate student, gave me an initial introduction to the scientific background of the lab during my rotation. His work in the ENCODE project also inspires me a lot. Jingtian Zhou, a very talented labmate with whom I feel lucky to collaborate, contributes many efforts to multiple projects included in my dissertation. Bang-An Wang is knowledgeable about numerous molecular technologies, and conversation with him is always helpful. Likewise, Peter Berube is always willing to discuss and share his in-depth knowledge about many experiments. Wei Tian, Wenliang Wang, and Zhuzhu zhang also contribute to many works related to my dissertation and provide suggestions on science and career choice. Next, our sequencing team at GALE has participated in producing almost all the data in my dissertation. This supportive and collaborative team includes Anna Bartlett, Joe Nery, Rosa Castanon, Mia Kenworthy, Jordan Altshul, Andrew Aldridge, Angeline Rivkin, and Stephen Gonzalez. Everything I can achieve in graduate school is based on their solid work. Besides, Cesar Barragan, Manoj Hariharan, Huaming Chen, Michael Nunn, Jiaying Xu, Sheng-Yong Niu, Bruce Jow, and all labmates working at the PBIO lab also provide numerous support in my daily lab work.

I also received countless help from people outside the lab throughout my graduate career. Jacinta Lucero, Julia Osteen, and Antonio Pinto-Duarte are essential collaborators in the Behrens lab that helped process the mouse brain tissues. In addition, Fangming Xie, Ethan Armand, and Wayne Doyle from Mukamel Lab are extremely helpful in analyzing snmCT-seq data. Yang Li, Sebastian Preissl, and Rongxin Fang from Ren lab generated the excellent

snATAC-seq data and kindly shared their results during our routine discussions. Finally, many other collaborators also gave me incredible support on various aspects, including works that are still ongoing and not included in this dissertation. Nicola Allen and her lab members Tao Tao and Isabella Farhy-Tselnicker led the developmental study of the visual circuit. Stuart Sealfon and Frederique Ruf-Zamojski's lab introduced me to the cellular diversity of pituitary.

I am very grateful to Carolyn O'Connor, Conor Fitzpatrick, Lara Boggeman, Michelle Liem, and Naomi Claffey at the Salk Flow Cytometry Core. Caz and her team provide constant support on all my projects. I also want to thank Frank Dwyer and his IT team at Salk for continuous maintenance of the local computational environments. I thank many Xsede and Texas Advanced Computing Center developers who have helped me set up our analysis pipelines on their supercomputers and developers in the NeMO archive and GEO, especially Emily Clough, for helping deposit and archiving our vast amount of data.

I want to thank all my fellow graduate students I met during my graduate study. I feel deeply fortunate to have these most brilliant and encouraging friends who can talk science and dreams. In addition, I want to thank the founding faculty of UCSD Biological Science, who awarded me the "Founding Faculty Award" for 2021. I also want to thank David Goeddel for providing generous support for my "David V. Goeddel Endowed Graduate Fellowship." The faculties, staff, and students in UCSD's biological science department have created such a prestigious Ph.D. program, and I always feel incredibly honored to be part of it. I also want to thank my undergraduate mentor, Minxin Guan, Frederick Roth, Min Chen, and Ye Chen. They initiated my interest in doing research and wrote me recommendation letters when I applied to graduate school. I am particularly grateful to them for having the patience to illustrate the scientific scope to a junior student.

Last but not least, for what I have achieved so far in my life, gratitude is due mostly to my family. I thank my mother and father for giving me life. My mother particularly taught me how to be confident and optimistic about any challenges. I thank my uncle, who provides me with generous financial support and has always been a road model in my life. I thank my grandparents for encouraging me and providing me with a strong willingness to get higher education. I believe my motivation for being the first one to get a Ph.D. degree in the family was deeply rooted in your education more than fifteen years ago.

Chapter 2, in full, is a reprint of the material as it appears in Nature 2021. "DNA methylation atlas of the mouse brain at single-cell resolution," Hanqing Liu, Jingtian Zhou, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, Julia K. Osteen, Joseph R. Nery, Huaming Chen, Angeline Rivkin, Rosa G. Castanon, Ben Clock, Yang Eric Li, Xiaomeng Hou, Olivier B. Poirion, Sebastian Preissl, Antonio Pinto-Duarte, Carolyn O'Connor, Lara Boggeman, Conor Fitzpatrick, Michael Nunn, Eran A. Mukamel, Zhuzhu Zhang, Edward M. Callaway, Bing Ren, Jesse R. Dixon, M. Margarita Behrens, and Joseph R. Ecker. The authors thank Yupeng He for the advice on the methylpy and REPTILE analysis; Terrence Sejnowski for the advice on the ANN analysis. This project is supported by NIMH U19MH11483 to Joseph Ecker and Edward Callaway and NHGRI R01HG010634 to Joseph R Ecker and Jesse Dixon. The Flow Cytometry Core Facility of the Salk Institute is supported by funding from NIH-NCI CCSG: P30 014195 and Shared Instrumentation Grant S10-OD023689. Joseph Ecker is an investigator of the Howard Hughes Medical Institute. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Cell Genomics, 2022. "Single nucleus multi-omics identifies human cortical cell regulatory genome diversity."

Chongyuan Luo, Hanqing Liu, Fangming Xie, Ethan J. Armand, Kimberly Siletti, Trygve E. Bakken, Rongxin Fang, Wayne I. Doyle, Tim Stuart, Rebecca D. Hodge, Lijuan Hu, Bang-An Wang, Zhuzhu Zhang, Sebastian Preissl, Dong-Sung Lee, Jingtian Zhou, Sheng-Yong Niu, Rosa Castanon, Anna Bartlett, Angeline Rivkin, Xinxin Wang, Jacinta Lucero, Joseph R. Nery, David A. Davis, Deborah C. Mash, Rahul Satija, Jesse R. Dixon, Sten Linnarsson, Ed Lein, M. Margarita Behrens, Bing Ren, Eran A. Mukamel, and Joseph R. Ecker. This work was supported by NIH grants: 5R21HG009274, 5R21MH112161, and 5U19MH114831 to Joseph Ecker; R01MH125252 and U01HG012079 to Chongyuan Luo; R01HG010634 to Joseph Ecker and Jesse Dixon; and U01MH114812 to Ed Lein. Joseph Ecker is an investigator of the Howard Hughes Medical Institute. Wayne Doyle is supported by an NIH training award 5T32MH020002. Postmortem human brain tissues were obtained from the NIH NeuroBioBank at the University of Maryland Brain and Tissue Bank and the University of Miami Brain Endowment Bank. The authors thank the tissue donors and their families for their invaluable contributions to the advancement of science. The authors thank the QB3 Macrolab at UC Berkeley for the purification of Tn5 transposase. Work at the Center for Epigenomics was supported in part by the UC San Diego School of Medicine. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is currently being prepared for submission for publication of the material. "ALLCools: a scalable and comprehensive single-cell DNA methylation analysis framework." Hanqing Liu, Jingtian Zhou, Wei Tian, Jiayin Xu, Joseph R. Ecker. The dissertation author was the primary investigator and author of this material.

VITA

2016 Bachelor of Science in Biotechnology, Zhejiang University, China

2022 Doctor of Philosophy in Biology, University of California San Diego

PUBLICATION

Hanqing Liu is the first or co-first author (*):

Liu, Hanqing*, Jingtian Zhou*, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, et al. 2021. "DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution." Nature 598 (7879): 120–28.

Yao, Zizhen*, **Hanqing Liu***, Fangming Xie*, Stephan Fischer*, Ricky S. Adkins, Andrew I. Aldridge, Seth A. Ament, et al. 2021. "A Transcriptomic and Epigenomic Cell Atlas of the Mouse Primary Motor Cortex." Nature 598 (7879): 103–10.

Luo, Chongyuan*, **Hanqing Liu***, Fangming Xie*, Ethan J. Armand, Kimberly Siletti, Trygve E. Bakken, Rongxin Fang, et al. 2022. "Single Nucleus Multi-Omics Identifies Human Cortical Cell Regulatory Genome Diversity." Cell Genomics 2 (3). https://doi.org/10.1016/j.xgen.2022.100107.

Liu, Hanqing*, Jingtian Zhou*, Wei Tian, Jiayin Xu, Joseph R. Ecker. ALLCools: a scalable and comprehensive single-cell DNA methylation analysis framework. In preparation.

Hanqing Liu is a co-author:

Bakken, Trygve E., Nikolas L. Jorstad, Qiwen Hu, Blue B. Lake, Wei Tian, Brian E. Kalmbach, Megan Crow, et al. 2021. "Comparative Cellular Analysis of Motor Cortex in Human, Marmoset and Mouse." Nature 598 (7879): 111–19.

BRAIN Initiative Cell Census Network (BICCN). 2021. "A Multimodal Cell Census and Atlas of the Mammalian Primary Motor Cortex." Nature 598 (7879): 86–102.

Cui, Limei, Jing Zheng, Qiong Zhao, Jia-Rong Chen, Hanqing Liu, Guanghua Peng, Yue Wu, et al. 2020. "Mutations of MAP1B Encoding a Microtubule-Associated Phosphoprotein Cause Sensorineural Hearing Loss." JCI Insight 5 (23). https://doi.org/10.1172/jci.insight.136046.

Farhy-Tselnicker, Isabella, Matthew M. Boisvert, Hanqing Liu, Cari Dowling, Galina A. Erikson, Elena Blanco-Suarez, Chen Farhy, Maxim N. Shokhirev, Joseph R. Ecker, and Nicola J. Allen. 2021. "Activity-Dependent Modulation of Synapse-Regulating Genes in Astrocytes." eLife 10 (September). https://doi.org/10.7554/eLife.70514.

Jiang, Pingping, Meng Wang, Ling Xue, Yun Xiao, Jialing Yu, Hui Wang, Juan Yao, et al. 2016. “A Hypertension-Associated tRNAAla Mutation Alters tRNA Metabolism and Mitochondrial Function.” Molecular and Cellular Biology 36 (14): 1920–30.

Li, Yang Eric, Sebastian Preissl, Xiaomeng Hou, Ziyang Zhang, Kai Zhang, Yunjiang Qiu, Olivier B. Poirion, et al. 2021. “An Atlas of Gene Regulatory Elements in Adult Mouse Cerebrum.” Nature 598 (7879): 129–36.

Ruf-Zamojski, Frederique, Zidong Zhang, Michel Zamojski, Gregory R. Smith, Natalia Mendelev, Hanqing Liu, German Nudelman, et al. 2021. “Single Nucleus Multi-Omics Regulatory Landscape of the Murine Pituitary.” Nature Communications 12 (1): 2677.

Wu, Yingzhou, Roujia Li, Song Sun, Jochen Weile, and Frederick P. Roth. 2021. “Improved Pathogenicity Prediction for Rare Human Missense Variants.” American Journal of Human Genetics 108 (10): 1891–1906.

Zhang, Zhuzhu, Jingtian Zhou, Pengcheng Tan, Yan Pang, Angeline C. Rivkin, Megan A. Kirchgessner, Elora Williams, et al. 2021. “Epigenomic Diversity of Cortical Projection Neurons in the Mouse Brain.” Nature 598 (7879): 167–73.

Zheng, Jing, Wen-Fang Meng, Chao-Fan Zhang, Han-Qing Liu, Juan Yao, Hui Wang, Ye Chen, and Min-Xin Guan. 2019. “New SNP Variants of MARVELD2 (DFNB49) Associated with Non-Syndromic Hearing Loss in Chinese Population.” Journal of Zhejiang University. Science. B 20 (2): 164–69.

ABSTRACT OF THE DISSERTATION

DNA methylation landscape of the mouse and human brain at single-cell resolution

by

Hanqing Liu

Doctor of Philosophy in Biology

University of California San Diego, 2022

Professor Joseph Ecker, Chair

Mammalian brain cells show a remarkable diversity, yet the regulatory DNA landscape underlying this extensive heterogeneity is poorly understood. Cytosine DNA methylation, a covalent modification in the genome, is a critical player in brain cell gene regulation. Recent advances in single-cell technologies allow studying this epigenomic modification at an unprecedented resolution. During my Ph.D., I have developed and used single-cell methylome and multi-omic technologies to profile mouse and human brains to establish the brain cell atlas. In chapter 1, I access the epigenomes of mouse brain cell types by applying single-nucleus DNA methylation sequencing to profile 103,982 nuclei from the mouse cerebrum. I constructed

cell taxonomies containing 161 epigenetic clusters and annotated each cluster with signature genes, regulatory elements, and transcription factors. The DNA methylation landscape of excitatory neurons in the cortex and hippocampus varied continuously along spatial gradients. By combining multi-omic datasets from single nuclei and annotating the regulatory genome of these cell types, the DNA methylation atlas establishes the epigenetic basis for neuronal diversity and spatial organization throughout the mouse cerebrum. In chapter 2, colleagues and I devised single-nucleus methylcytosine, chromatin accessibility, and transcriptome sequencing (snmCAT-seq) and applied it to postmortem human frontal cortex tissue. We developed a cross-validation approach using multi-modal information to validate fine-grained cell types. We reconstructed regulatory lineages for cortical cell populations and found specific enrichment of genetic risk for neuropsychiatric traits, enabling the prediction of cell types associated with diseases. In chapter 3, I developed a comprehensive computational package called ALLCools to solve the computational challenges brought up by the single-cell methylome and multi-omic dataset. This package includes data formats specialized for the single-cell DNA methylation data and various functions covering cellular and genomic analysis. As a result, my Ph.D. study has produced high-resolution single-cell DNA methylome datasets in the mouse brain, developed new multi-omic technologies applicable to the human brain, and established comprehensive analysis frameworks to analyze the extensive data. Together, these works create concrete deliverables that extend the understanding of brain epigenome diversity and enable future studies to understand the molecular source of brain cellular diversity and link the psychological diseases to their molecular basis.

CHAPTER 1

INTRODUCTION

Epigenomic dynamics are associated with cell differentiation and maturation in the mammalian brain and have an essential role in regulating neuronal functions and animal behaviour[1,2]. Cytosine DNA methylation (5mC) is a stable covalent modification that persists in post-mitotic cells throughout their lifetime and is critical for proper gene regulation[2]. In mammalian genomes, 5mC occurs predominantly at CpG sites (mCG), showing dynamic patterns at regulatory elements with tissue and cell-type specificity[2–5], modulating binding affinity of transcription factors[6] and controlling gene transcription[1]. Non-CpG cytosines are also abundantly methylated (mCH, H denotes A, C, or T)—uniquely in neurons—in the mouse and human brain[2,7], which can directly affect DNA binding of methyl CpG binding protein 2 (MeCP2)[8–10], causing Rett Syndrome[11]. Levels of mCH at gene bodies are anti-correlated with gene expression and show high heterogeneity across neuronal cell types[3,4].

Recently, we and others have developed single-cell technologies to study DNA methylation in individual cells at single-base resolution[3,12–14], which greatly extended our knowledge of DNA methylation dynamics between different cells in heterogeneous tissue[3,15,16] or across developmental trajectories[17,18]. Besides, we and others also developed methylation-based multi-omic technologies, profiling DNA methylome together with RNA expression[19–22], chromatin accessibility[20,21,23], and conformation[24,25] within the same cell. The direct association of different molecular modalities allows linking the regulatory effect of single cytosine methylation to chromatin status and gene expression at a cell-specific level, revealing the regulatory role of DNA methylation in an unprecedented resolution[17,21]. Detail technology reviews are available[26,27].

During my Ph. D. study, I focused on three aspects to extend the understanding of DNA methylation diversity of the mammalian brain. In the first project (chapter 2), I will describe using the single-nucleus methylome sequencing (snmC-seq2) technology to establish a mouse brain cell atlas with base-resolution epigenomic profiles. In the second project (chapter 3), I will describe a novel single-nucleus multi-omic technology snmCAT-seq that is applicable to frozen human brain tissue and the human brain multi-modality dataset we generated with this technology. In the third project (chapter 4), I will describe a comprehensive analysis package that I developed to cover the analysis of single-cell methylome dataset with high scalability and reproducibility.

1.1 Building A Mouse Brain Cell Atlas with Single-nucleus Methylome Sequencing

A deeper understanding of epigenomic diversity in the mouse brain provides a complementary approach to transcriptome-based profiling methods for identifying brain cell types and allows genome-wide prediction of the regulatory elements and transcriptional networks underlying this diversity. Previous studies have demonstrated the utility of studying brain cell types and regulatory diversity using single-nucleus methylome sequencing (snmC-seq)[3]. This study uses snmC-seq2[28] to perform thorough methylome profiling with detailed spatial dissection in the adult postnatal day 56 (P56) male mouse brain. In Li et al.[29], the same tissue samples were profiled using single-nucleus assay for transposase-accessible chromatin using sequencing (snATAC-seq) to identify genome-wide accessible chromatin[30], providing complementary epigenomic information to aid in cell-type-specific regulatory genome annotation. Moreover, to further study *cis*-regulatory elements and their potential target genes across the genome, we applied single-nucleus methylation and chromosome conformation capture sequencing (sn-m3C-seq)[24] to profile the methylome and chromatin conformation in the same cells.

These epigenomic datasets provide a detailed and comprehensive census of the diversity of cell types across mouse brain regions, allowing identification of cell-type-specific regulatory elements and their candidate target genes and upstream transcription factors. Here we construct a single-cell base-resolution DNA methylation dataset containing 103,982 methylomes from 45 dissected brain regions and use an iterative analysis framework to identify 161 predicted mouse brain subtypes. Comparing subtype-level methylomes enables us to identify 3.9 million genomic regions showing cell-type-specific mCG variation, covering approximately 50% (1,240 Mb) of the mouse genome. We show that differentially methylated transcription factor genes and binding motifs can be associated with subtype taxonomy branches, allowing the prediction of cell-type gene regulatory programs specific for each developmental lineage. Integration of these data with cell clusters identified on the basis of chromatin accessibility validates most methylome-derived subtypes, enabling the prediction of 1.6 million enhancer-like genomic regions. We identify *cis*-regulatory interactions between enhancers and genes using computational prediction and single-cell chromatin conformation profiling (in the hippocampus (HIP)). We also identify spatial methylation gradients in cortical excitatory neurons and dentate gyrus granule cells and associated transcription factors and motifs. We apply an artificial neural network (ANN) model to precisely predict single-neuron cell-type identity and brain area spatial location using its methylome profile as input and develop the brain cell methylation viewer (http://neomorph.salk.edu/omb) as a portal for querying and visualization of cell- and cluster-level methylation data.

1.2 Multi-omic Profiling in the Human Brain

Single-cell transcriptome, cytosine DNA methylation (mC) and chromatin profiling techniques have been successfully applied for cell-type classification and studies of gene expression and regulatory diversity in complex tissues[31,32]. The broad range of targeted molecular

signatures, as well as technical differences between measurement platforms, presents a challenge for integrative analysis. For example, mouse cortical neurons have been studied using single-cell assays that profile RNA, mC or chromatin accessibility[3,30,33–35], with each study reporting its own classification of cell types. Although it is possible to correlate the major cortical cell types identified by transcriptomic and epigenomic approaches, it remains unclear whether fine subtypes can effectively be integrated across different datasets and fused between modalities. Recently, computational methods based on Canonical Correlation Analysis[36], mutual nearest neighbors[37] or matrix factorization[38] have been developed to fuse molecular data types. However, validating the results of computational data fusion requires multi-omic reference data comprising different types of molecular measurements made in the same cell.

Single-cell multi-omics profiling provides a unique opportunity to evaluate cell type classification using multiple molecular signatures[31]. Most single-cell studies rely on clustering analysis to identify cell types. However, it is challenging to objectively determine whether the criteria used to distinguish cell clusters are statistically appropriate and whether the resulting clusters reflect biologically distinct cell types[39]. We reasoned that genuine cell types should be distinguished by concordant molecular signatures of cell regulation at multiple levels, including RNA, mC and open chromatin, in individual cells. Moreover, multi-omic data can uncover subtle interactions among transcriptomic and epigenomic levels of cellular regulation.

Existing methods for joint profiling of transcriptome and mC, such as scM&T-seq and scMT-seq, rely on the physical separation of RNA and DNA followed by parallel sequencing library preparation[19,20,40]. Generating separate transcriptome and mC sequencing libraries leads to a complex workflow and increases cost. Moreover, it is unclear if these methods can be applied to single nuclei, which contain much less polyadenylated RNA than whole cells. Since the cell

membrane is ruptured in frozen tissues, the ability to produce robust transcriptome profiles from single nuclei is critical for applying a multi-omic assay for cell-type classification in frozen human tissue specimens.

Here we describe a single nucleus multi-omic method snmCAT-seq (single-nucleus methylCytosine, chromatin Accessibility and Transcriptome sequencing) that simultaneously interrogates transcriptome, mC and chromatin accessibility without requiring the physical separation of RNA and DNA. We applied snmCAT-seq to cultured human cells and postmortem human frontal cortex tissues. We further generated an additional 23,005 single-nucleus, droplet-based RNA-seq profiles (snRNA-seq) and 12,557 single-nucleus, snATAC-seq-based open chromatin profiles using frozen human frontal cortex tissue[30]. Using this comprehensive multimodal dataset, we developed computational strategies to tackle two challenges in single-cell biology: 1) how to assess the statistical and biological validity of clustering analyses, and 2) how to validate computational approaches to fuse multiple single-cell data types. We then performed integrated analyses of single-cell methylomes for the human frontal cortex comprised of 15,030 cells, including two multi-omic data sets generated by snmCAT-seq and the previously published sn-m3C-seq, a method to simultaneously profile chromatin conformation and mC[24]. These large datasets enabled the identification of gene regulatory diversity for 63 finely defined brain cell types at an unprecedented level of data fusion using four levels of molecular signatures (i.e., transcriptome, methylome, chromatin accessibility, and conformation) to define their unique regulatory genomes with cell-type specificity and link them to genetic disease risk variants.

1.3 Analysis Framework for Single-Cell Methylome and Multi-omic Datasets

Despite the tremendous advancement in technology development, computational challenges raised by single-cell methylome and multiomic technologies haven't been fully solved.

First, single-cell methylome technologies have scaled the size of the samples by orders of magnitudes, from dozens[5,41] to hundreds of thousands of methylomes[15,17,42] in a single study. Noteworthy, the single-cell RNA-seq (scRNA) and single-cell ATAC-seq (scATAC) also create a vast amount of samples, which are well handled by corresponding analysis packages[36,43–46]. However, packages developed for these read-count-based sequencing assays are not suitable for the read-fraction-based mC data due to fundamental differences in the data quantification and analysis goals. Therefore, algorithmic and engineering innovations are needed to store and process single-cell mC data before incorporating existing single-cell computational tools. Besides, many single-cell data analysis packages have been mainly focused on cellular analysis, for example, dimension reduction, clustering[47,48], differential gene analysis between cells[44], cell-cell integration[49], and trajectory analysis[50]. Although all these analyses are also essential for the mC data, the additional goal of studying single-cell DNA methylation is to characterize regulatory elements in the genome at single cytosine resolution. Therefore, a comprehensive analysis package bringing the cellular and genomic analysis together will maximize the utility of the informative single-cell genome-wide DNA methylation assays.

My colleagues and I describe ALLCools (ALL Cytosine tools), a versatile python package designed for single-cell mC and mC-based multi-omic assays. ALLCools defines one tab-separated text format (ALLC format) to handle raw unprocessed methylation count tables. From the ALLC table, ALLCools then generates two data structures for storing huge cell-by-region (MCDS) and region-by-annotation (RegionDS) matrices generated by recent single-cell technologies, with a scalable capacity of more than one million cells or ten million genomic regions. Besides, both data structures natively support handling quantification from multiple molecular modalities, making it suitable for storing all kinds of mC or mC-based multi-omic technology data.

For cell clustering analysis, we developed two methods tailored for different types of methylation data and feature sets, providing comprehensive support to datasets generated from a wide range of tissue sources (See below). Furthermore, ALLCools provides an analysis framework for cellular analysis and genomic analysis, including functions from cell clustering to post-clustering regulatory element analysis. ALLCools's Python-based implementation is convenient to interface with other existing single-cell packages such as Scanpy[44] or those with more specific purposes like performing multiomic analysis[48,51], cell-cell integration[37,52], enhancer prediction[53], and enhancer gene interaction[54].

To demonstrate the utilities of ALLCools, I used an integrated mouse hippocampus dataset to demonstrate the integrative pipeline of identifying cell-type-specific Cis-Regulatory Elements (CRE) using the combination of DNA methylation and chromatin accessibility data, and then further link them to potential target genes with the chromatin conformation. This example showcases the power of single-cell methylome and multiome datasets in characterizing the complex cell-type-specific gene regulation in adult neurons. Together, the ALLCools package is a full-stack solution for single-cell DNA methylation and multi-omic data analysis. Detailed documentation of the ALLCools package is available in https://lhqing.github.io/ALLCools/intro.html.

CHAPTER 2

DNA Methylation Atlas of the Mouse Brain at Single-Cell Resolution

2.1 Abstract

Mammalian brain cells show remarkable diversity in gene expression, anatomy and function, yet the regulatory DNA landscape underlying this extensive heterogeneity is poorly understood. Here we carry out a comprehensive assessment of the epigenomes of mouse brain cell types by applying single-nucleus DNA methylation sequencing[3,28] to profile 103,982 nuclei (including 95,815 neurons and 8,167 non-neuronal cells) from 45 regions of the mouse cortex, hippocampus, striatum, pallidum and olfactory areas. We identified 161 cell clusters with distinct spatial locations and projection targets. We constructed taxonomies of these epigenetic types, annotated with signature genes, regulatory elements and transcription factors. These features indicate the potential regulatory landscape supporting the assignment of putative cell types and reveal repetitive usage of regulators in excitatory and inhibitory cells for determining subtypes. The DNA methylation landscape of excitatory neurons in the cortex and hippocampus varied continuously along spatial gradients. Using this deep dataset, we constructed an artificial neural network model that precisely predicts single neuron cell-type identity and brain area spatial location. Integration of high-resolution DNA methylomes with single-nucleus chromatin accessibility data[29] enabled prediction of high-confidence enhancer–gene interactions for all identified cell types, which were subsequently validated by cell-type-specific chromatin conformation capture experiments[24]. By combining multi-omic datasets (DNA methylation, chromatin contacts, and open chromatin) from single nuclei and annotating the regulatory genome

of hundreds of cell types in the mouse brain, our DNA methylation atlas establishes the epigenetic basis for neuronal diversity and spatial organization throughout the mouse cerebrum.

2.2 Single-cell DNA methylome atlas

We used snmC-seq2[28] to profile genome-wide 5mC at single-cell resolution (Fig. 2.1a) across the cortex, HIP, striatum and pallidum (or cerebral nuclei, CNU), and olfactory areas (OLF) (Fig. 2.2.1b–e) using adult male C57BL/6 mice[55]. In total, we analysed 45 dissected regions in two replicates (Supplementary Fig. 2.1, Supplementary Table 2.2). Fluorescence-activated nuclei sorting (FANS) of antibody-labelled nuclei was applied to capture NeuN-positive neurons ($NeuN^+$, 92% of neurons), while also sampling a smaller number of NeuN-negative ($NeuN^-$, 8% of neurons) non-neuronal cells (Fig. 2.1a). In total, we profiled the DNA methylomes of 103,982 single nuclei, yielding, on average, 1.5 million stringently filtered reads per cell ($1.5 \times 10^6 \pm 0.58 \times 10^6$, mean ± s.d.) covering 6.2 ± 2.6% of the cytosines in the mouse genome in each cell. These enabled reliable quantification of the DNA methylation fraction for 25,905 ± 1,090 (95 ± 4%) 100-kb bins and 44,944 ± 4,438 (81 ± 8%) gene bodies (Supplementary Fig. 2.2a). The global methylation levels range from 0.2% to 7.6% in non-CpG sites and 61.6% to 88.8% in CpG sites (Supplementary Fig. 2.2b, c).

On the basis of the mCH and mCG profiles in 100-kb bins throughout the genome, we performed a three-level iterative clustering analysis to categorize the epigenomic cell populations (Fig. 2.1f, g). After quality control and preprocessing (Methods), in the first level (cell class), we clustered 103,982 cells as 67,472 (65%) excitatory neurons, 28,343 (27%) inhibitory neurons, and 8,167 (8%) non-neurons (Supplementary Table 2.3). The second round of iterative analysis of each cell class identified 41 cell major types in total (cluster size range 95–11,919), and the third round separated these major types further into 161 cell subtypes (cluster size range 12–6,551). All

subtypes are highly conserved across replicates, and replicates from the same brain region are co-clustered compared with samples from other brain regions (Supplementary Fig. 2.2d–g).

The spatial distribution of each cell type is assessed based on where the cells were dissected (Supplementary Table 2.5). Here we used uniform manifold approximation and projection (UMAP)[56] to visualize cell spatial locations (Fig. 2.1f, Supplementary Fig. 2.3) and major cell types (Supplementary Fig. 2.2h). Major non-neuronal cell types have a similar distribution across brain regions (Supplementary Fig. 2.1g), except adult neuron progenitors (ANP). We found two subtypes of ANPs, presumably corresponding to neuronal precursors in the subgranular zone of the dentate gyrus (DG)[57] (ANP anp-dg) and the rostral migratory stream[57] in CNU and OLF (ANP anp-olf-cnu). Excitatory neurons from isocortex, OLF and HIP formed different major types, with some exceptions, potentially owing to overlaps in dissected regions (Supplementary Table 2.2). Cells from the isocortex were further separated on the basis of their projection types[3,35,58]. The intratelencephalic (IT) neurons from all cortical regions contain four major types corresponding to the laminar layers (L2/3, L4, L5 and L6), each of which includes cells from all cortical regions, except L4, which lack cells from the prefrontal cortex (PFC) and anterior cingulate area (ACA). Excitatory neurons from the HIP were further partitioned into major types corresponding to DG granule cells and different subfields of cornus ammonis (CA). We also identified major types from cortical subplate structures, including the claustrum (CLA) and endopiriform nucleus (EP) from isocortex and OLF dissections. GABAergic inhibitory neurons from isocortex and HIP cluster together into five major types, whereas interneurons from CNU and OLF group into nine major types.

In total, we identified 68 excitatory and 77 inhibitory subtypes (Fig. 2.2a, b, Supplementary Table 2.7). Although there is no one-to-one correspondence between subtypes and brain regions,

individual subtypes show differential regional enrichment (Fig. 2.2a, b, top right) and distinct global mCH levels, ranging from 0.98% (DG dg-all) to 4.64% (PAL-Inh Chat, an inhibitory subtype in pallidum (PAL)PAL-Inh Chat) (Fig. 2.2a, b, bottom right). Specifically, isocortical excitatory subtypes usually consist of cells majorly derived from either the sensorimotor (primary motor (MOp), secondary motor (MOs), primary somatosensory (SSp), and secondary somatosensory (SSs) cortex), medial (PFC and ACA), or frontal areas (orbital (ORB) and agranular insular (AI) area (AI)). In the OLF, excitatory cells from the anterior olfactory nucleus (AON) and main olfactory bulb (MOB) are enriched in the subtype OLF-Exc *Bmpr1b*, whereas cells from the piriform area (PIR) are relatively enriched in the other OLF-Exc subtypes. Similarly, some inhibitory subtypes in CNU and OLF also correspond to different substructures in these two regions (Supplementary Note 2.1), indicating substantial spatial-related methylation diversity among CNU and OLF interneurons. By contrast, most caudal (CGE) or medial (MGE) ganglionic eminence-derived inhibitory subtypes contain cells derived predominantly from all cortical or hippocampal regions. To better demonstrate the unprecedented level of neuronal subtype and spatial diversity in their DNA methylomes, we provide a web application to interactively display this information at different granularity (http://neomorph.salk.edu/omb). We also provide a detailed discussion of how exemplified subtypes correspond to cell types with known functional and spatial features (Supplementary Fig. 2.4) in Supplementary Note 2.1.

2.3 Consensus epigenomic profiles

Integrating single-cell datasets collected using different molecular profiling modalities can help to establish a consensus cell-type atlas[39,58]. By integrating the methylome data with the chromatin accessibility data profiled using snATAC-seq on the same brain samples from a parallel study[29], the two modalities validated each other at the subtype level (Fig. 2.2c, d, Supplementary

Fig. 2.5a–f, Supplementary Table 2.10). We then calculated overlap scores between the original methylation subtypes (m-types) and the chromatin accessibility subtypes (a-types), which further quantified the matching of subtypes between the two modalities (Fig. 2.2e, Supplementary Fig. 2.5e, Methods). Moreover, the mCG DMRs (see below) highly overlap with open chromatin peaks in the hippocampal subtypes (Fig. 2.2f). Their mCG fractions and chromatin accessibility levels show similar cell-type-specificity across hippocampal subtypes, confirming the correct match of cell-type identities (Supplementary Fig. 2.5f).

2.4 Projection specificity of ET-L5 neurons

To further infer the projection targets of cell subtypes, we integrated our extra-telencephalic (ET) L5 neurons with Epi-Retro-Seq data[42]. Epi-Retro-Seq uses retrograde viral labelling to select neurons projecting to specific brain regions, followed by methylome analysis of their epigenetic subtypes. Cells from the same brain region of the two datasets are colocalized on *t*-distributed stochastic neighbour embedding (*t*-SNE) analysis, validating the subtypes' spatial distribution (Fig. 2.2g–i, Supplementary Fig. 2.5g–i). The overlap scores between unbiased (snmC-seq2) and targeted (Epi-Retro-Seq) profiling experiments (Supplementary Fig. 2.5j) indicate that some subtypes identified from the same cortical area show different projection specificity. For example, SSp and MOp neurons were mainly enriched in three subtypes marked by *Kcnh1*, *Tmtc2* and *Nectin1*, respectively. However, neurons projecting to the medulla in the MOp and SSp only integrate with the subtype marked with *Kcnh1* (Fig. 2.2j), suggesting that the subtypes identified in unbiased methylome profiling have distinct projection specificities.

2.5 Regulatory taxonomy of neuronal subtypes

Having developed a consensus map of cell types based on their DNA methylomes, we identified 16,451 differentially CH-methylated (where H correspond to A, T or C) genes (CH-

DMGs) and 3.9 million CG-differentially methylated regions (CG-DMRs, 624 ± 176 base pairs (bp) mean ± s.d.) between the subtypes (Supplementary Fig. 2.6, Methods, Supplementary Note 2.2). snmC-seq2 captures both cell-type-specific gene expression and predicted regulatory events[3,28]. Specifically, both gene body mCH and mCG negatively correlate with gene expression in neurons, with mCH showing a stronger correlation than mCG[2–4,10]. CG-DMRs provide predictions about cell-type-specific regulatory elements and transcription factors whose motifs enriched in these CG-DMRs predict the crucial regulators of the cell type[3–5].

To further explore the gene regulatory relationship between neuronal subtypes, we constructed taxonomy trees for excitatory and inhibitory subtypes, based on gene body mCH of CH-DMGs (Supplementary Fig. 2.7a, b, Methods). The dendrogram structures represent the similarities between these discrete subtypes and may reflect the developmental history of neuronal type specification[35,59]. Next, we used both CH-DMGs and CG-DMRs to annotate the tree and explore the features specifying cell subtypes (excitatory in Fig. 2.3a–c and inhibitory in Supplementary Fig. 2.7c, d). Specifically, we calculated a branch-specific methylation impact score for each gene or transcription factor motif that summarizes all of the pairwise comparisons related to that branch (Supplementary Fig. 2.7e; Methods). The impact score ranges from 0 to 1, with a higher score predicting stronger functional relevance to the branch. We assign 6,038 unique genes to branches within the excitatory taxonomy (5,975 in inhibitory taxonomy), including 406 transcription factor genes (412 in inhibitory taxonomy) using genes with impact scores greater than 0.3. For example, motifs from the ROR (also known as NR1F) family were assigned to the branch that separates superficial layer IT neurons from deeper layer IT neurons (Fig. 2.3d–f, node 9), whereas motifs from the CUX family were assigned to the IT-L2/3 branch, separating it from IT-L4/5 neurons (Fig. 2.3d–f, node 11). Both of these families contain members, such as *Cux1*,

Cux2 and *Rorb*, that show laminar expression in the corresponding layers and regulate cortical layer differentiation during development[35].

After impact score assignment, each branch of this taxonomy was associated with multiple transcription factor genes and motifs, which potentially function in combination to shape cell-type identities[60] (Fig. 2.3e, f). For example, we focused on two brain structures of interest: the CLA and the EP[61,62]. At the major-cell-type level, distinct clusters are marked by *Npsr1* (EP) and *B3gat2* (CLA). The known EP and CLA marker transcription factor *Nr4a2*[61] also shows hypomethylation in both clusters compared to other clusters. Accordingly, the NR4A2 motif is also associated with a branch that splits CLA neurons from IT-L6 neurons (Fig. 2.3d–f, node 6). On another branch separating EP from CLA and IT-L6 neurons, genes for several transcription factors, including NF-1 family members *Nfia* and *Nfib* and the RFX family member *Rfx3*, together with corresponding motifs (Fig. 2.3d–f, node 5) rank near the top. Our findings suggest that these transcription factors may function together with *Nr4a2*, potentially separating EP neurons from CLA and IT-L6 neurons.

Beyond identifying specific cell-subtype characteristics, we derived total impact (TI) scores to summarize the methylation variation of genes and motifs to understand their relative importance in cell type diversification and function (Supplementary Fig. 2.7f–k, Supplementary Note 2.3). By comparing the TI scores of genes and motifs calculated from the inhibitory and excitatory taxonomies, we found that there were more transcription factor genes and motifs having large TI scores in both cell classes than in either one or the other (Supplementary Fig. 2.7f, i). For instance, *Bcl11b* distinguishes OLF-Exc and IT neurons in the excitatory lineage and distinguishes CGE-*Lamp5* and CGE-*Vip* in the inhibitory lineage. Similarly, *Satb1* separates IT–L4 from IT-L2/3 and MGE from CGE in excitatory and inhibitory cells. These findings indicate broad

repurposing of transcription factors for cell-type specification among distinct developmental lineages.

2.6 Enhancer–gene Interactions

To systematically identify enhancer-like regions in specific cell types, we predicted enhancer-DMRs (eDMR) by integrating matched DNA methylome and chromatin accessibility profiles[29] of 161 subtypes (Fig. 2.4a, Methods). We identified 1,612,198 eDMR (34% of CG-DMRs), 73% of which overlapped with separately identified snATAC-seq peaks (Fig. 2.4b). Fetal-enhancer DMRs (feDMR) (that is, eDMRs between development time points) of forebrain bulk tissues[5] show high (88%) overlap with eDMRs. Surprisingly, the eDMRs also cover 74% of the feDMRs from other fetal tissues[5], indicating extensive reuse of enhancer-like regulatory elements across mammalian tissue types (Fig. 2.4b).

Next, we examined the relationship between the cell-type-signature genes and their potential regulatory elements. We calculated the partial correlation between all DMG–DMR pairs within 1 Mb distance using methylation levels across 145 neuronal subtypes (Methods). We identified a total of 1,038,853 (64%) eDMRs that correlated with at least one gene (correlation >0.3 with empirical P <0.005, two-sided permutation test, Supplementary Fig. 2.8a). Notably, for those strongly positive-correlated DMR–DMG pairs (correlation >0.5), the DMRs are largely (63%) within 100 kb of the transcription start sites (TSSs) of the corresponding genes but are depleted from ±1 kb (Fig. 2.4c, Supplementary Fig. 2.8b), whereas for the negatively correlated DMR–DMG pairs, only 11% of DMRs are found within 100 kb of the TSS (Supplementary Fig. 2.8c).

Using the gene–enhancer interactions predicted by this correlation analysis, we assigned eDMRs to their target genes. The percentages of feDMR-overlapping eDMRs vary markedly

among genes (Fig. 2.4d, Supplementary Fig. 2.8d, e). Of note, DMRs assigned to the same gene show different mCG specificity among subtypes. For example, *Tle4*-correlated eDMR could be partitioned into three groups (Supplementary Fig. 2.8e–g). One group (G2) of elements that displayed little diversity in bulk data showed highly specific mCG and open-chromatin signals in MSN–D1/D2 neurons, whereas another group (G3) was specific to CT–L6 neurons. These two groups of DMRs suggest that possible alternative regulatory elements are used to regulate the same gene in different cell types, although further experiments are required to validate this hypothesis.

Together, these analyses allow us to carefully chart the specificity of regulatory elements identified in bulk tissues to the subtype level. Besides, we identified many regulatory elements that show more restricted specificity (for example, eDMRs correlated with *Tle4* in MSN-D1/D2), providing abundant candidates for further pursuing enhancer-driven adeno-associated viruses (AAVs) that target highly specific cell types[63].

2.7 3D genome structure of hippocampus

Distal enhancers typically regulate gene expression through physical interaction with promoters[64]. Therefore, to examine whether physical chromatin contacts support our correlation-based predictions of enhancer–gene associations, we generated sn-m3C-seq[24] data for 5,142 single nuclei from the HIP (152,000 contacts per cell on average). We assigned these cells, on the basis of the sn-m3C-seq data, to eight major cell types based on integration with the snmC-seq2 HIP data. In total, 19,151 chromosome loops were identified in at least one of the cell types at 25-kb resolution (range from 1,173 to 12,614 chromosome loops per cell type).

Using DG and CA1 as examples, a notably higher correlation was observed between enhancers and genes at loop anchors than between random enhancer–gene pairs (Supplementary Fig. 2.8h). Reciprocally, the enhancer–gene pairs showing stronger correlation with methylation

were more likely to be found linked by chromosome loops or within the same looping region (Supplementary Fig. 2.8i). We also compared the concordance of methylation patterns between genes and enhancers linked by different methods and found the pairs linked by loop anchors or closest genes had the highest correlation of methylation (Supplementary Fig. 2.8j). Together, these analyses validate the physical proximity of enhancer–gene pairs predicted by our correlation-based method in specific cell types.

Additionally, we observed significant cell-type-specific 3D genome structures. The major cell types could be distinguished on UMAP embedding on the basis of chromosome interaction (Fig. 2.4e), indicating the dynamic nature of genome architecture across cell types. Among the 19,151 chromosome loops, 48.7% showed significantly different contact frequency between cell types (Fig. 2.4f). eDMRs were highly enriched at these differential loop anchors (Supplementary Fig. 2.8k). mCG levels at distal *cis*-elements are typically anti-correlated with enhancer activity[5]. Thus, we hypothesized that enhancers at differential loop anchors might also be hypomethylated in the corresponding cell type. Indeed, using the loops identified in DG and CA1 as examples, we observed that enhancers at the anchor of cell-type-specific loops show corresponding hypomethylation in the same cell type that the loop is specific to (Supplementary Fig. 2.8l).

Many differential loops were observed near marker genes of the corresponding cell type. For example, *Foxp1*, a gene for a CA1-specific transcription factor[65], has chromosome loops surrounding its gene body in CA1 but not DG (Fig. 2.4g, h). eDMRs and open chromatin were observed at these loop anchors. Notably, three loops in CA1 anchored at the TSS of the same transcript of *Foxp1* (Fig. 2.4h). Stronger demethylation and chromatin accessibility were also observed at the same transcript than in other transcripts (Fig. 2.4h, box E). These epigenetic patterns might suggest a specific transcript of *Foxp1* (Foxp1-225) is selectively activated in CA1.

by contrast, *Lrrtm4*, encoding a DG specific presynaptic protein that mediates excitatory synapse development[66], shows extensive looping to distal elements in DG but not CA1 (Supplementary Fig. 2.8n). Notably, among 34 genes showing alternative loop usage, 20 genes expressed in both DG and CA1[67]; for example, the TSS of *Grm7* interacts with an upstream enhancer in DG and gene body enhancers in CA1 (Supplementary Fig. 2.8o).

2.8 mC Gradients in IT neurons

Cortical excitatory IT neurons are classified into major types corresponding to their laminar layers: L2/3, L4, L5 and L6 (Fig. 2.5a). In agreement with the anti-correlation between transcript levels and DNA methylation, we found hypomethylation in IT neurons of the layer marker genes[35] (Supplementary Fig. 2.9a). Furthermore, UMAP embedding (Fig. 2.5a) reveals a continuous gradient of IT neurons resembling the medial–lateral distribution of the cortical regions (Fig. 2.5b), strongly suggesting that the arealization information is well preserved in the DNA methylome.

To systematically explore the spatial gradient of DNA methylation, we merged the cells into spatial groups on the basis of their cortical layer and region and generated a taxonomy between them (Methods). The taxonomy split the cells into four layer groups, followed by cortical-region separation within each layer (Supplementary Fig. 2.9c), providing a clear structure for investigating layer-related or region-related methylation variation. Specifically, the layer-related transcription factors included many known laminar marker genes and their DNA-binding motifs (Supplementary Fig. 2.9d), whereas some also show regional specific methylation differences. For example, *Cux1*, encoding a homeobox transcription factor specific to L2/3 and L4 neurons, is hypomethylated in motor (MO) and somatosensory (SS) cortex, but is hypermethylated in L2/3 of other regions, in agreement with patterns from in situ hybridization[68]. *Cux2*, which encodes another homeobox transcription factor, does not show the same regional specificity

(Supplementary Fig. 2.9a). We also identified genes for many additional transcription factors that showed cortical region specificity (Fig. 2.5c, Supplementary Fig. 2.9e). For example, *Etv6* is only hypomethylated in medial dissection regions across layers, whereas *Zic4* is hypermethylated in those regions. By contrast, *Rora* shows an anterior–posterior methylation gradient within the L4 and L5 cells. Together, these observed methylome spatial gradients demonstrated the value of our dataset for further exploring the cortical arealization with cell-type resolution.

2.9 mC gradients in DG granule cells

Global methylation gradients are observed within large cell types. For example, DG granule cells were continuously distributed in the UMAP embedding from low to high global mCH and mCG (Fig. 2.5d, global mCH fraction 0.5–1.9%, mCG fraction 69–79%). This gradient correlated with the anterior–posterior position of brain sections. Granule cells from the most posterior DG regions had higher global methylation than cells from anterior regions (Fig. 2.5d).

mCH accumulates throughout the genome during postnatal brain development[2,5]. We reasoned that DG granule cells, which are continuously replenished by ongoing neurogenesis throughout the lifespan, may accumulate mCH during their post-mitotic maturation. If so, global mCH should correlate with the age and maturity of granule cells. To investigate this, we divided DG granule cells into four groups on the basis of their global mCH levels and investigated regions of differential methylation between the groups. We identified 219,498 gradient CG-DMRs between the four groups, among which 139,387 showed a positive correlation with global mCH (+DMR), and 80,111 were negatively correlated (−DMR) (Fig. 2.5e). Notably, genes overlapping +DMRs or −DMRs have different annotated functions: genes enriched in +DMRs (+DMRgenes, $n = 328$) were associated with developmental processes, whereas those enriched in −DMRs (−DMRgenes, $n = 112$) were related to synaptic function (Supplementary Fig. 2.10a, b).

To further test the relationship between the +DMRgenes, −DMRgenes and DG development, we examined the expression patterns of these genes across time using a single-cell RNA-seq dataset that grouped DG cells into eight cell types, along their developmental trajectory from radial glia to mature granule cells[69]. The +DMRgenes were more highly expressed in immature cell types than in mature cell types (for example, *Tcf4*; Fig. 2.5f, Supplementary Fig. 2.10c), whereas the −DMRgenes showed the reverse trend (for example, *Rfx3*; Fig. 2.5g, Supplementary Fig. 2.10d). These results are consistent with the hypothesis that young DG granule cells have low global mCH and low methylation at genes associated with neural precursors. Conversely, older DG granule cells accumulate greater global mCH and have low methylation at genes associated with mature neurons. Notably, the global mCH levels also correlate with the brain dissections (Fig. 2.5d), indicating that the spatial axis can partially explain the methylation gradient (Supplementary Note 2.4).

Next, we investigated whether the global methylation level is correlated with 3D genome architecture. By plotting the chromatin interaction strength against the anchors' genomic distance, we observed a higher proportion of short-range contacts and a smaller proportion of long-range contacts in the groups with higher global mCH (Supplementary Fig. 2.10f). Although compartment strengths were not correlated with the global methylation changes (Supplementary Fig. 2.10g), the number of intra-domain contacts was positively correlated with global mCH across single cells (Supplementary Fig. 2.10h). After normalizing for the effect of decay, we found that insulation scores at domain boundaries were significantly lower in the groups with high global mCH levels (Supplementary Fig. 2.10i; all $P < 1 \times 10^{-10}$, two-sided Wilcoxon signed-rank test). Together these suggest that local structures may be more condensed over flanking regions in the high-mCH cell groups.

2.10 Cell type and spatial prediction model

To further quantify the spatial and cell-type information encoded in a single cell's DNA methylome, we built a multi-task deep ANN using cell-level methylome profiles from this study (Fig. 2.6a). Specifically, mCH levels of 100-kb bins were used to train and test the network with fivefold cross-validation (Method). The ANN predicted neuronal subtype identity and spatial location simultaneously for each testing cell with 95% and 89% accuracy, respectively (Fig. 2.6b–d). Notably, the location prediction accuracy of the ANN was higher than using only the spatial distribution information of subtypes (overall increased by 38%, Supplementary Fig. 2.11c), suggesting that spatial diversity is well-preserved in the neuronal DNA methylome. We also notice higher levels of errors in location prediction of some cell types, especially in the cortical MGE and CGE inhibitory neurons (Fig. 2.6c, Supplementary Fig. 2.11c). This finding is consistent with previous transcriptome-based studies[35,67], suggesting these neurons do not display strong cortical region specificity. Many cell-type marker genes are also enriched in features that capture most spatial information (Fig. 2.6e, f). For example, besides distinguishing CT-L6 neurons from other cell types, *Foxp2* shows notable mCH differences among dissected regions within CT-L6 (Fig. 2.6f). Notably, we also observed the moderate spatial specificity of astrocytes and oligodendrocytes using a separate model trained with methylomes of non-neuronal cells (Supplementary Note 2.5).

2.11 Discussion

In this Article, we present a single-cell DNA methylomic atlas of the mouse brain with detailed spatial dissection encompassing a large methylation dataset. This comprehensive dataset enables high-throughput cell-type classification, marker gene prediction and identification of regulatory elements. The three-level iterative clustering defined 161 subtypes representing

excitatory (68), inhibitory (77) and non-neuronal cells (16). The development of a hierarchical taxonomic architecture for cell subtypes on the basis of CH-DMGs allowed us to assign specific genes and transcription factor binding motifs to taxonomy branches using the methylation impact score. These assignments describe cell-type specificity at different levels, potentially relating to different developmental stages of each neuronal lineage. Notably, we found that transcription factor genes and their corresponding DNA-binding motifs were co-associated with the same branch in the taxonomy, providing a rich source of candidate transcription factors for future study.

Through integration with snATAC-seq[29], we matched subtypes classified in both epigenomic modalities and used the combined information to predict 1.6 million active-enhancer-like eDMRs, including 72% of cell-type-specific elements missed from previous tissue-level bulk studies[5]. To examine the associations of eDMRs and their targeting genes, we applied multi-omic methods to establish an eDMR–gene landscape using correlation-based prediction and chromatin conformation profiling using sn-m3C-seq, resulting in the identification of chromatin loops between eDMRs and their potential targeting genes in specific cell types.

Our brain-wide epigenomic dataset reveals extraordinary spatial diversity encoded in the DNA methylomes of neurons. The ANN trained on the single-cell methylome profiles accurately reproduced the detailed brain-dissection information within most subtypes, indicating the existence of large spatial methylation gradients throughout the brain. Echoing cortex development studies[70], glutamatergic neurons are regionalized by a protomap formed from an early developmental gradient of transcription factor expression. Similarly, we observed that many transcription factor genes and their corresponding DNA-binding motifs showed gradients of DNA methylation in adult IT neurons from distinct cortical regions. Additionally, we also found intra-subtype methylation gradients in DG granule cells that correlate with the spatial axis in the DG.

These gradient-related CG-DMRs are enriched in essential neurodevelopmental and synaptic genes[69,71], suggesting that these spatially resolved DNA methylation gradients reflect past regulatory events occurring during brain maturation. We qualify our findings by noting that snmC-seq2 is a sodium bisulfite-based method and cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine, which has been shown to accumulate in some brain regions[72]. New methods will be needed to simultaneously measure the full complement of cytosine base modifications at the single-cell level.

Overall, our analysis highlights the power of this dataset power for characterizing cell types using gene activity information from both coding regions and the regulatory elements in the non-coding regions of the genome. This comprehensive epigenomic dataset provides a valuable resource for answering fundamental questions about gene regulation in specifying cell-type spatial diversity and provides the raw material to develop new genetic tools for targeting specific cell types and functional testing.

2.12 METHODS

2.12.1 Mouse brain tissues

All experimental procedures using live animals were approved by the Salk Institute Animal Care and Use Committee under protocol number 18-00006. Adult (P56) C57BL/6J male mice were purchased from Jackson Laboratories and maintained in the Salk animal barrier facility on 12 h dark-light cycles with food ad libitum for a maximum of 10 days. Brains were extracted and sliced coronally at 600 μm from the frontal pole across the whole brain (for a total of 18 slices) in an ice-cold dissection buffer containing 2.5 mM KCl, 0.5 mM $CaCl_2$, 7 mM $MgCl_2$, 1.25 mM NaH_2PO_4, 110 mM sucrose, 10 mM glucose and 25 mM $NaHCO_3$. The solution was kept ice-cold and

bubbled with 95% O_2, 5% CO_2 for at least 15 min before starting the slicing procedure. Slices were kept in 12-well plates containing ice-cold dissection buffers (for a maximum of 20 min) until dissection aided by an SZX16 Olympus microscope equipped with an SDF PLAPO 1XPF objective. Olympus cellSens Dimension 1.8 was used for image acquisition. Each brain region was dissected from slices along the anterior-posterior axis according to the Allen Brain reference Atlas CCFv3[55] (see Supplementary Fig. 2.1 for the depiction of a posterior view of each coronal slice). 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, each dissected region was pooled from 6–30 animals, and two biological replicas were processed for each slice.

2.12.2 Fluorescence-activated nuclei sorting

Nuclei were isolated as previously described[2,3]. Isolated nuclei were labelled by incubation with 1:1,000 dilution of Alexa Fluor 488-conjugated anti-NeuN antibody (MAB377X, Millipore) and a 1:1,000 dilution of Hoechst 33342 at 4 °C for 1 h with continuous shaking. FANS of single nuclei was performed using a BD Influx sorter with an 85-µm nozzle at 22.5 PSI sheath pressure. Single nuclei were sorted into each well of a 384-well plate preloaded with 2 µl of proteinase K digestion buffer (1 µl M-Digestion Buffer (Zymo-Cat#D5021-9), 0.1 µl of 20 µg μl^{-1} proteinase K and 0.9 µl H_2O). The alignment of the receiving 384-well plate was performed by sorting sheath flow into wells of an empty plate and making adjustments based on the liquid drop position. Single-cell (one-drop single) mode was selected to ensure the stringency of sorting. For each 384-well plate, columns 1–22 were sorted with NeuN+ (488+) gate, and column 23-24 with NeuN− (488−) gate, reaching an 11:1 ratio of NeuN+ to NeuN- nuclei. BD Influx Software v1.2.0.142 was used to select cell populations.

2.12.3 Library preparation and Illumina sequencing

Detailed methods for bisulfite conversion and library preparation were previously described for snmC-seq2[3,28]. The snmC-seq2 and sn-m3C-seq (see below) libraries generated from mouse brain tissues were sequenced using an Illumina Novaseq 6000 instrument with S4 flow cells using the 150-bp paired-end mode. Freedom EVOware v2.7 was used for library preparation, and Illumina MiSeq control software v3.1.0.13 and NovaSeq 6000 control software v1.6.0/Real-Time Analysis (RTA) v3.4.4 were used for sequencing.

2.12.4 The sn-m3C-seq specific steps of library preparation

Single-nucleus methyl-3C sequencing (sn-m3C-seq) was performed as previously described[24]. In brief, the same batch of dissected tissue samples from the dorsal dentate gyrus (DG-1 and DG-2, Supplementary Table 2.2), ventral dentate gyrus (DG-3 and DG-4), dorsal HIP (CA-1 and CA-2), and ventral HIP (CA-3 and CA-4), were frozen in liquid nitrogen. The samples were then pulverized while frozen using a mortar and pestle, and then immediately fixed with 2% formaldehyde in DPBS for 10 min. The samples were quenched with 0.2 M glycine and stored at −80 °C until ready for further processing. After isolating nuclei as previously described[24], nuclei were digested overnight with NlaIII and ligated for 4 h. Nuclei were then stained with Hoechst 33342 (but not stained with NeuN antibody) and filtered through a 0.2-μm filter, and sorted similarly to the snmC-seq2 samples. Libraries were generated using the snmC-seq2 method.

2.12.5 Mouse brain region nomenclature

The mouse brain dissection and naming of anatomical structures in this study followed the Allen Mouse Brain common coordinate framework (CCF)[55]. On the basis of the hierarchical structure of the Allen CCF, we used a three-level spatial region organization to facilitate description: (1) the major region, for example, isocortex, HIP; (2) the sub-region, for example,

MOp, SSp, within isocortex; (3) the dissection region, for example, MOp-1 and MOp-2, within MOp. Supplementary Table 2.1 contains the full names of all abbreviations used in this study. All brain atlas images were created based on Wang et al.[55] and ©2017 Allen Institute for Brain Science. Allen Brain Reference Atlas. Available from: http://atlas.brain-map.org/.

2.12.6 Analysis stages

The following method sections were divided into three stages. The first stage, 'Mapping and feature generation', describes mapping and generating files in the single-cell methylation-specific data format. The second stage, 'Clustering related', describes clustering, identifying DMGs, or integrating other datasets, which all happened at the single-cell level. The third stage, 'Cell-type-specific regulatory elements', describes the identification of putative cell-type-specific regulatory elements using cluster-merged methylomes. Other figure-specific analysis topics may combine results from more than one stage.

2.12.7 Mapping and feature generation

2.12.7.1 Mapping and feature-count pipeline

We implemented a versatile mapping pipeline, YAP (https://hq-1.gitbook.io/mc/), for all the single-cell-methylome-based technologies developed by our group[3,21,28]. The main steps of this pipeline include: (1) demultiplexing FASTQ files into single cells; (2) reads level quality control (QC); (3) mapping; (4) BAM file processing and QC; and (5) final molecular profile generation. The details of the five steps for snmC-seq2 were previously described[28]. We mapped all of the reads to the mouse mm10 genome. We calculated the methylcytosine counts and total cytosine counts for two sets of genomic regions in each cell after mapping. Non-overlapping chromosome 100-kb bins of the mm10 genome (generated by "bedtools makewindows -w 100000") were used

for clustering analysis and ANN model training, and the gene body regions ±2 kb defined by the mouse GENCODE vm22 were used for cluster annotation and integration with other modalities.

2.12.7.2 sn-m3C-seq-specific steps or read mapping and chromatin contact analysis

Methylome sequencing reads were mapped following the TAURUS-MH pipeline, as previously described[24]. Specifically, reads were trimmed for Illumina adaptors, and then an additional 10 bp was trimmed on both sides. Then R1 and R2 reads were mapped separately to the mm10 genome using Bismark with Bowtie. The unmapped reads were collected and split into shorter reads representing the first 40 bp, the last 40 bp, and the middle part of the original reads (if read length >80 bp after trimming). The split reads were mapped again using Bismark with Bowtie. The reads with MAPQ <10 were removed. The filtered bam files from split and unsplit R1 and R2 reads were deduplicated with Picard and merged into a single bam file to generate the methylation data. Methylpy (v1.4.2)[41] was used to generate an ALLC file (base-level methylation counts) from the bam file for every single cell. We paired the R1 and R2 bam files where each read-pair represents a potential contact to generate the Hi-C contact map. For generating contact files, read pairs where the two ends mapped within 1 kbp of each other, were removed.

2.12.8 Clustering-related methods

2.12.8.1 Single-cell methylome data quality control and preprocessing

Cell filtering. We filtered the cells on the basis of these main mapping metrics: (1) mCCC level <0.03; (2) overall mCG level >0.5; (3) overall mCH level <0.2; (4) total final reads >500,000; and (5) Bismark mapping rate >0.5. Other metrics such as genome coverage, PCR duplicates rate and index ratio were also generated and evaluated during filtering. However, after removing

outliers with the main metrics 1–5, few additional outliers were found. Note the mCCC level is used as the estimation of the upper bound of bisulfite non-conversion rate[3].

Feature filtering. 100 kb genomic bin features were filtered by removing bins with mean total cytosine base calls <250 (low coverage) or >3,000 (unusually high-coverage regions). Regions that overlap with the ENCODE blacklist[73] were also excluded from further analysis.

Computation and normalization of the methylation level. For CG and CH methylation, the methylation level computation from the methylcytosine and total cytosine matrices contains two steps: (1) prior estimation for the beta-binomial distribution, and (2) posterior level calculation and normalization per cell.

Step 1: for each cell, we calculated the sample mean m and variance v of the raw methylcytosine (mc) level (mc/cov), where cov is the total cytosine base coverage and mc is the methylcytosine base coveragecoefficient of variance, for each sequence context (CG or CH). The shape parameters (α, β) of the beta distribution were then estimated using the method of moments:

$$\alpha = m\left(m\left(1-m\right)/v-1\right)$$

$$\beta = \left(1-m\right)\left(m\left(1-m\right)/v-1\right)$$

This approach used different priors for different methylation types for each cell and used weaker beforepriors to cells with more information (higher raw variance).

Step 2: we then calculated the posterior: $\hat{\text{mc}} = \dfrac{\alpha + \text{mc}}{\alpha + \beta + \text{cov}}$ for all bins in each cell. Like the counts per million bases reads (CPM) normalization in the single-cell RNA-seq analysis,

we normalized this posterior methylation ratio by the cell's global mean methylation, $m = \alpha/(\alpha + \beta)$. Thus, all the posterior $\widehat{mc}$ values with 0 cov will have a constant value of 1 after normalization. The resulting normalized mc level matrix contains no NA (not available) value, and features with lower cov tend to have a mean value close to 1.

Selection of highly variable features. Highly variable methylation features were selected with a modified approach using the scanpy.pp.highly_variable_genes function from the scanpy 1.4.3 package[44]. In brief, the scanpy.pp.highly_variable_genes function normalized the dispersion of a gene by scaling with the mean and standard deviation of the dispersions for genes falling into a given bin for mean expression of genes. In our modified approach, we reasoned that both the mean methylation level and the mean cov of a feature (100 kb bin or gene) could impact mc level dispersion. We grouped features that fall into a combined bin of mean and cov. We then normalized the dispersion within each mean–cov group. After dispersion normalization, we selected the top 3,000 features based on normalized dispersion for clustering analysis.

Dimension reduction and combination of different mC types. For each selected feature, mc levels were scaled to unit variance and zero mean. We then performed principal component analysis (PCA) on the scaled mc level matrix. The number of principal components (PCs) was selected by inspecting the variance ratio of each PC using the elbow method. The CH and CG PCs were then concatenated together for further analysis in clustering and manifold learning (Supplementary Table 2.6 for parameters of PCA and clustering analysis).

2.12.8.2 Consensus clustering

Consensus clustering on concatenated PCs. We used a consensus clustering approach based on multiple Leiden clustering[74] over *k*-nearest neighbour (KNN) graph to account for the

randomness of the Leiden clustering algorithms. After selecting dominant PCs from PCA in both mCH and mCG matrices, we concatenated the PCs together to construct a KNN graph using scanpy.pp.neighbours with Euclidean distance. Given fixed resolution parameters, we repeated the Leiden clustering 300 times on the KNN graph with different random starts and combined these cluster assignments as a new feature matrix, where each single Leiden result is a feature. We then used the outlier-aware DBSCAN algorithm from the scikit-learn package to perform consensus clustering over the Leiden feature matrix using the hamming distance. Different epsilon parameters of DBSCAN are traversed to generate consensus cluster versions with the number of clusters that range from the minimum to the maximum number of clusters observed in the multiple Leiden runs. Each version contained a few outliers; these usually fall into three categories: (1) cells located between two clusters had gradient differences instead of clear borders, for example, border of IT layers; (2) cells with a low number of reads potentially lack information in essential features to determine the specific cluster; and (3) cells with a high number of reads that were potential doublets. The number of type 1 and 2 outliers depends on the resolution parameter and is discussed in the choice of the resolution parameter section. The type 3 outliers were very rare after cell filtering. The supervised model evaluation below then determined the final consensus cluster version.

Supervised model evaluation on the clustering assignment. We performed a recursive feature elimination with cross-validation (RFECV)[75] process from the scikit-learn package to evaluate clustering reproducibility for each consensus clustering version. We first removed the outliers from this process, and then we held out 10% of the cells as the final testing dataset. For the remaining 90% of the cells, we used tenfold cross-validation to train a multiclass prediction model using the input PCs as features and sklearn.metrics.balanced_accuracy_score[76] as an

evaluation score. The multiclass prediction model is based on BalancedRandomForestClassifier from the imblearn package, which accounts for imbalanced classification problems[77]. After training, we used the 10% testing dataset to test the model performance using the score from balanced_accuracy_score. We kept the best model and corresponding clustering assignments as the final clustering version. Finally, we used this prediction model to predict outliers' cluster assignments. We rescued the outlier with prediction probability >0.3, otherwise labelling them as outliers.

Manifold learning for visualization. In each round of clustering analysis, the *t*-SNE[78,79] and UMAP[56] embedding were run on the PC matrix the same as the clustering input using the implementation from the scanpy[44] package. The coordinates from both algorithms were in Supplementary Table 2.5.

Choice of resolution parameter. Choosing the resolution parameter of the Leiden algorithm is critical for determining the final number of clusters. We selected the resolution parameter by three criteria: (1). the portion of outliers <0.05 in the final consensus clustering version; (2) the ultimate prediction model accuracy >0.9; and (3) the average cell per cluster ≥ 30, which controls the cluster size to reach the minimum coverage required for further epigenome analysis such as DMR calls. All three criteria prevented the over-splitting of the clusters; thus, we selected the maximum resolution parameter under meeting the criteria using a grid search.

Cell class (level 1 clustering) annotation. We annotated non-neuron cells based on both the NeuN- gate origin and low global mCH fraction. Given the strong anti-correlation between CH methylation and gene expression, we used hypo-CH-methylation at gene bodies ±2 kb of pan-excitatory markers such as *Slc17a7* and *Sv2b*, and pan-inhibitory markers such as *Gad1* and *Gad2* to annotate excitatory and inhibitory cell classes, respectively.

Major type (level 2) and subtypes (level 3) annotations. We used both gene body ±2 kb hypo-CH-methylation (or hypo-CG-methylation for non-neurons) of well-known marker genes and the dissection information to annotate neuron and non-neuron clusters. All cluster marker genes are listed in Supplementary Table 2.7, together with the description of the cluster names, references to the marker gene information, and the URL to the data browser. The major cell types were annotated based on well-known marker genes reported in the previous studies[3,35,67,80–82]. Whenever possible, we name these clusters with canonical names (for example, IT-L23, L6b) or using descriptive names that reflect the specific spatial location of the cluster (for example, EP, CLA, IG-CA2). For subtypes, we named the clusters via its parent major type name followed by a subtype marker gene name.

2.12.8.3 Pairwise DMG identification

We used a pairwise strategy to calculate DMGs for each pair of clusters within the same round of analysis. We used the gene body ±2 kb regions of all the protein-coding and long non-coding RNA genes with evidence level 1 or 2 from the mouse GENCODE vm22. We used the single-cell level mCH fraction normalized by the global mCH level (as in 'Computation and normalization of the methylation level' in the clustering step above) to calculate markers between all neuronal clusters. We compared non-neuron clusters separately using the mCG fraction normalized by the global mCG level. For each pairwise comparison, we used the Wilcoxon rank-sum test to select genes with a significant decrease (hypo-methylation). Marker genes were chosen based on adjusted $P < 10^{-3}$ with multitest correction using the Benjamini–Hochberg procedure, delta-normalized methylation level change <−0.5 (hypo-methylation) and area under the receiver-operating curve (AUROC) >0.8. We required each cluster to have ≥5 DMGs compared to any other cluster. Otherwise, the smallest cluster that did not meet this criterion was merged to the

closest cluster based on Euclidean distance between cluster centroids in the PC matrix used for clustering. Then the marker identification process was repeated until all clusters found enough marker genes.

2.12.8.4 Three levels of iterative clustering analysis

On the basis of the consensus clustering steps described above, we used an iterative approach to cluster the data into three levels of categories. In the first level, termed CellClass, clustering analysis is done using all cells and then manually merged into three canonical classes: excitatory neurons, inhibitory neurons, and non-neurons based on marker genes. Within each CellClass, we performed all the preprocessing and clustering steps again to obtain clusters for the MajorType level using the same stop criteria. Furthermore, within each MajorType, we obtained clusters for the SubType level. All clusters' annotations and relationships are presented in Supplementary Table 2.7.

2.12.8.5 Subtype taxonomy tree

To build the taxonomy tree of subtypes, we selected the top 50 genes that showed the most significant changes for each subtypes' pairwise comparisons. We then used the union of these genes from all subtypes and obtained 2,503 unique genes. We calculated the median mCH level of these genes in each subtype and applied bootstrap resampling-based hierarchical clustering with average linkage and the correlation metric using the R package pvclust (v.2.2)[83].

2.12.8.6 Impact score and total impact score

We defined the impact score (IS) to summarize pairwise comparisons for two subtype groups, where one group, A, contains M clusters and the other group, B, contains N clusters. For each gene or motif, the number of total related pairwise comparisons is $M \times N$, the number of

significant comparisons with desired change (hypo-methylation for gene or enrichment for motif) is a in group A and b in group B. The IS is then calculated as $\mathrm{IS_A} = \frac{a-b}{M \times N}$ and $\mathrm{IS_B} = \frac{b-a}{M \times N}$ for the two directions. For either group, IS ranges from −1 to 1, and 0 means no impact, 1 means full impact and −1 means full impact in the other group (Supplementary Fig. 2.7e).

We explored two scenarios using the IS to describe cluster characteristics (Supplementary Fig. 2.7e). The first scenario is considering each pair of branches in the subtype taxonomy tree as comprising group A and group B. Thus, the IS can quantify and rank genes or motifs to the upper nodes based on the leaves' pairwise comparisons (Fig. 2.3d–f). The second scenario summarizes the total impact for specific genes or motifs regarding the taxonomy tree based on the calculation in the first scenario (Supplementary Fig. 2.7f–k). In a subtype taxonomy tree with n subtypes, the total non-singleton node was $n - 1$, and each node i had a height h_i and associated $\mathrm{IS_A}$ for one of the branches ($\mathrm{IS_B} = -\mathrm{IS_A}$). The node-height-weighted total IS ($\mathrm{IS_{total}}$) was then calculated as:

$$\mathrm{IS_{total}} = \sum_{i=1}^{n-1} h_i \times \left|\mathrm{IS_A}\right|$$

The larger total IS indicated that a gene or motif shows more cell-type-taxonomy-related significant changes. The total IS can also be calculated in a sub-tree or any combination of interests to rank genes and motifs most related to that combination (See 'Figure-specific methods' for Fig. 2.5 regarding calculating layer and region total IS from the same tree).

2.12.8.7 Integration with snATAC-seq data

A portion of the same brain tissue sample used in this study for methylome profiling was also processed using snATAC-seq in a parallel study of chromatin accessibility[29]. The final high-quality snATAC-seq cells were assigned to 160 chromatin accessibility clusters (a-types). The snATAC-seq-specific data analysis steps are described in Li et al.[29]. Here, we performed cross-modality data integration and label-transferring to assign the 160 a-types to the 161 methylome subtypes in the following steps:

(1) We manually grouped both modalities into five integration groups (for example, all IT neurons as a group) and only performed the integration of cells within the same group to decrease computation time. These groups were distinct in the clustering steps of both modalities and can be matched with great confidence using known marker genes. Steps 2–6 were repeated for each group. See Supplementary Fig. 2.5 for the group design.

(2) We used a similar approach as described above to identify pairwise differential accessible genes (DAGs) between all pairs of a-types. The cut-off for DAG is adjusted $P < 10^{-3}$, fold change >2 and AUROC >0.8.

(3) We then gathered DMGs from comparisons of related subtypes in the same group. Both DAGs and DMGs were filtered according to whether they recurred in >5 pairwise comparisons. The intersection of the remaining genes was used as the feature set of integration.

(4) After identifying DAGs using cell-level snATAC-seq data, we merged the snATAC-seq cells into pseudo-cells to increase snATAC-seq data coverage. Within each a-type, we did a *k*-means clustering (k = no. of cells in that cluster/50) on the same PCs used in snATAC-seq clustering. We discarded small *k*-means clusters with less than 10 cells (about 5% of the cells) and

merged each remaining *k*-means cluster into a pseudo-cell. On average, a pseudo-cell had about 50 times more fragments than a single cell.

(5) We then used the MNN based Scanorama[37] method with default parameters to integrate the snmC-seq cells and snATAC-seq pseudo-cells using genes from step 3. After Scanorama integration, we did co-clustering on the integrated PC matrix using the clustering approaches described above.

(6) We used the intermediate clustering assignment from step 5 to calculate the overlap score (below) between the original methylome subtypes and the a-types. We used overlap score >0.3 to assign a-types to each methylome subtype. For those subtypes that have no match under this threshold, we assigned the top a-type ranked by the overlap score (Supplementary Tables 2.10, 2.11).

2.12.8.8 Overlap score

We used the overlap score to match a-type and methylome subtypes. The overlap score, range from 0 to 1, was defined as the sum of the minimum proportion of samples in each cluster overlapped within each co-cluster[84]. A higher score between one methylome subtype and one a-cluster indicates they consistently co-clustered within one or more co-clusters. Besides matching clusters in integration analysis, the overlap score (OS) was also used in two other cases: (1) to quantify replicates and region overlaps over methylome subtypes (Supplementary Fig. 2.2e–g); and (2) to quantify the overlap of each L5-ET subtype overlapping with 'soma location' and 'projection target' labels from epi-retro-seq cells (Supplementary Fig. 2.5j) through integration with the epi-retro-seq dataset.

2.12.9 Cell-type-specific regulatory elements

2.12.9.1 DMR analysis

After clustering analysis, we used the subtype cluster assignments to merge single-cell ALLC files into the pseudo-bulk level and then used methylpy (v1.4.2)[41] DMRfind function to calculate mCG DMRs across all subtypes. The base calls of each pair of CpG sites were added before analysis. In brief, the methylpy function used a permutation-based root mean square test of goodness of fit to identify differentially methylated sites (DMS) simultaneously across all samples (subtypes in this case), and then merge the DMS within 250 bp into the DMR. We further excluded DMS calls that have low absolute mCG level differences by using a robust-mean-based approach. For each DMR merged from the DMS, we ordered all the samples by their mCG fraction and calculated the robust mean m using the samples between 25th and 75th percentiles. We then reassigned hypo-DMR and hyper-DMR to each sample when a region met two criteria: (1) the sample mCG fraction of this DMR is lower than ($m - 0.3$) for hypo-DMR or ($m + 0.3$) for hyper-DMR, and (2) the DMR is originally a significant hypo- or hyper-DMR in that sample judged by methylpy. DMRs without any hypo- or hyper-DMR assignment were excluded from further analyses. On the basis of these filtering criteria, we estimate the false discovery rate of calling DMRs is 2.7% (Supplementary Note 2.2, Supplementary Fig. 2.6).

2.12.9.2 Enhancer prediction using DNA methylation and chromatin accessibility

We performed enhancer prediction using the REPTILE[53] algorithm. REPTILE is a random-forest-based supervised method that incorporates different epigenomic profiles with base-level DNA methylation data to learn and then distinguish the epigenomic signatures of enhancers and genomic background. We trained the model in a similar way as in the previous studies[5,53], using

CG methylation, chromatin accessibility of each subtype and mouse embryonic stem cells (mouse ES cells). The model was first trained on mouse ES cell data and then predicted a quantitative score that we termed enhancer score for each subtype's DMRs. The positives were 2 kb regions centred at the summits of the top 5,000 EP300 peaks in mouse ES cells. Negatives include randomly chosen 5,000 promoters and 30,000 2-kb genomic bins. The bins have no overlap with any positive region or gene promoter[5].

Methylation and chromatin accessibility profiles in bigwig format for mouse ES cells were from the GEO database (GSM723018). The mCG fraction bigwig file was generated from subtype-merged ALLC files using the ALLCools package (https://github.com/lhqing/ALLCools). For chromatin accessibility of each subtype, we merged all fragments from snATAC-seq cells that were assigned to this subtype in the integration analysis and used deeptools bamcoverage to generate CPM normalized bigwig files. All bigwig file bin sizes were 50 bp.

2.12.9.3 Motif-enrichment analysis

We used 719 motif PWMs from the JASPAR 2020 CORE vertebrates database[85], where each motif was able to assign corresponding mouse transcription factor genes. The specific DMR sets used in each motif-enrichment analysis are described in figure specific methods below. For each set of DMRs, we standardized the region length to the centre ±250bp and used the FIMO tool from the MEME suite[86] to scan the motifs in each enhancer with the log-odds score $P < 10^{-6}$ as the threshold. To calculate motif enrichment, we use the adult non-neuronal mouse tissue DMRs[7] as background regions unless expressly noted. We subtracted enhancers in the region set from the background and then scanned the motifs in background regions using the same approach. We then used Fisher's exact test to find motifs enriched in the region set and the Benjamini–Hochberg

procedure to correct multiple tests. We used the TFClass[87] classification to group transcription factors with similar motifs.

2.12.9.4 DMR–DMG partial correlation

To calculate DMR–DMG partial correlation, we used the mCG fraction of DMRs and the mCH fraction of DMGs in each neuronal subtype. We first used linear regression to regress out variance due to global methylation difference (using scanpy.pp.regress_out function), then use the residual matrix to calculate the Pearson correlation between DMR and DMG pairs where the DMR centre is within 1 Mb of the TSSs of the DMG. We shuffled the subtype orders in both matrices and recalculated all pairs 100 times to generate the null distribution.

2.12.9.5 Identification of loops and differential loops from sn-m3C-seq data

After merging the chromatin contacts from cells belonging to the same type, we generated a .hic file of the cell-type with Juicer tools pre. HICCUPS[88] was used to identify loops in each cell type. The loops from eight major cell types were concatenated and deduplicated and used as the total samples for differential loop calling. A loop-by-cell matrix was generated, in which each element represents the number of contacts supporting each loop in each cell. The matrix was used as input of EdgeR to identify differential interactions with ANOVA tests. Loops with FDR $<10^{-5}$ and minimum–maximum fold change >2 were used as differential loops. Note that the abundance of cell types is highly variable, leading to different coverages of contact maps after merging all the cells from each cell type. Since *HICCUPS* loop calling is sensitive to the coverage, more loops were identified in the abundant cell types (for example, 12,614 loops were called in DG, containing 1,933 cells) compared to the less abundant ones (for example, 1,173 loops were called in MGE,

containing 145 cells). Therefore, we do not compare the feature counts related to the loops across cell types directly in our analyses.

2.12.10 Figure-specific methods

3D model of dissection regions (Fig. 2.1b–e). We created in silico dissection regions based on the Allen CCF[55] 3D model using Blender 2.8 that precisely follow our dissection plan. To ease visualization of all different regions, we modified the layout and removed some of the symmetric structures, but all the actual dissections were applied symmetrically to both hemispheres.

Calculating the genome feature detected ratio (Supplementary Fig. 2.2a). The detected ratio of chromosome 100-kb bins and gene bodies is calculated as the percentage of bins with >20 total cytosine coverage. Non-overlapping chromosome 100-kb bins were generated by bedtools makewindows -w 100000; gene bodies were defined by GENCODE vm22.

Integration with epi-retro-seq L5-ET cells (Fig. 2.2g–j, Supplementary Fig. 2.5g–j). Epi-retro-seq is an snmC-seq2-based method that combines retrograde AAV labelling[42]. The L5-ET cells' non-overlapping chromosome 100-kb bin matrix gathered by the epi-retro-seq dataset was concatenated with all the L5-ET cells from this study for co-clustering and embedding as described in 'Clustering-related methods'. We then calculated the OS between subtypes in this study and the 'soma location' or 'projection target' labels of epi-retro-seq cells. The first OS helped quantify how consistent the spatial location is between the two studies; the second OS allowed us to impute the projection targets of subtypes in this study.

Pairwise DMR and motif-enrichment analysis (Fig. 2.3c, f). The total subtype DMRs were identified as described in 'Cell-type-specific regulatory elements' by comparing all subtypes. We then assigned DMRs to each subtype pair if the DMRs were: (1) significantly hypomethylated

in only one of the subtypes; and (2) the mCG fraction difference between the two subtypes is >0.4. Each subtype pair was associated with two exclusive sets of pairwise DMRs. We carried out motif-enrichment analysis described in 'Cell-type-specific regulatory elements' on each DMR set using the other set as background. Motifs enriched in either direction were then used to calculate the impact score and were associated with upper nodes of the taxonomy.

Overlapping eDMR with genome regions (Fig. 2.4b). The cluster-specific snATAC-seq peaks were identified in Li et al.[29]. We used bedtools merge to aggregate the total non-overlap peak regions and bedtools intersect to calculate the overlap between peaks and eDMRs. The developing forebrain and other tissue feDMRs were identified in He et al.[5] using methylC-seq[89] for bulk whole-genome bisulfite sequencing. All of the genome features used in Fig. 2.4b were defined as in He et al.[5], except using an updated mm10 CGI region and RepeatMaster transposable elements lists (UCSC table browser downloaded on 9 October 2019).

Heat maps of the gene–enhancer landscape (Supplementary Fig. 2.8e). The eDMRs for each gene were selected by eDMR–gene correlation of >0.3. Sections of the heat maps in were gathered by (1) mCG fraction of each eDMR in 161 subtypes from this study; (2) snATAC-seq subtype-level fragments per kilobase of transcript per million mapped reads (FPKM) of each eDMR in the same subtype orders. The subtype snATAC profiles were merged from integration results as described in 'Clustering-related methods'; (3) mCG fraction of each eDMR in forebrain tissue during ten developing time points from embryonic day 10.5 (E10.5) to P0 (data from He et al.[5]); (4) H3K27ac FPKM of each eDMR in 7 developing time points from E11.5 to P0 (data from Gorkin et al.[90]); (5) H3K27ac FPKM of each eDMR in P56 frontal brain tissue (data from Lister et al.[2]); and (6) eDMR is overlapped with forebrain feDMR using bedtools intersect.

Embedding of cells with chromosome interactions (Fig. 2.4e). scHiCluster[91] was used to generate the *t*-SNE embedding of the sn-m3C-seq cells. Specifically, a contact matrix at 1-Mb resolution was generated for each chromosome of each cell. The matrices were then smoothed by linear convolution with pad = 1 and random walk with restart probability = 0.5. The top 20th percentile of strongest interactions on the smoothed map was extracted, binarized and used for PCA. The first 20 PCs were used for *t*-SNE.

IT layer dissection region group DMG and DMR analysis (Fig. 2.5a–c). To collect enough cells for dissection region analysis, we used only the major types (corresponding to L2/3, L4, L5 and L6) of IT neurons. We grouped cells into groups according to layer dissection region and kept groups with >50 cells for further analysis (Supplementary Fig. 2.9b). We performed pairwise DMG, DMR and motif-enrichment analysis, the same as the subtype analysis in Fig. 2.3, but using the layer dissection region group labels. We then built a spatial taxonomy for these groups and used it to calculate impact scores. To rank layer-related or dissection-region-related genes and motifs separately, we used two sets of the branches (Supplementary Fig. 2.9c, top set for layers, bottom set for regions) in the taxonomy and calculated two total impact scores using the equations above.

DG cell group and gradient DMR analysis (Fig. 2.5e). DG cells were grouped into four evenly sized groups according to the cells' global mCH levels, with cut-off thresholds at 0.45%, 0.55% and 0.69%. We then randomly chose 400 cells from each group to call gradient-DMRs using methods described in 'Clustering-related methods'. To ensure the DMRs identified between intra-DG groups were not due to stochasticity, we also randomly sampled 15 groups of 400 cells from all DG cells regardless of their global mCH and called DMRs among them as control-DMRs (2,003 using the same filtering condition). Only 0.04% of gradient DMRs overlapped with the

control DMRs; these were removed from further analysis. Pearson correlations (ρ) of mCG fractions of each gradient DMR was calculated against a linear sequence (1, 2, 3, 4) to quantify the gradient trend. DMRs with $\rho < -0.75$ or $\rho > 0.75$ were considered to be significantly correlated. Weakly correlated DMRs (10% of DMRs) were not included in further analysis.

DMR- and DMS-enriched genes (Fig. 2.5f, g). To investigate the correlated DMR or DMS enrichment in specific gene bodies, we compared the number of DMS and cytosine inside the gene body with the number of DMS and cytosine in the ±1 Mb regions using Fisher's exact test. We chose genes passing two criteria: (1) adjusted $P < 0.01$ with multitest correction using the Benjamini–Hochberg procedure, and (2) overlap with >20 DMSs. Gene ontology analysis of DMR and DMS enriched genes was carried out using GOATOOLS[92]. All protein-coding genes with gene body length >5 kb were used as background to prevent gene-length bias.

Compartment strength analysis (Supplementary Fig. 2.10g). We normalized the total chromosome contacts by *z*-score in each 1-Mb bin of the DG contact matrix, and the bins with normalized coverage between −1 and 2 were kept for the analysis. After filtering, the PC1 of the genome-wide Knight-RuizKR-normalized contact matrix was used as the compartment score. The score was divided into 50 categories with equal sizes from low to high, and bins were assigned to the categories. The intra-chromosomal observation/expectation (ove) matrices of each group were used to quantify the compartment strength. We computed the average ove values within each pair of categories to generate the 50 × 50 saddle matrices. The compartment strength was computed with the average of the upper left and lower right 10 × 10 matrices divided by the average of the upper right and lower left 10 × 10 matrices[93].

Domain analysis (Supplementary Fig. 2.10i). We identified 4,580 contact domains at 10-kb resolution in DG using Arrowhead[88]. For bin i, the insulation score *I* is computed by

$$I_i = \frac{\text{mean}_{i-10\le i'<i;\ i\le j'<i+10} A_{i'j'}}{\max\left(\text{mean}_{i-10\le i'<i;\ i-10\le j'<i} A_{i'j'},\ \text{mean}_{i\le i'<i+10;\ i\le j'<i+10} A_{i'j'}\right)},$$

where A is the ove of Knight-RuizKR-normalized matrices. For each group, insulation scores of domain boundaries and 100-kb flanking regions were computed and averaged across all boundaries.

2.12.11 Prediction model description

Related to Fig. 2.6. To reduce the computing complexity, we applied PCA on the dataset of 100-kb bin mCH features to obtain the first 3,000 PCs, which retain 61% of the variance of the original data. These 3,000 PCs were then used to train and test the predicting model. We used an ANN with two hidden layers to simultaneously predict cell subtypes and their dissection regions. The input layer contains 3,000 nodes, followed by a shared layer with 1,000 nodes. The shared layer is further connected simultaneously to two branch hidden layers of the subsection region's subtype, each containing 200 nodes. The corresponding one-hot encoding output layers follow branch hidden layers. We used fivefold cross-validation to access the model performance. We applied the dropout technique[94] with a dropout rate $P = 0.5$ on each hidden layer to prevent overfitting during the training. Adam optimization[95] was used to train the network with a cross-entropy loss function. The training epoch number and batch size are 10 and 100, respectively. The training and testing processes were conducted via TensorFlow 2.0[96].

Model performance. The two output layers generate two probabilistic vectors for each single cell input as the prediction results for cell subtypes and dissection regions, respectively. The subtype and dissection region label with the highest probabilities were used as the prediction results for each cell to calculate accuracy. When calculating the cell dissection region accuracy

(Fig. 2.6c), we defined two kinds of accuracy with different stringency: (1) the exact accuracy using the predicted label, and (2) the fuzzy accuracy using predicted labels or its potential overlap neighbours. The potential overlap neighbours curated based on Allen CCF (Supplementary Fig. 2.11a, Supplementary Table 2.2) stood for adjacent regions of a particular dissection region. The exact accuracy of the ANN model is 69% and the fuzzy accuracy is 89%. To evaluate how much of the dissection region accuracy was improved via ANN, we calculated fuzzy accuracy based only on naive guesses in each subtype based on the dissection region composition (grey dots in Supplementary Fig. 2.11c). We also trained additional models using logistic regression and random forest for benchmarks. The performance of ANN on subtype prediction is comparable with logistic regression and random forest. By contrast, the performance in location prediction is substantially improved against the other two models (Supplementary Fig. 2.11b), suggesting that distinguishing the cells from different dissected regions may require nonlinear relationships between genomic regions. We used scikit-learn (v0.23) for logistic regression and random forest implementation and the multinomial objective function for multi-class classification. N_estimators were set to 1,000 for the random forest.

Biological feature importance for dissection region prediction (Fig. 2.6e). To assess which DNA regions store information of cell spatial origins that is distinguishable using our model, we evaluated the importance of PC features by examining how permutation of each PC feature across cells affects prediction accuracy. We tested five permutations for each feature and used decreasing average accuracy to indicate PC feature importance. We examined genes contained in the 100-kb bins with the top 1% PCA factor loadings for the most important PC feature for a given cell type.

2.13 Data availability

Single-cell raw and processed data included in this study were deposited to NCBI Gene Expression Omnibus and Sequence Read Archive with accession number GSE132489 (each experiment has a separate accession number recorded in GSE132489; see Supplementary Table 2.13), and to the NeMO archive: https://assets.nemoarchive.org/dat-vmivr5x. Single-cell methylation data can be visualized at the Brain Cell Methylation Viewer: http://neomorph.salk.edu/omb/home. Cluster merged methylome profiles can be visualized at http://neomorph.salk.edu/mouse_brain.php. Other datasets used in the paper include single-nuclei ATAC-seq data[29] from http://catlas.org, mouse embryo forebrain development data[5] from the ENCODE portal (https://www.encodeproject.org/), the developing hippocampal single-cell RNA-seq data from GSE104323, DNA methylation and chromatin accessibility profiles for mouse ES cells from GSM723018 and the JASPAR 2020 CORE vertebrates database from http://jaspar.genereg.net/.

2.14 Code availability

The mapping pipeline for snmC-seq2 data is available at https://hq-1.gitbook.io/mc/; the ALLCools package for post-mapping analysis and snmC-seq2 related data structure are available at https://github.com/lhqing/ALLCools; the jupyter notebooks for reproducing specific analysis are at https://github.com/lhqing/mouse_brain_2020; and the source code of the Brain Cell Methylation Viewer is at https://github.com/lhqing/omb.

2.15 Author Contributions

J.R.E., H.L., B.R., M.M.B., C.L. and J.R.D. conceived the study. H.L., J.Z. and W.T. analysed the snmC-seq data and drafted the manuscript. J.R.E., C.L., E.A.M., J.R.D. and M.M.B. edited the manuscript. J.R.E., H.L., M.M.B., A.B., J.L. and S.P. coordinated the research. M.M.B.,

A.B., A.A., H.L., J.L., J.R.N., A.R., J.K.O., A.P.-D., C.O., L.B., C.F., C.L. and J.R.E. generated the snmC-seq2 data. J.R.D., B.C., A.B., J.L., J.Z., A.A., J.K.O., C.L., J.R.N., C.O., L.B., C.F., R.G.C., M.M.B. and J.R.E. generated the sn-m3C-seq data. S.P., M.M.B., X.H., J.L., O.B.P., Y.E.L., J.K.O. and B.R. generated the snATAC-seq data. Z.Z., J.Z., E.M.C., M.M.B., J.R.E., A.B., A.A., J.R.N., C.O., L.B., C.F., R.G.C. and A.R. generated the Epi-Retro-Seq data. H.L., H.C., E.A.M., M.N. and C.L. contributed to data archive/infrastructure. J.R.E. supervised the study.

2.16 Acknowledgements

Chapter 2, in full, is a reprint of the material as it appears in Nature 2021. "DNA methylation atlas of the mouse brain at single-cell resolution," Hanqing Liu, Jingtian Zhou, Wei Tian, Chongyuan Luo, Anna Bartlett, Andrew Aldridge, Jacinta Lucero, Julia K. Osteen, Joseph R. Nery, Huaming Chen, Angeline Rivkin, Rosa G. Castanon, Ben Clock, Yang Eric Li, Xiaomeng Hou, Olivier B. Poirion, Sebastian Preissl, Antonio Pinto-Duarte, Carolyn O'Connor, Lara Boggeman, Conor Fitzpatrick, Michael Nunn, Eran A. Mukamel, Zhuzhu Zhang, Edward M. Callaway, Bing Ren, Jesse R. Dixon, M. Margarita Behrens, and Joseph R. Ecker. The authors thank Yupeng He for the advice on the methylpy and REPTILE analysis; Terrence Sejnowski for the advice on the ANN analysis. This project is supported by NIMH U19MH11483 to Joseph Ecker and Edward Callaway and NHGRI R01HG010634 to Joseph R Ecker and Jesse Dixon. The Flow Cytometry Core Facility of the Salk Institute is supported by funding from NIH-NCI CCSG: P30 014195 and Shared Instrumentation Grant S10-OD023689. Joseph Ecker is an investigator of the Howard Hughes Medical Institute. The dissertation author was the primary investigator and author of this paper.

2.17 Figures

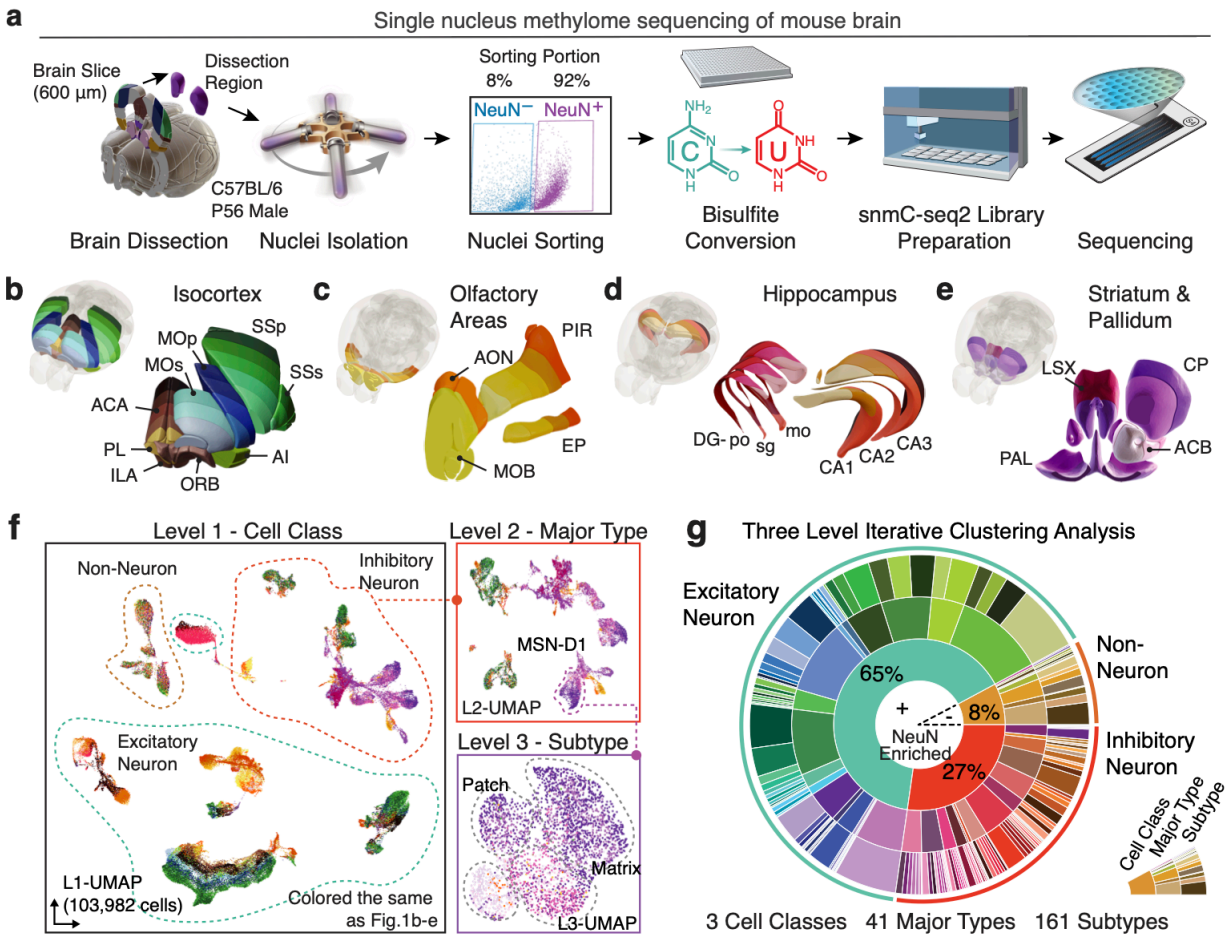


Figure 2.1 A survey of single-cell DNA methylomes in the mouse brain.

a, The workflow of dissection, FANS and snmC-seq2 sequencing. **b**–**e**, Dissected regions of isocortex (**b**), OLF (**c**), HIP (**d**) and CNU (**e**). **f**, Three-level UMAP from iterative analysis, colour coded as in **b**–**e**, panels show an example in which MSN-D1 neurons are separated into subtypes. **g**, Proportions of cells in clusters defined in the three-level iterative analysis. Brain atlas images in **a**–**d** were created based on Wang et al.[55] and © 2017 Allen Institute for Brain Science. Allen Brain Reference Atlas. Available from: atlas.brain-map.org.

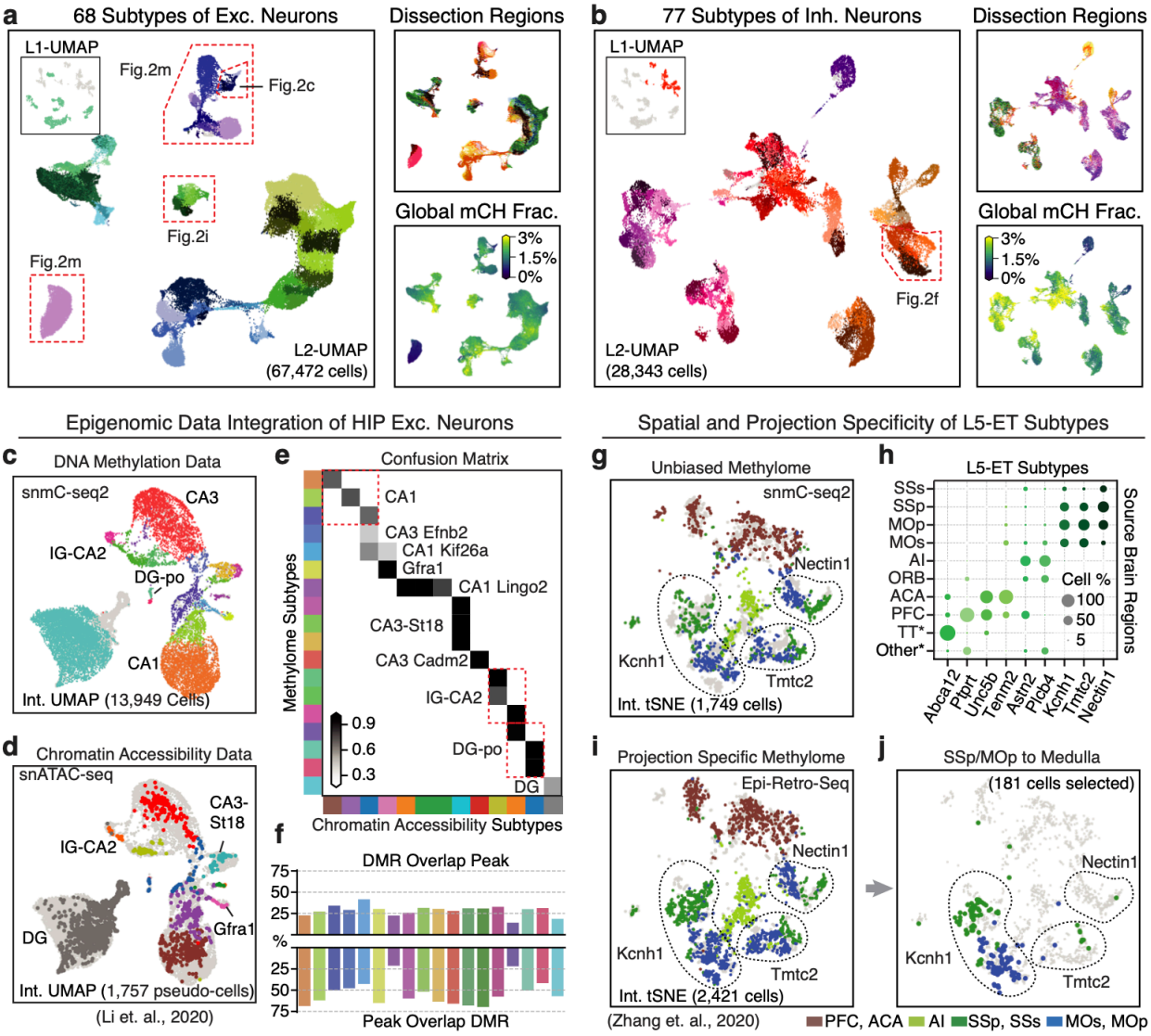


Figure 2.2 Epigenomic diversity of neurons.

a, **b**, Level 2 UMAP of excitatory (**a**) and inhibitory (**b**) neurons, coloured by subtype, dissection region and global mCH fraction. **c**, **d**, Integration UMAP of the HIP excitatory neurons profiled by snmC-seq2 (**c**) and snATAC-seq (**d**; shows pseudo-cells). **e**, Overlap score of a-types and m-types. **f**, Overlap of CG-DMR and ATAC peaks in matched subtypes. **g**, **i**, **j**, Integration *t*-SNE of ET-L5 neurons profiled by snmC-seq2 (**g**) and Epi-Retro-Seq (**i**, **j**), coloured by dissection region. Three SSp- and MOp-enriched subtypes are labelled by their marker gene. **j**, Medulla projecting neurons from SSp or MOp only. **h**, Spatial composition of ET-L5 subtypes.

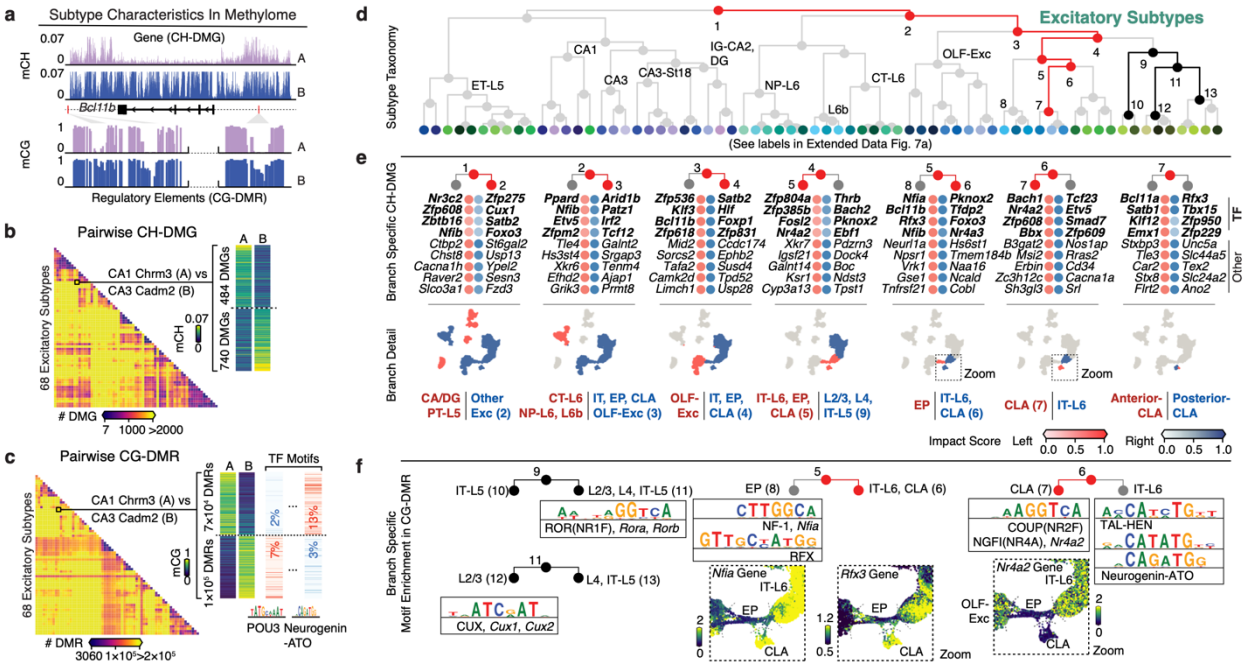


Figure 2.3. Relating genes and regulatory elements to cell subtype taxonomy.

a, Schematic of the two characteristics contained in the methylome profiles. **b, c**, Pairwise CH-DMG (**b**) and CG-DMR (**c**) counts between 68 excitatory subtypes. **c**, In each CG-DMR set, we further identify differentially enriched motifs (left). TF, transcription factor. **d**, Excitatory subtype taxonomy tree. **e**, Top impact scores of ranked genes for the left and right branches of nodes 1–7 in **d**. The top four genes are transcription factor genes (bold); these are followed by other protein-coding genes. The scatter plots below show cells involved in each branch. **f**, Branch-specific transcription factor motif families. The zoomed UMAPs show individual transcription factor genes in those families, whose differential mCH fractions are concordant with their motif enrichment.

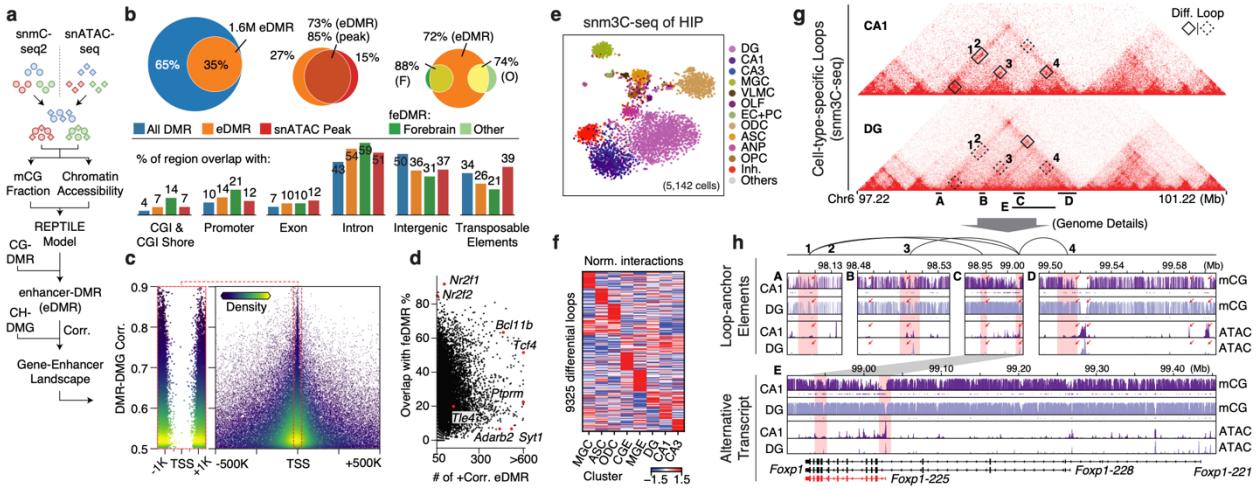


Figure 2.4 Gene–enhancer landscapes in neuronal subtypes.

a, Schematic of enhancer calling using matched DNA methylome and chromatin accessibility subtype profiles. Corr., correlation. **b**, Overlap of regulatory elements identified in this study and other epigenomic studies (ATAC peaks[29] and feDMRs[5]). **c**, DMR–DMG correlation and the distance between DMR centre and gene TSS; each point is a DMR–DMG pair coloured by kernel density. **d**, Percentage of positively correlated eDMRs that overlap with forebrain feDMRs in each gene. **e**, *t*-SNE of cells analysed by sn-m3C-seq coloured by assigned major cell types. **f**, Interaction level (*z*-score across rows and columns) of differential loops in eight clusters at 25-kb resolution. **g**, **h**, Epigenomic signatures surrounding *Foxp1*. **g**, Triangle heat maps showing CA1 and DG chromatin contacts and differential loops. **h**, Genome browser sections showing detailed mCG and ATAC profiles near anchors of four CA1-specific loops. Red rectangles indicate loop anchors and red arrows indicate notable regulatory elements.

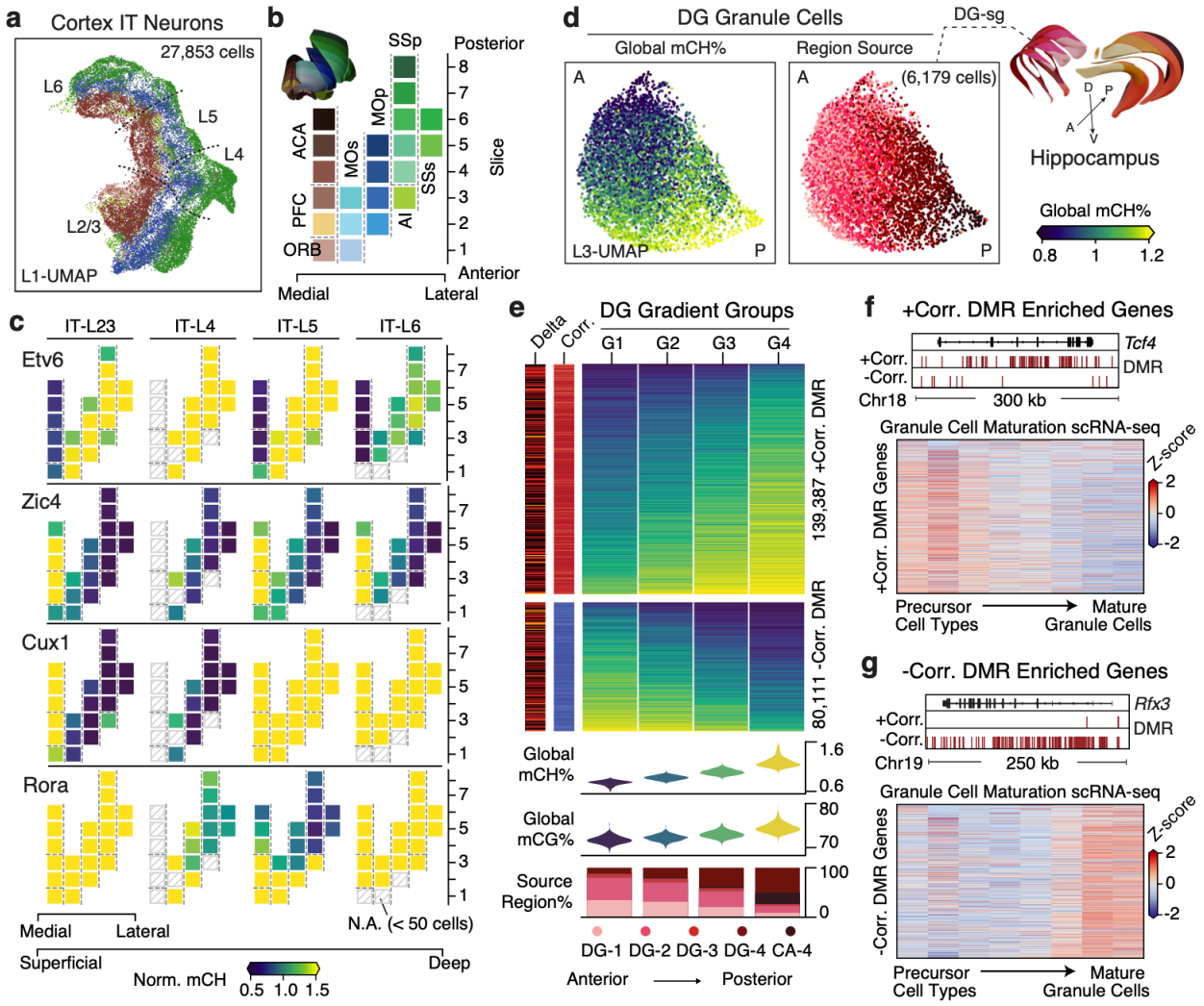


Figure 2.5 Brain-wide spatial gradients of DNA methylation.

a, UMAP for cortex IT neurons coloured by dissection regions. **b**, The 21 cortical dissection regions organized by spatial axis. **c**, Normalized mCH fraction of spatial CH-DMGs, with the same layout as **b**. **d**, UMAPs for DG granule cells coloured by their cell global mCH fractions and dissection regions. A, anterior; P, posterior; D, dorsal; V, ventral. **e**, Compound figure showing four cells groups organized according to DG gradient and the two gradient DMR groups separated according to the sign of the correlation to the cell's global mCH level. **f**, **g**, Bottom, expression (*z*-score across rows) of genes positively (**f**) or negatively (**g**) correlated with DMRs across cell types along the granule cell maturation pathway[69]. Top, genome browser views of representative genes.

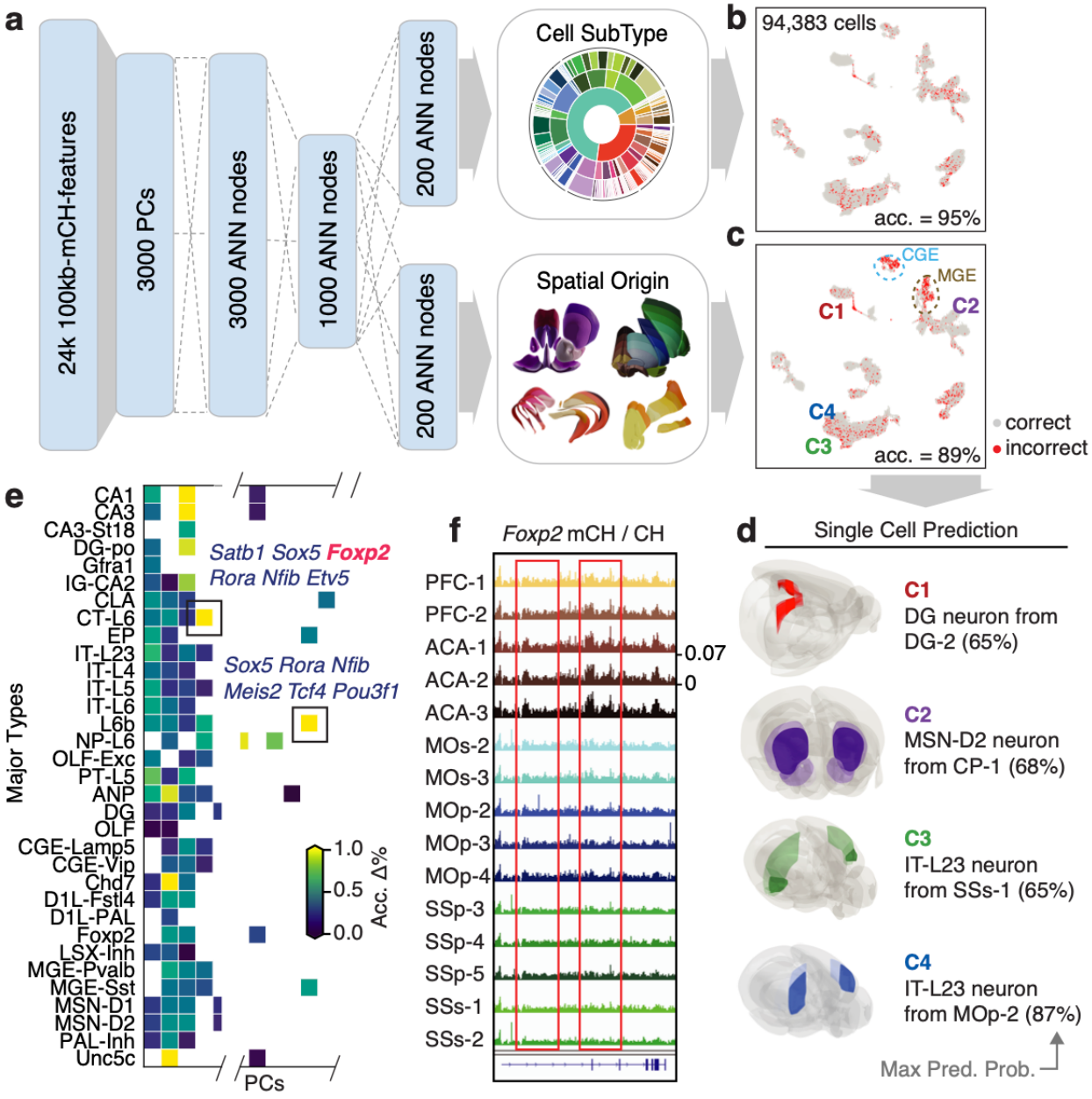


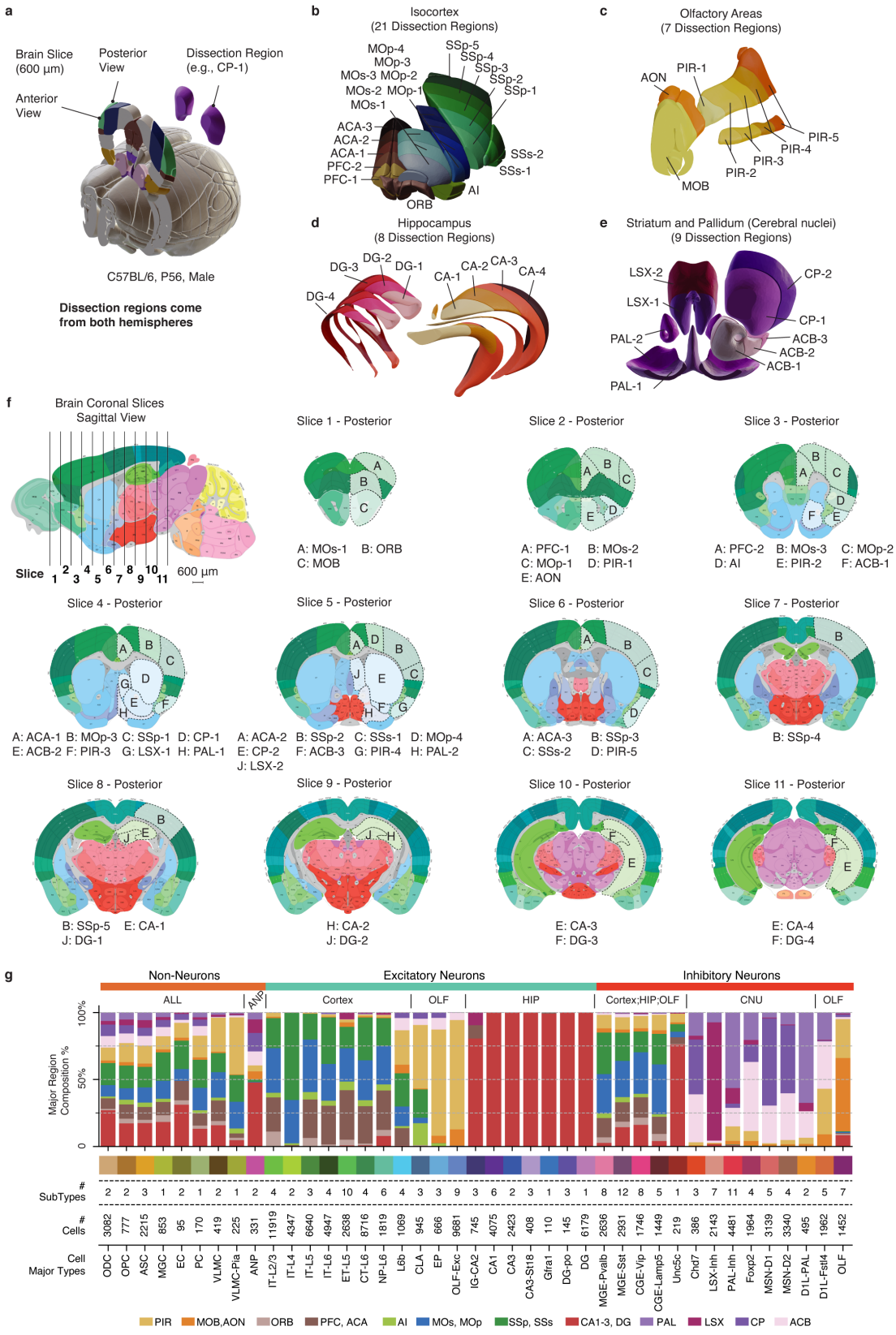
Figure 2.6 A methylome-based predictive model captures both cellular and spatial characteristics of neurons.

a, Schematic of the model that predicts both cell-type identity and spatial origin. **b, c**, Model performance on the prediction of cell subtypes (**b**), and dissection regions (**c**). Per cent accuracy is shown. **d**, Examples of using the model to predict cell spatial origin (maximum prediction probability in parentheses). **e**, Evaluation of importance of features (principal components) for spatial origin prediction. **f**, The *Foxp2* gene body mCH fraction in each cortical dissection region group.

2.18 Supplementary Figures

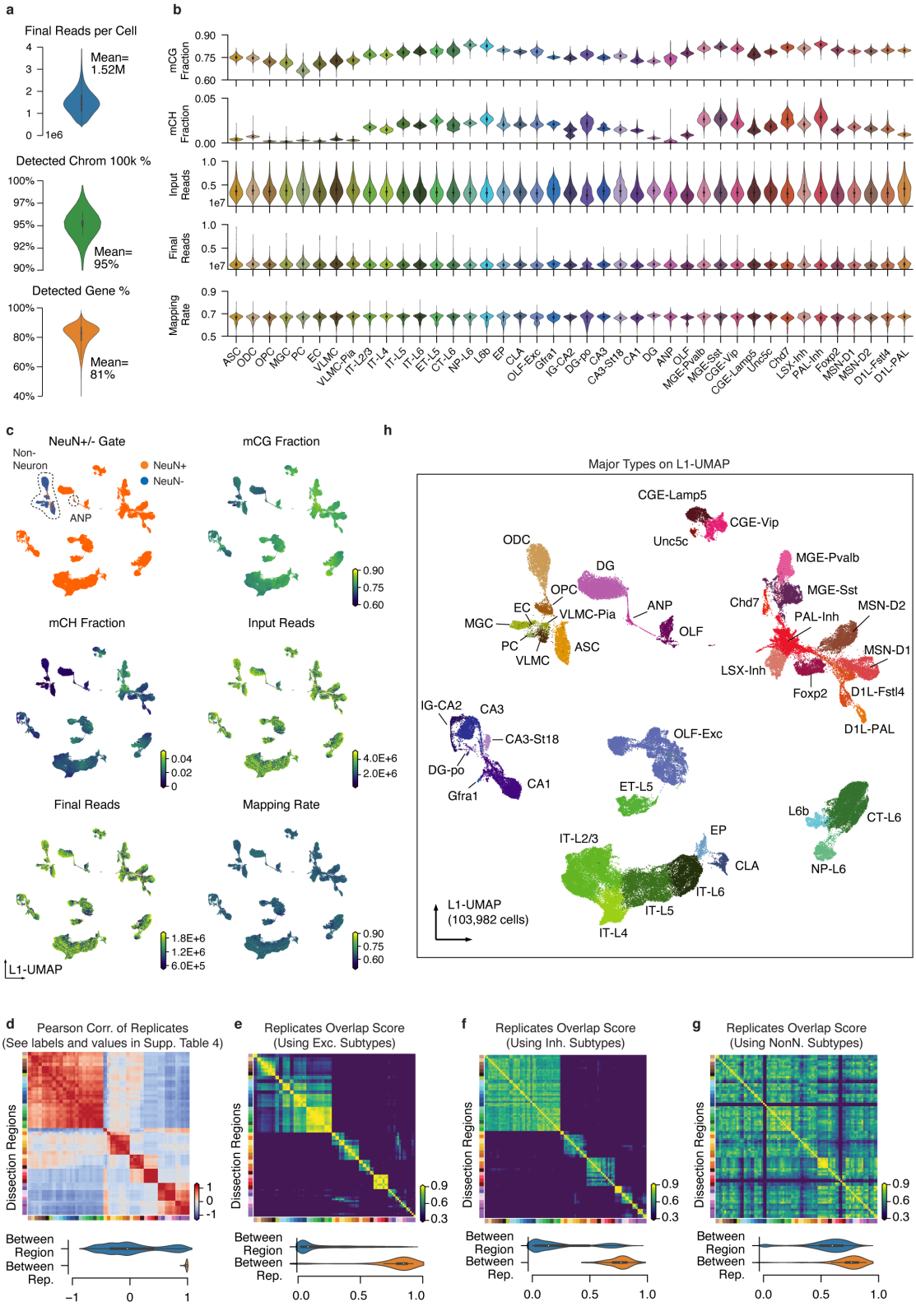
Supplementary Figure 2.1 Brain dissection regions.

a, Schematic of brain dissection steps. Each male C57BL/6 mouse brain (age P56) was dissected into 600-μm slices. We then dissected brain regions from both hemispheres within a specific slice. **b**–**e**, 3D mouse brain schematic adapted from Allen CCFv3 to display the four major brain regions and 45 dissection regions. Each colour represents a dissection region. **f**, 2D mouse brain atlas adapted from Allen Mouse Brain Reference Atlas, the first sagittal image showing the location of each coronal slice, followed by 11 posterior view images of all coronal slices, the same 45 dissection regions are labelled on the corresponding slice. All coronal images follow the same scale as the sagittal image. The posterior view of each slice is the anterior view of the next slice. **g**, An integrated overview of brain region composition, subtype and cell numbers of the major types. All brain atlas images were created based on Wang et al.[55] and © 2017 Allen Institute for Brain Science. Allen Brain Reference Atlas. Available from: http://www.atlas.brain-map.org.


a
Brain Slice (600 µm)
Posterior View
Dissection Region (e.g., CP-1)
Anterior View
C57BL/6, P56, Male
Dissection regions come from both hemispheres
b
Isocortex (21 Dissection Regions)
MOp-4
MOp-3
MOs-3 MOp-2
MOs-2 MOp-1
MOs-1
ACA-3
ACA-2
ACA-1
PFC-2
PFC-1
SSp-5
SSp-4
SSp-3
SSp-2
SSp-1
SSs-2
SSs-1
AI
ORB
c
Olfactory Areas (7 Dissection Regions)
AON
PIR-1
PIR-5
PIR-4
PIR-3
PIR-2
MOB
d
Hippocampus (8 Dissection Regions)
DG-3
DG-2
DG-1
DG-4
CA-1
CA-2
CA-3
CA-4
e
Striatum and Pallidum (Cerebral nuclei) (9 Dissection Regions)
LSX-2
LSX-1
PAL-2
PAL-1
CP-2
CP-1
ACB-3
ACB-2
ACB-1
f
Brain Coronal Slices Sagittal View
Slice 1 2 3 4 5 6 7 8 9 10 11
600 µm
Slice 1 - Posterior
A: MOs-1 B: ORB C: MOB
Slice 2 - Posterior
A: PFC-1 B: MOs-2 C: MOp-1 D: PIR-1 E: AON
Slice 3 - Posterior
A: PFC-2 B: MOs-3 C: MOp-2 D: AI E: PIR-2 F: ACB-1
Slice 4 - Posterior
A: ACA-1 B: MOp-3 C: SSp-1 D: CP-1 E: ACB-2 F: PIR-3 G: LSX-1 H: PAL-1
Slice 5 - Posterior
A: ACA-2 B: SSp-2 C: SSs-1 D: MOp-4 E: CP-2 F: ACB-3 G: PIR-4 H: PAL-2 J: LSX-2
Slice 6 - Posterior
A: ACA-3 B: SSp-3 C: SSs-2 D: PIR-5
Slice 7 - Posterior
B: SSp-4
Slice 8 - Posterior
B: SSp-5 E: CA-1 J: DG-1
Slice 9 - Posterior
H: CA-2 J: DG-2
Slice 10 - Posterior
E: CA-3 F: DG-3
Slice 11 - Posterior
E: CA-4 F: DG-4
g
Non-Neurons
Excitatory Neurons
Inhibitory Neurons
ALL
ANP
Cortex
OLF
HIP
Cortex;HIP;OLF
CNU
OLF
Major Region Composition %
100%
50%
0
SubTypes
Cells
Cell Major Types
ODC OPC ASC MGC EC PC VLMC VLMC-Pia ANP IT-L2/3 IT-L4 IT-L5 IT-L6 ET-L5 CT-L6 NP-L6 L6b CLA EP OLF-Exc IG-CA2 CA1 CA3 CA3-St18 Gfra1 DG-po DG MGE-Pvalb MGE-Sst CGE-Vip CGE-Lamp5 Unc5c Chd7 LSX-Inh PAL-Inh Foxp2 MSN-D1 MSN-D2 D1L-PAL D1L-Fstl4 OLF
PIR MOB,AON ORB PFC, ACA AI MOs, MOp SSp, SSs CA1-3, DG PAL LSX CP ACB

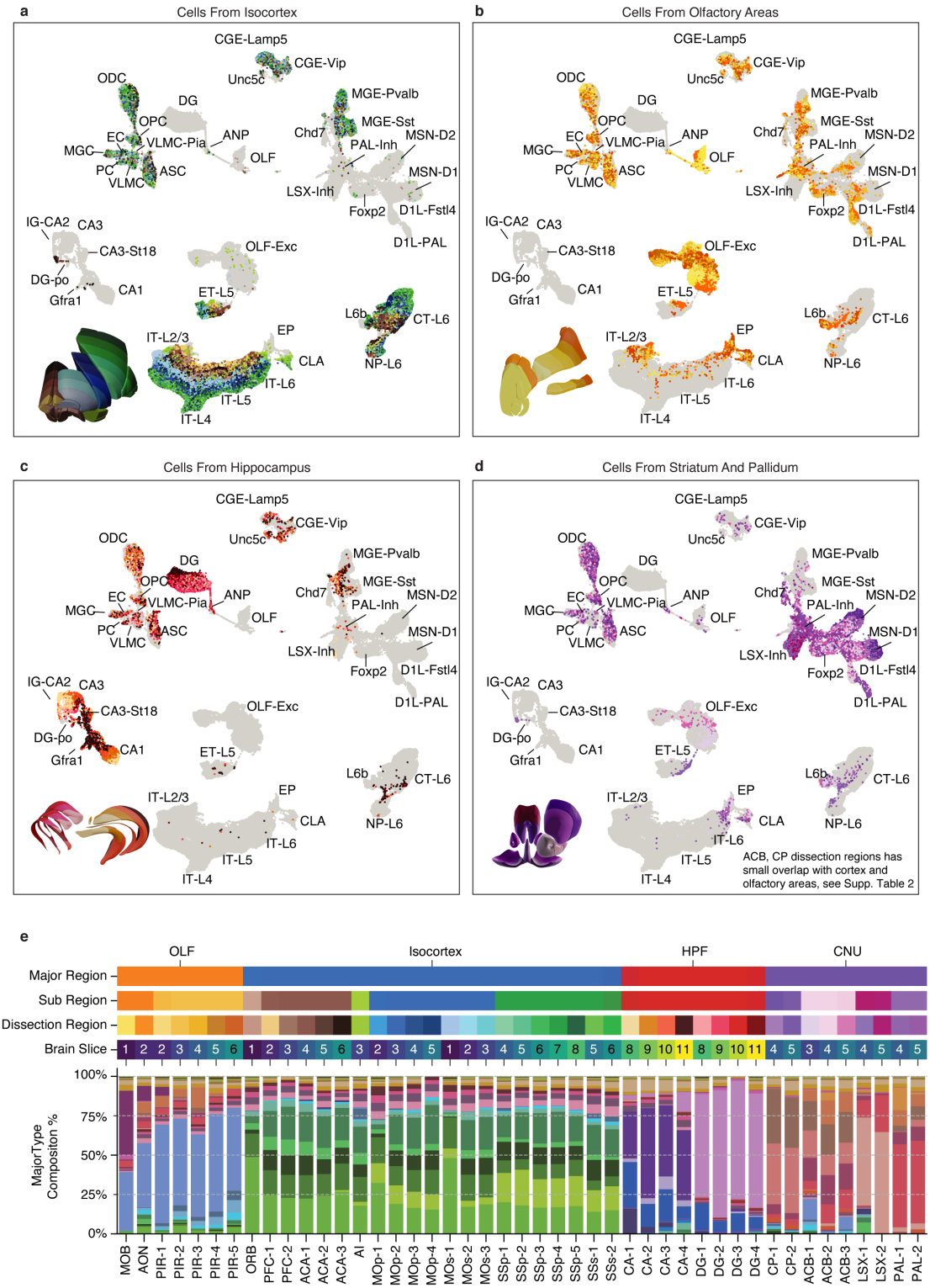
Supplementary Figure 2.2 Major Type labelling and basic mapping metrics of snmC-seq2.

a, The number of final pass QC reads, the percentage of non-overlapping chromosome 100-kb bins detected, and the percentage of GENCODE vm22 genes detected per cell. **b**, Violin plots for all of the key metrics, group by major types. **c**, L1 UMAP coloured by NeuN antibody FACS gates and other snmC-seq2 key read mapping metrics. **d**, Heat map of Pearson correlation between the average methylome profiles (mean mCH and mCG fraction of all chromosome 100-kb bins across all cells belong to a replicate sample) of the 92 replicates from 45 brain regions. The violin plot below summarizes the value between replicates within the same brain region or between different brain regions. **e**–**g**, Pairwise overlap score (measuring co-clustering of two replicates) of excitatory subtypes (**e**), inhibitory subtypes (**f**), and non-neuronal subtypes (**g**). The violin plots summarize the subtype overlap score between replicates within the same brain region or between different brain regions. **h**, L1 UMAP coloured and labelled by major cell types.


a
Final Reads per Cell
Mean= 1.52M
1e6
Detected Chrom 100k %
Mean= 95%
Detected Gene %
Mean= 81%
b
mCG Fraction
mCH Fraction
Input Reads
Final Reads
Mapping Rate
ASC ODC OPC MGC PC EC VLMC VLMC-Pia IT-L2/3 IT-L4 IT-L5 IT-L6 ET-L5 CT-L6 NP-L6 L6b EP CLA OLF-Exc IG-CA2 DG-po CA3 CA3-St18 CA1 DG ANP OLF MGE-Pvalb MGE-Sst CGE-Vip CGE-Lamp5 Unc5c Chd7 LSX-Inh PAL-Inh Foxp2 MSN-D1 MSN-D2 D1L-Fstl4 D1L-PAL
c
NeuN+/- Gate
Non-Neuron
NeuN+
NeuN-
ANP
mCG Fraction
mCH Fraction
Input Reads
Final Reads
Mapping Rate
L1-UMAP
h
Major Types on L1-UMAP
CGE-Lamp5
CGE-Vip
Unc5c
ODC
DG
OPC
EC
VLMC-Pia
ANP
MGC
PC
VLMC
ASC
OLF
MGE-Pvalb
Chd7
MGE-Sst
MSN-D2
PAL-Inh
MSN-D1
LSX-Inh
Foxp2
D1L-Fstl4
D1L-PAL
IG-CA2
CA3
CA3-St18
DG-po
CA1
Gfra1
OLF-Exc
ET-L5
L6b
CT-L6
NP-L6
IT-L2/3
EP
CLA
IT-L6
IT-L5
IT-L4
L1-UMAP
(103,982 cells)
d
Pearson Corr. of Replicates
(See labels and values in Supp. Table 4)
Dissection Regions
Between Region
Between Rep.
e
Replicates Overlap Score
(Using Exc. Subtypes)
f
Replicates Overlap Score
(Using Inh. Subtypes)
g
Replicates Overlap Score
(Using NonN. Subtypes)

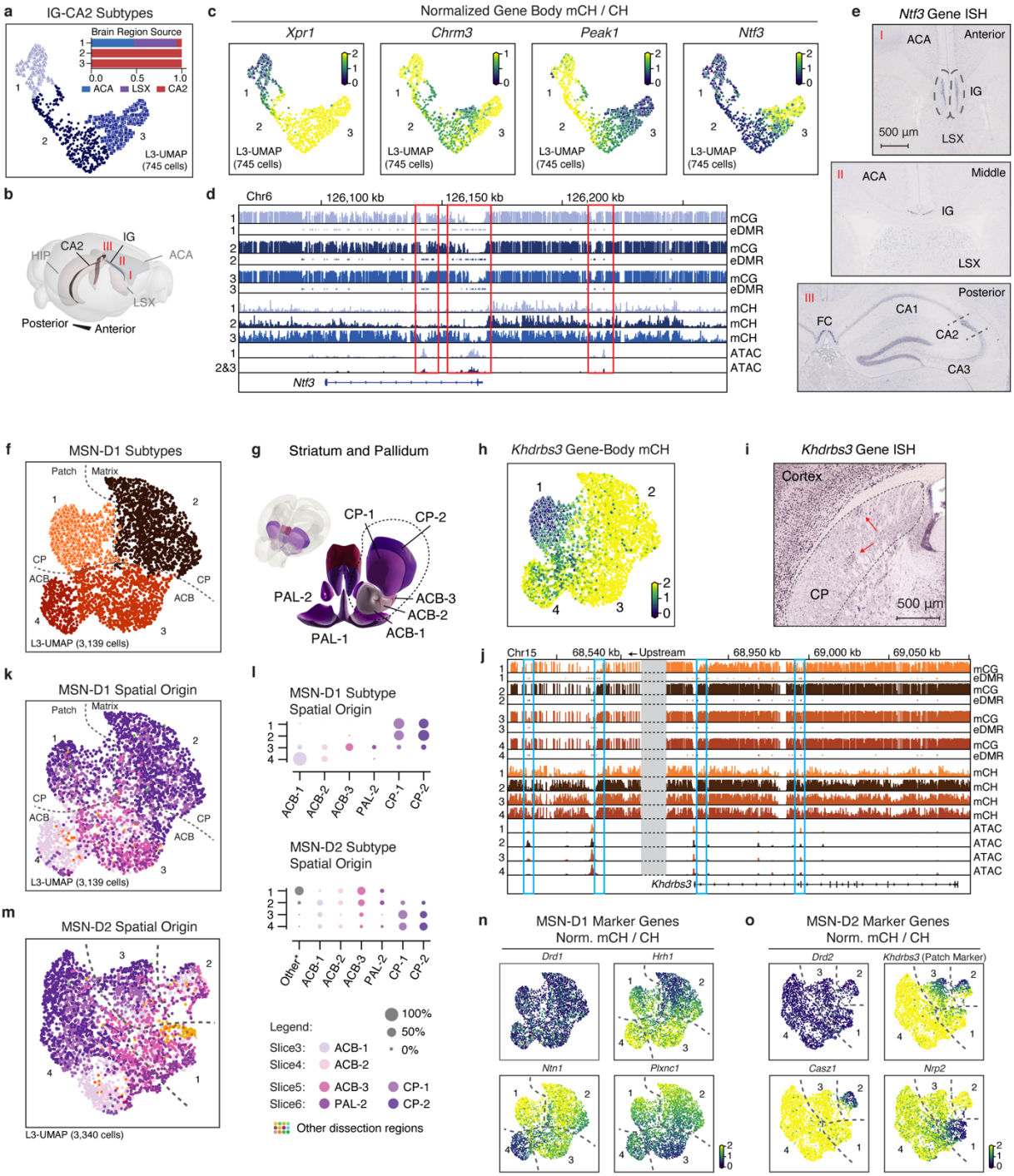
Supplementary Figure 2.3 Cell-type composition of dissection regions.

a-d, L1 UMAP labelled by major types and partially coloured by dissection regions for cells from isocortex (**a**), OLF (**b**), HIP (**c**) and cerebral nucleus (**d**). Other cells are shown in grey as background. **e**, Similar compound bar plot as Supplementary Fig. 2.1g, arranged top to bottom, showing the organization of dissection regions and the major type composition of each dissection region.


a
Cells From Isocortex
b
Cells From Olfactory Areas
c
Cells From Hippocampus
d
Cells From Striatum And Pallidum
CGE-Lamp5
CGE-Vip
Unc5c
ODC
DG
OPC
EC
VLMC-Pia
ANP
MGC
OLF
PC
VLMC
ASC
MGE-Pvalb
Chd7
MGE-Sst
PAL-Inh
MSN-D2
MSN-D1
LSX-Inh
Foxp2
D1L-Fstl4
D1L-PAL
IG-CA2
CA3
CA3-St18
DG-po
CA1
Gfra1
OLF-Exc
ET-L5
L6b
CT-L6
NP-L6
IT-L2/3
EP
CLA
IT-L6
IT-L5
IT-L4
ACB, CP dissection regions has small overlap with cortex and olfactory areas, see Supp. Table 2
e
OLF
Isocortex
HPF
CNU
Major Region
Sub Region
Dissection Region
Brain Slice
MajorType Compositon %
100%
75%
50%
25%
0%
MOB
AON
PIR-1
PIR-2
PIR-3
PIR-4
PIR-5
ORB
PFC-1
PFC-2
ACA-1
ACA-2
ACA-3
AI
MOp-1
MOp-2
MOp-3
MOp-4
MOs-1
MOs-2
MOs-3
SSp-1
SSp-2
SSp-3
SSp-4
SSp-5
SSs-1
SSs-2
CA-1
CA-2
CA-3
CA-4
DG-1
DG-2
DG-3
DG-4
CP-1
CP-2
ACB-1
ACB-2
ACB-3
LSX-1
LSX-2
PAL-1
PAL-2

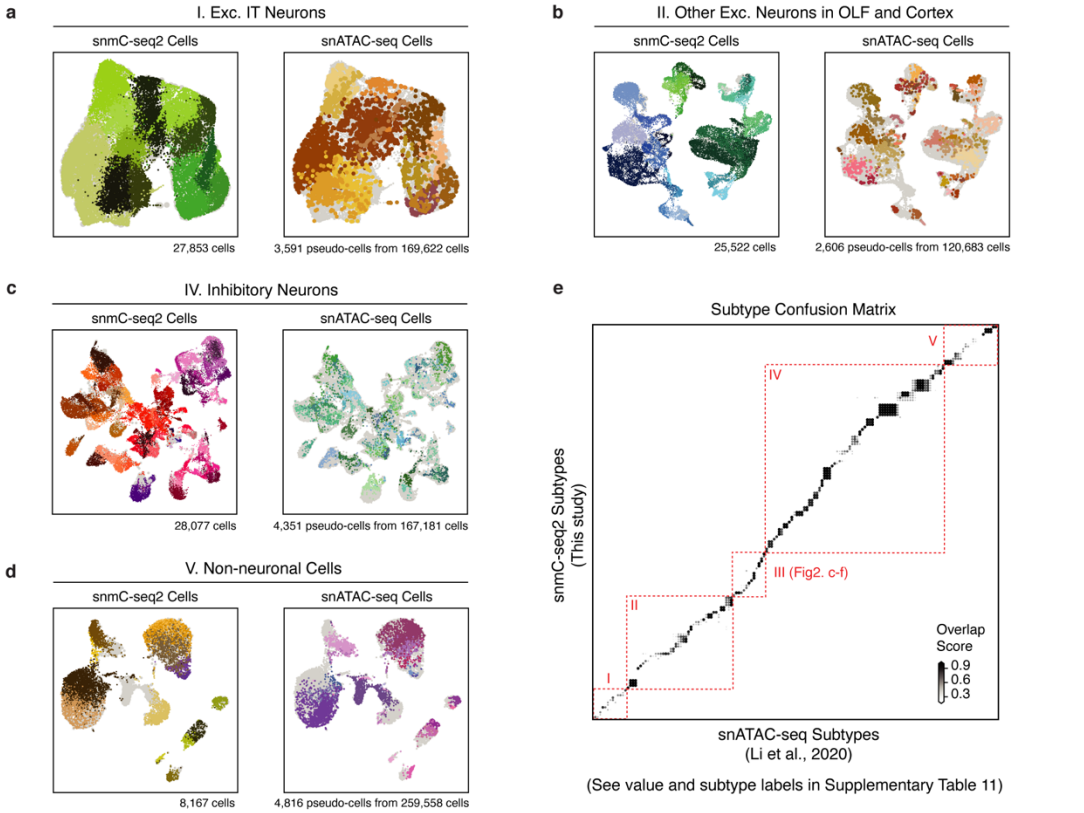
Supplementary Figure 2.4 Supporting details of cellular and spatial diversity of neurons at the subtype level.

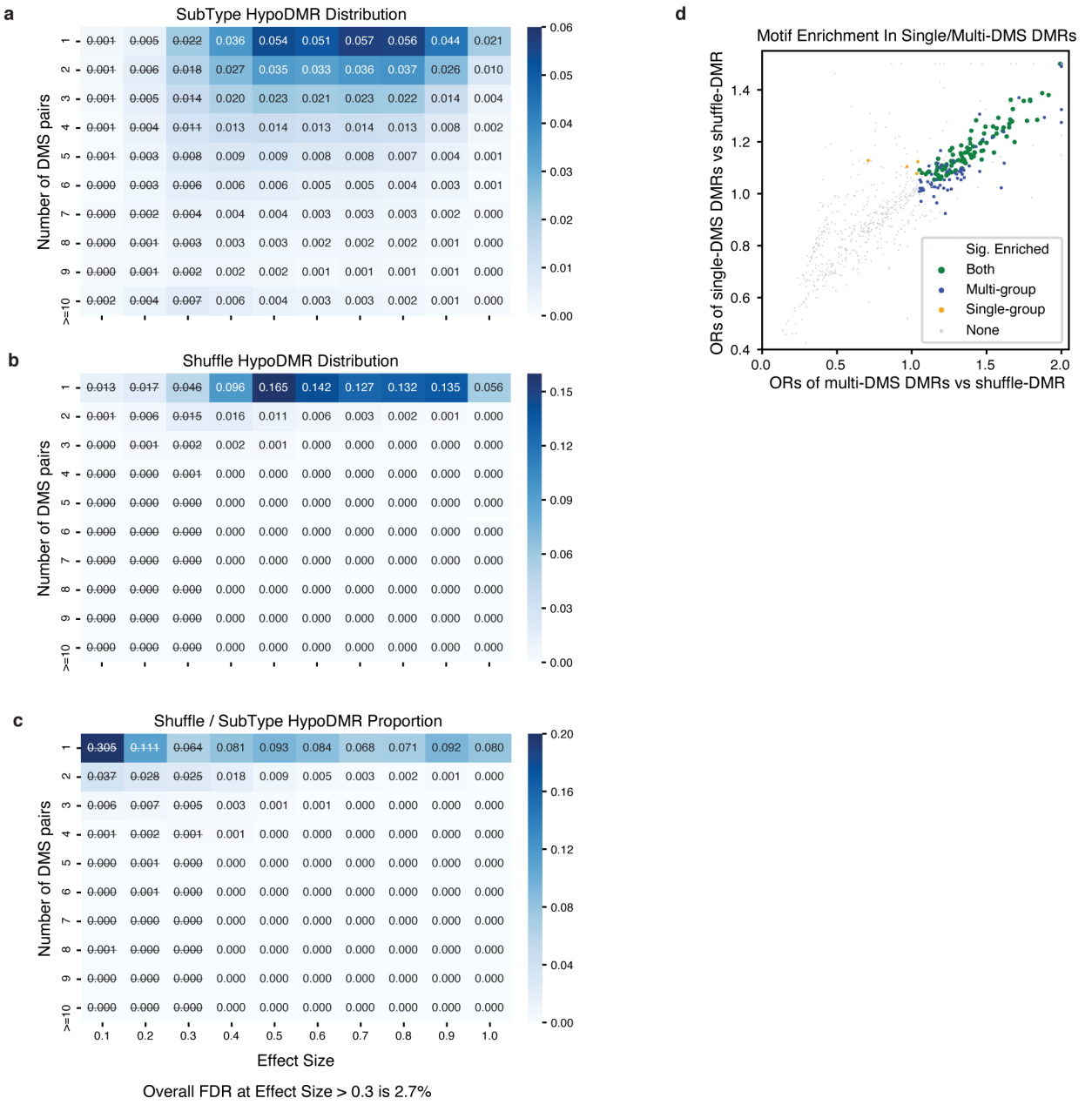
a, Level 3 UMAP of IG-CA2 neurons coloured by subtypes. Bar plot showing sub-region composition of the subtypes: (1) Xpr1, (2) Chrm3, and (3) Peak1. **b**, The 3D model illustrates the spatial relationships between related anatomical structures. **c**, mCH fraction of marker genes in IG-CA2 cells. **d**, Methylome, and chromatin accessibility genome browser view of *Ntf3* genes and its upstream regions. ATAC and eDMR information are from Fig. 2.4 analysis. **e**, Three different views of the in situ hybridization experiment (source: https://mouse.brain-map.org/gene/show/17972; the same patterns were shown in three biological replicates) from Allen Brain Atlas[81], showing the *Ntf3* gene expressed in both IG and CA2. **f**, Level 3 UMAP of MSN-D1 neurons coloured by subtype. Numbers indicate four subtypes: (1) Khdrbs3, (2) Hrh1 (3) Plxnc1, and (4) Ntn1. **g**, The 3D model of related striatum dissection regions. **h**, **i**, mCH fraction (**h**), and an in situ hybridization experiment (source: https://mouse.brain-map.org/gene/show/13769; the same patterns were shown in two biological replicates) from Allen Brain Atlas[81] (**i**) of the *Khdrbs3* gene, red arrows indicate patch regions in CP. **j**, Genome browser view of *Khdrbs3* genes similar to **d**. **k**, Level 3 UMAP of MSN-D1 neurons coloured by dissection regions. **l**, The region composition of each subtype of MSN-D1 and MSN-D2. **m**, MSN-D2 subtypes: (1) Nrp2, (2) Casz1, (3) Col14a1, and (4) Slc24a2. **n**, **o**, mCH fraction of MSN-D1 (**n**) and MSN-D2 (**o**) subtype marker genes. All brain atlas images (**b**, **g**) were created based on Wang et al.[55] and © 2017 Allen Institute for Brain Science. Allen Brain Reference Atlas. Available from: http://atlas.brain-map.org.


a
IG-CA2 Subtypes
Brain Region Source
ACA
LSX
CA2
L3-UMAP
(745 cells)
b
HIP
CA2
IG
ACA
LSX
Posterior
Anterior
c
Normalized Gene Body mCH / CH
Xpr1
Chrm3
Peak1
Ntf3
d
Chr6
126,100 kb
126,150 kb
126,200 kb
mCG
eDMR
mCH
ATAC
e
Ntf3 Gene ISH
ACA
Anterior
IG
LSX
Middle
Posterior
CA1
CA2
CA3
FC
f
MSN-D1 Subtypes
Patch
Matrix
CP
ACB
L3-UMAP (3,139 cells)
g
Striatum and Pallidum
CP-1
CP-2
ACB-3
ACB-2
ACB-1
PAL-2
PAL-1
h
Khdrbs3 Gene-Body mCH
i
Khdrbs3 Gene ISH
Cortex
CP
500 µm
j
Chr15
68,540 kb
Upstream
68,950 kb
69,000 kb
69,050 kb
Khdrbs3
k
MSN-D1 Spatial Origin
l
MSN-D1 Subtype
Spatial Origin
MSN-D2 Subtype
Spatial Origin
Other*
Legend:
100%
50%
0%
Slice3: ACB-1
Slice4: ACB-2
Slice5: ACB-3
Slice6: PAL-2
CP-1
CP-2
Other dissection regions
m
MSN-D2 Spatial Origin
L3-UMAP (3,340 cells)
n
MSN-D1 Marker Genes
Norm. mCH / CH
Drd1
Hrh1
Ntn1
Plxnc1
o
MSN-D2 Marker Genes
Norm. mCH / CH
Drd2
Khdrbs3 (Patch Marker)
Casz1
Nrp2

Supplementary Figure 2.5 Integration with snATAC-seq and epi-retro-seq.

a–**d**, Integration UMAP for snmC-seq2 cells and snATAC-seq pseudo-cells from each cell group: excitatory IT neurons (**a**), other excitatory neurons (**b**), inhibitory neurons (**c**) and non-neuronal cells (**d**). Each panel is coloured by subtypes from the corresponding study, the other dataset is shown in grey in the background. **e**, Overlap score matrix matching the 160 a-types to the 161 m-types. **f**, mCG fraction (left), and chromatin accessibility (right) of cluster-specific CG-DMRs (columns) in HIP subtypes (rows). **g**, **h**, Same integration *t*-SNE as Fig. 2.2g, i coloured by the dissection regions but using all cells profiled by Epi-Retro-Seq (**g**) or snmC-seq2 (**h**), cells from brain regions that have only been profiled via one of the methods are circled out. **i**, Same *t*-SNE as (**h**) coloured by snmC-seq2 subtypes. **j**, Overlap score matrix matching the subtypes to the 'Soma Location (source)' and 'Projection target' information labels of Epi-Retro-Seq cells.


a
I. Exc. IT Neurons
snmC-seq2 Cells
snATAC-seq Cells
27,853 cells
3,591 pseudo-cells from 169,622 cells
b
II. Other Exc. Neurons in OLF and Cortex
snmC-seq2 Cells
snATAC-seq Cells
25,522 cells
2,606 pseudo-cells from 120,683 cells
c
IV. Inhibitory Neurons
snmC-seq2 Cells
snATAC-seq Cells
28,077 cells
4,351 pseudo-cells from 167,181 cells
d
V. Non-neuronal Cells
snmC-seq2 Cells
snATAC-seq Cells
8,167 cells
4,816 pseudo-cells from 259,558 cells
e
Subtype Confusion Matrix
snmC-seq2 Subtypes (This study)
snATAC-seq Subtypes (Li et al., 2020)
I
II
III (Fig2. c-f)
IV
V
Overlap Score
0.9
0.6
0.3
(See value and subtype labels in Supplementary Table 11)
f
mCG / CG (snmC-seq2)
Chromatin Accessibility (snATAC-seq)
0
0.5
1
Hippocampus SubTypes
CA3-St18 Tead1
CA3-St18 Epha5
CA3-St18 Nuak1
CA3 Cadm2
CA3 Efnb2
CA1 Ak5
CA1 Ptprg
CA1 Chrm3
CA1 Kif26a
Gfra1 Gfra1
CA1 Lingo2
DG-po Kctd8
DG-po Calb2
DG-po Bcl11a
IG-CA2 Xpr1
IG-CA2 Peak1
IG-CA2 Chrm3
DG dg-all
The same set of cluster-specific DMR
g
All ET-L5 Cells by Epi-Retro-Seq
MOp
SSp
ACA
AI
AUD
RSP
PTLp
VIS
Cells from regions not profiled in Epi-Retro-Seq
(4,176 cells)
h
All ET-L5 Cells by snmC-seq2
MOp MOs
SSp SSs
PFC ACA
AI
ORB
LSX
TT*
Cells from regions not profiled in snmC-seq2
(2,637 cells)
i
ET-L5 SubTypes by snmC-seq2
Unc5b
Abca12
Tenm2
Ptprt
Nectin1
Plcb4
Astn2
Kcnh1
Tmtc2
L3-TSNE (4,176 cells)
j
Subtype Overlapping Epi-Retro-Seq Informations
Soma Location (Source)
ACA
AI
MOp
SSp
RSP
PTLp
AUD
VIS
Projection Target
MOp
SSp
ACA
VISp
STR
TH
SC
VTA
Pons
MY
Overlap Score
0.5
0.3
0.1
Abca12
Astn2
Kcnh1
Nectin1
Plcb4
Ptprt
Tenm2
Tmtc2
Unc5b
SubTypes by snmC-seq2

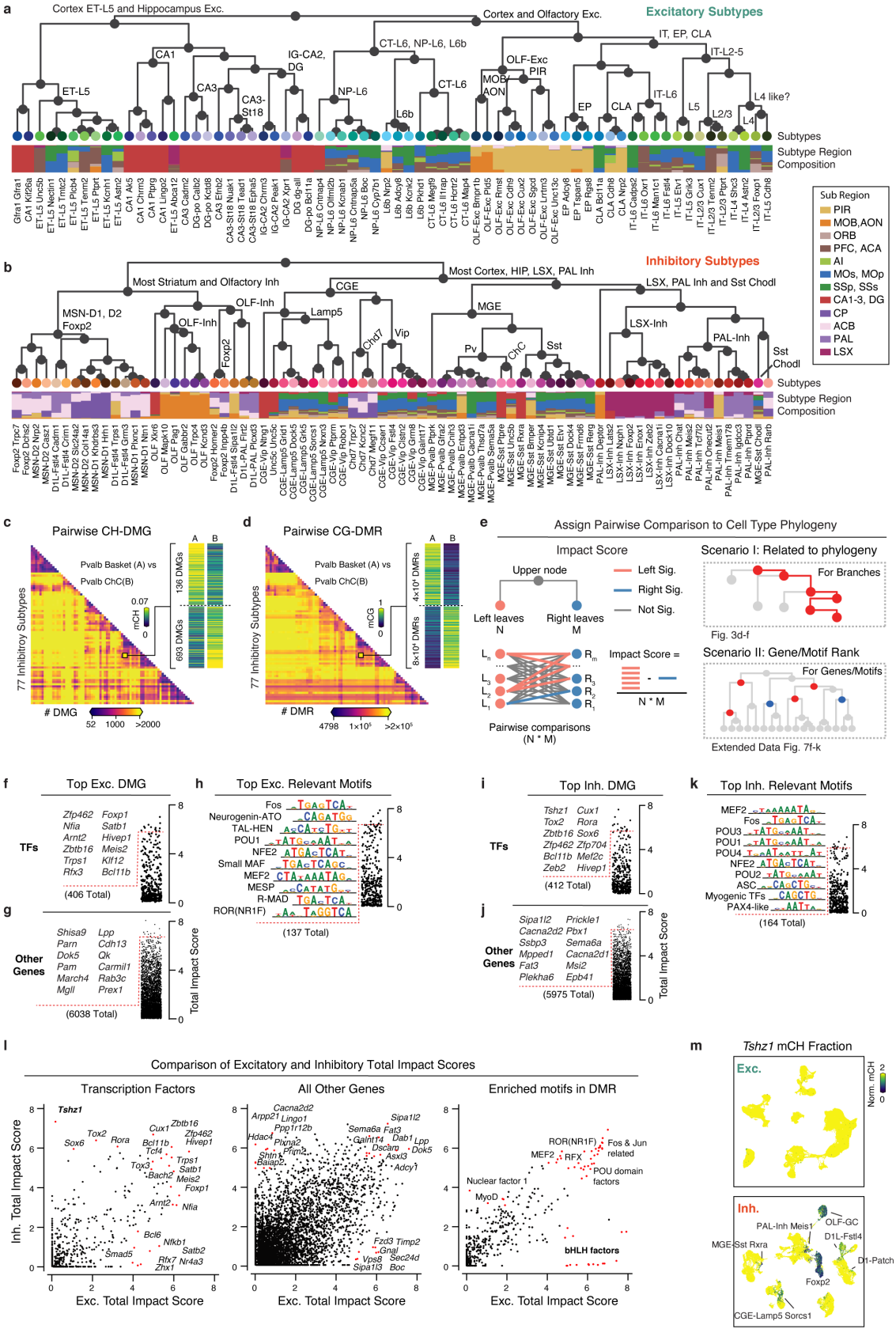


Supplementary Figure 2.6 Controlling the FDR of CG-DMRs.

a, **b**, The proportion of subtype (**a**) or shuffled DMRs (**b**) in each block of specific effect size and number of DMSs. **c**, The empirical FDR of each block, calculated by (no. of shuffle-DMRs/no. of subtype-DMRs) in each block. DMRs with effect size <0.3 were excluded in further analyses. **d**, The odds ratio of transcription factor motif enrichment in single-DMS DMRs (sDMRs) and multi-DMS DMRs (mDMRs). Each dot represents a transcription factor. Transcription factors whose motifs are significantly enriched in both sDMRs and mDMRs are coloured in green, and transcription factors that are significant only in sDMR or mDMRs are coloured in red or blue, respectively. Non-significant transcription factors are shown in grey.

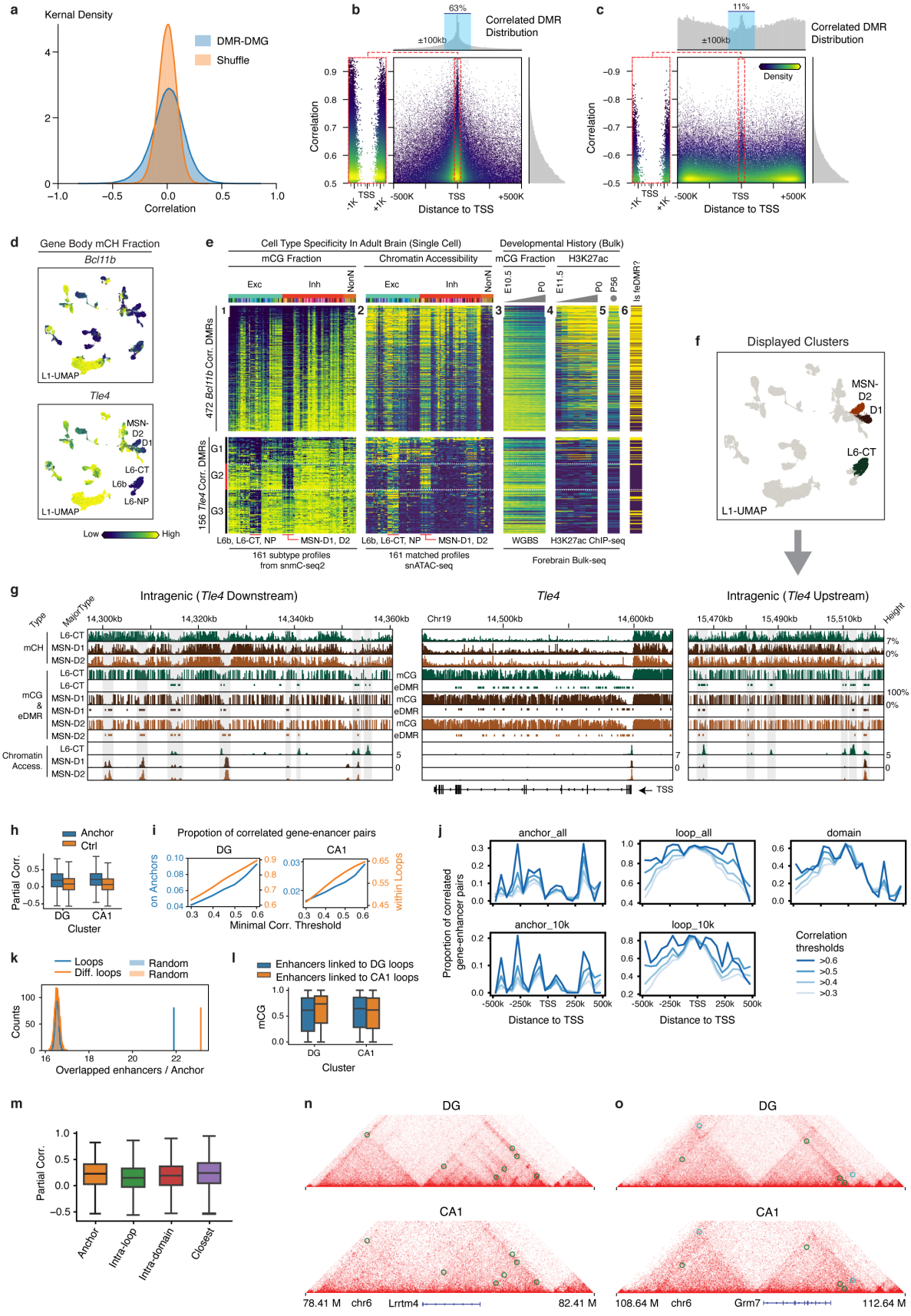
Supplementary Figure 2.7 Subtype taxonomy with related genes and motifs.

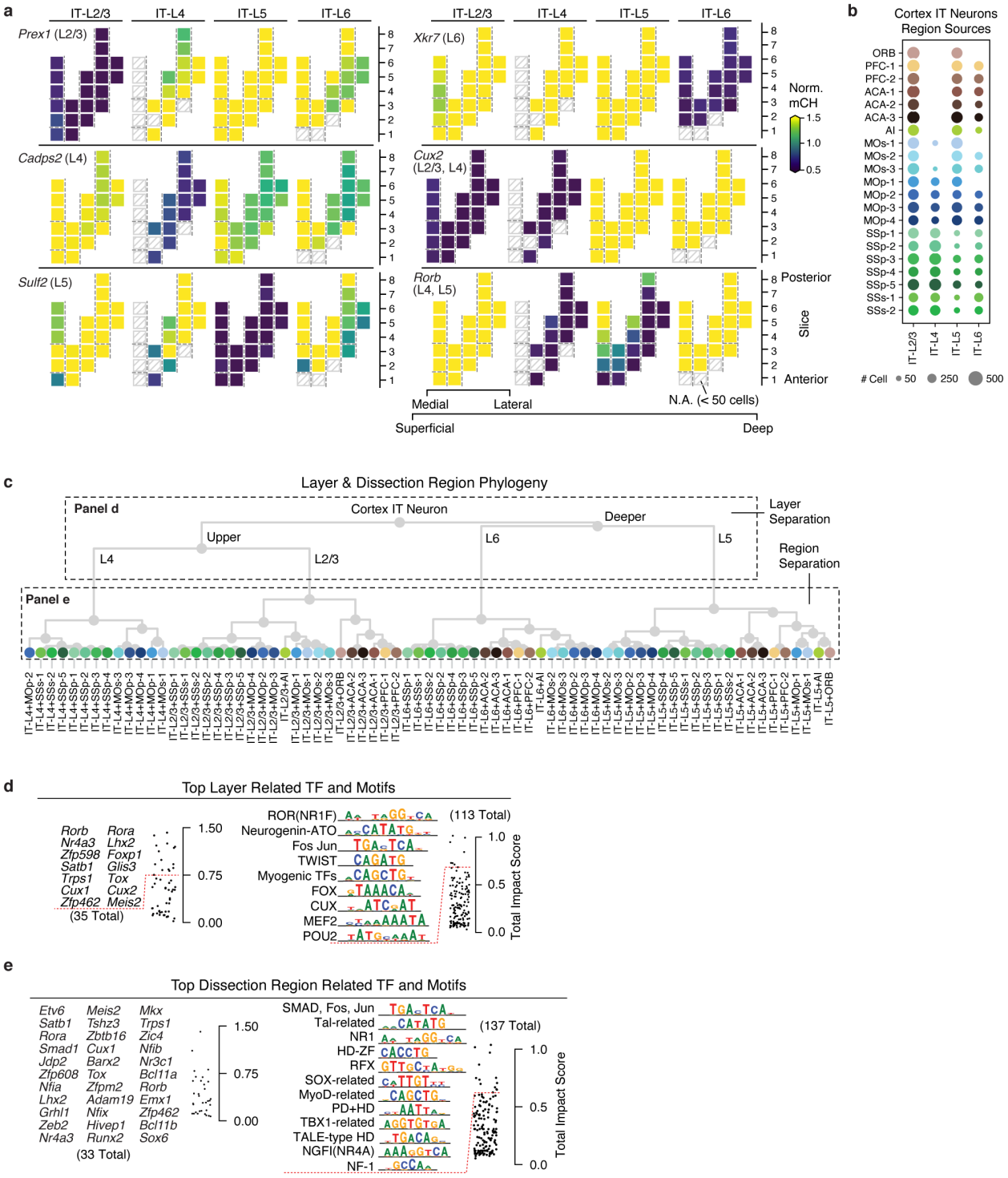
a, **b**, Subtype taxonomy of excitatory (**a**) and inhibitory (**b**) neurons. Leaf nodes are coloured by subtypes, and the bar plot shows subregion composition. **c**, **d**, Counts heat map of pairwise CH-DMG (**c**) and CG-DMR (**d**) between 77 inhibitory subtypes. **e**, Schematic of impact score calculation (left), and two scenarios of discussing impact scores (right). **f**–**h**, Top transcription factors (**f**), other genes (**g**) and enriched motifs (**h**) ranked by total impact score based on the excitatory subtype taxonomy. **i**–**k**, Top transcription factors (**i**), other genes (**j**) and enriched motifs (**k**) ranked by total impact score based on the inhibitory subtype taxonomy. **l**, An example gene *Tshz1* only shows subtype diversity in inhibitory subtypes but not in excitatory subtypes. **m**, Comparison of the total impact scores calculated from either excitatory subtype taxonomy (*x*-axis) or inhibitory subtype taxonomy (*y*-axis) for transcription factors, other genes and enriched motifs.


a
Cortex ET-L5 and Hippocampus Exc.
Cortex and Olfactory Exc.
Excitatory Subtypes
ET-L5
CA1
CA3
CA3-St18
IG-CA2, DG
NP-L6
CT-L6, NP-L6, L6b
L6b
CT-L6
MOB/AON
OLF-Exc PIR
EP
CLA
IT, EP, CLA
IT-L6
IT-L2-5
L5
L2/3
L4
L4 like?
Subtypes
Subtype Region Composition
b
Most Striatum and Olfactory Inh
Most Cortex, HIP, LSX, PAL Inh
LSX, PAL Inh and Sst Chodl
Inhibitory Subtypes
MSN-D1, D2 Foxp2
OLF-Inh
CGE
Lamp5
Chd7
Vip
Foxp2
MGE
Pv
ChC
Sst
LSX-Inh
PAL-Inh
Sst Chodl
Sub Region
PIR
MOB,AON
ORB
PFC, ACA
AI
MOs, MOp
SSp, SSs
CA1-3, DG
CP
ACB
PAL
LSX
c
Pairwise CH-DMG
Pvalb Basket (A) vs
Pvalb ChC(B)
77 Inhibitroy Subtypes
136 DMGs
693 DMGs
mCH
DMG
52 1000 >2000
d
Pairwise CG-DMR
4×10⁴ DMRs
8×10⁴ DMRs
mCG
DMR
4798 1×10⁵ >2×10⁵
e
Assign Pairwise Comparison to Cell Type Phylogeny
Impact Score
Upper node
Left leaves N
Right leaves M
Left Sig.
Right Sig.
Not Sig.
Pairwise comparisons (N * M)
Impact Score =
N * M
Scenario I: Related to phylogeny
For Branches
Fig. 3d-f
Scenario II: Gene/Motif Rank
For Genes/Motifs
Extended Data Fig. 7f-k
f
Top Exc. DMG
TFs
(406 Total)
g
Other Genes
(6038 Total)
Total Impact Score
h
Top Exc. Relevant Motifs
(137 Total)
i
Top Inh. DMG
(412 Total)
j
(5975 Total)
k
Top Inh. Relevant Motifs
(164 Total)
l
Comparison of Excitatory and Inhibitory Total Impact Scores
Transcription Factors
All Other Genes
Enriched motifs in DMR
Inh. Total Impact Score
Exc. Total Impact Score
Nuclear factor 1
MyoD
MEF2
RFX
ROR(NR1F)
Fos & Jun related
POU domain factors
bHLH factors
m
Tshz1 mCH Fraction
Exc.
Norm. mCH
Inh.
PAL-Inh Meis1
OLF-GC
D1L-Fstl4
MGE-Sst Rxra
D1-Patch
Foxp2
CGE-Lamp5 Sorcs1

Supplementary Figure 2.8 Gene-Enhancer landscape related.

a, Distribution of actual DMR–DMG partial correlation compared to the shuffled null distribution. **b**, **c**, DMR–DMG correlation (*y*-axis), and the distance between DMR centre and gene TSS (*x*-axis), each point is a DMR–DMG pair, colour represents points kernel density. The positively (**b**) and negatively (**c**) correlated DMRs are shown separately, owing to very different genome location distributions that are plotted on the top histograms. **d**, The gene body mCH fraction of *Bcl11b* (top) and *Tle4* (bottom) gene. **e**, The predicted enhancer landscape of *Bcl11b* (top) and *Tle4* (bottom). Each row is a correlated eDMR to the gene, columns from left to right are: (1) mCG fraction and (2) ATAC FPKM in 161 subtypes; (3) bulk developing forebrain tissue mCG fraction and (4) H3K27ac FPKM; (5) adult frontal cortex H3K27ac FPKM; and (6) feDMR or not. **f**, detailed view of surrounding eDMRs that are correlated with *Tle4* gene body mCH. Alternative eDMRs appear only in either CT-L6 or MSN-D1/D2 can be seen both upstream and downstream of the gene. **g**, Level 1 UMAP coloured by corresponding cell major types shown in **f**. **h**, Partial correlation between mCG of enhancers and mCH of genes on separated loop anchors of DG (left) and CA1 (right) compared to random anchors with comparable distance (n = 4,171, 4,036, 4,326, 5,133 (left to right)), $P = 5.9 \times 10^{-74}$ for DG and 3.0×10^{-158} for CA1, two-sided Wilcoxon rank-sum tests. **i**, Proportion of loop supported enhancer-gene pairs among the pairs linked by correlation analyses surpassing different correlation thresholds in DG (left) and CA1 (right). The proportion of pairs that the gene and enhancer located on separated anchors of the same loop (blue, left *y*-axis) or within the same loop (orange, right *y*-axis) is shown. **j**, Proportion of loop supported enhancer-gene pairs among those linked by correlation analyses surpassing different correlation thresholds at each specific distance. **k**, Number of enhancers per loop anchor (blue) or per differential loop anchor (orange) compared to randomly selected 25-kb regions across the genome.; $P < 0.005$, two-sided permutation test with 2,000 times repeats. **l**, mCG of enhancers linking to DG specific loops (blue, n = 13,854) and CA1-specific loops (orange, n = 14,373) in DG (left, $P = 2.9\times 10^{-3}$) or CA1 (right, $P = 3.5 \times 10^{-5}$). P values were computed with two-sided Wilcoxon rank-sum tests. **m**, Partial correlation between mCG of enhancers and mCH of genes linked by different methods (n = 4,171, 127,730, 28,203, 10,058 (left to right)). The elements of box plots are defined as: centre line, median; box limits, first and third quartiles; whiskers, 1.5 × interquartile range. **n**, **o**, Interaction maps, mCH, mCG, ATAC and differential loops tracks surrounding *Lrrtm4* (**n**) and *Grm7* (**o**). Circles on the interaction maps represent differential loops between DG and CA1, where green represents DG loops, and cyan represents CA1 loops.


a
Kernal Density
DMR-DMG
Shuffle
Correlation
b
63%
±100kb
Correlated DMR Distribution
Correlation
-1K TSS +1K
-500K TSS +500K
Distance to TSS
c
11%
±100kb
Correlated DMR Distribution
Density
Correlation
-1K TSS +1K
-500K TSS +500K
Distance to TSS
d
Gene Body mCH Fraction
Bcl11b
L1-UMAP
Tle4
MSN-D2
D1
L6-CT
L6b
L6-NP
L1-UMAP
Low
High
e
Cell Type Specificity In Adult Brain (Single Cell)
Developmental History (Bulk)
mCG Fraction
Chromatin Accessibility
mCG Fraction
H3K27ac
Exc
Inh
NonN
E10.5
P0
E11.5
P0
P56
Is eDMR?
472 Bcl11b Corr. DMRs
156 Tle4 Corr. DMRs
G1
G2
G3
L6b, L6-CT, NP
MSN-D1, D2
WGBS
H3K27ac ChIP-seq
161 subtype profiles from snmC-seq2
161 matched profiles snATAC-seq
Forebrain Bulk-seq
f
Displayed Clusters
MSN-D2
D1
L6-CT
L1-UMAP
g
Intragenic (Tle4 Downstream)
Tle4
Intragenic (Tle4 Upstream)
Type
MajorType
Height
mCH
mCG & eDMR
Chromatin Access.
L6-CT
MSN-D1
MSN-D2
mCG
eDMR
TSS
h
Anchor
Ctrl
Partial Corr.
DG
CA1
Cluster
i
Propotion of correlated gene-enhancer pairs
DG
CA1
on Anchors
within Loops
Minimal Corr. Threshold
j
anchor_all
loop_all
domain
anchor_10k
loop_10k
Proportion of correlated gene-enhancer pairs
Distance to TSS
Correlation thresholds
>0.6
>0.5
>0.4
>0.3
k
Loops
Diff. loops
Random
Random
Counts
Overlapped enhancers / Anchor
l
Enhancers linked to DG loops
Enhancers linked to CA1 loops
mCG
DG
CA1
Cluster
m
Partial Corr.
Anchor
Intra-loop
Intra-domain
Closest
n
DG
CA1
78.41 M chr6 Lrrtm4 82.41 M
o
DG
CA1
108.64 M chr6 Grm7 112.64 M

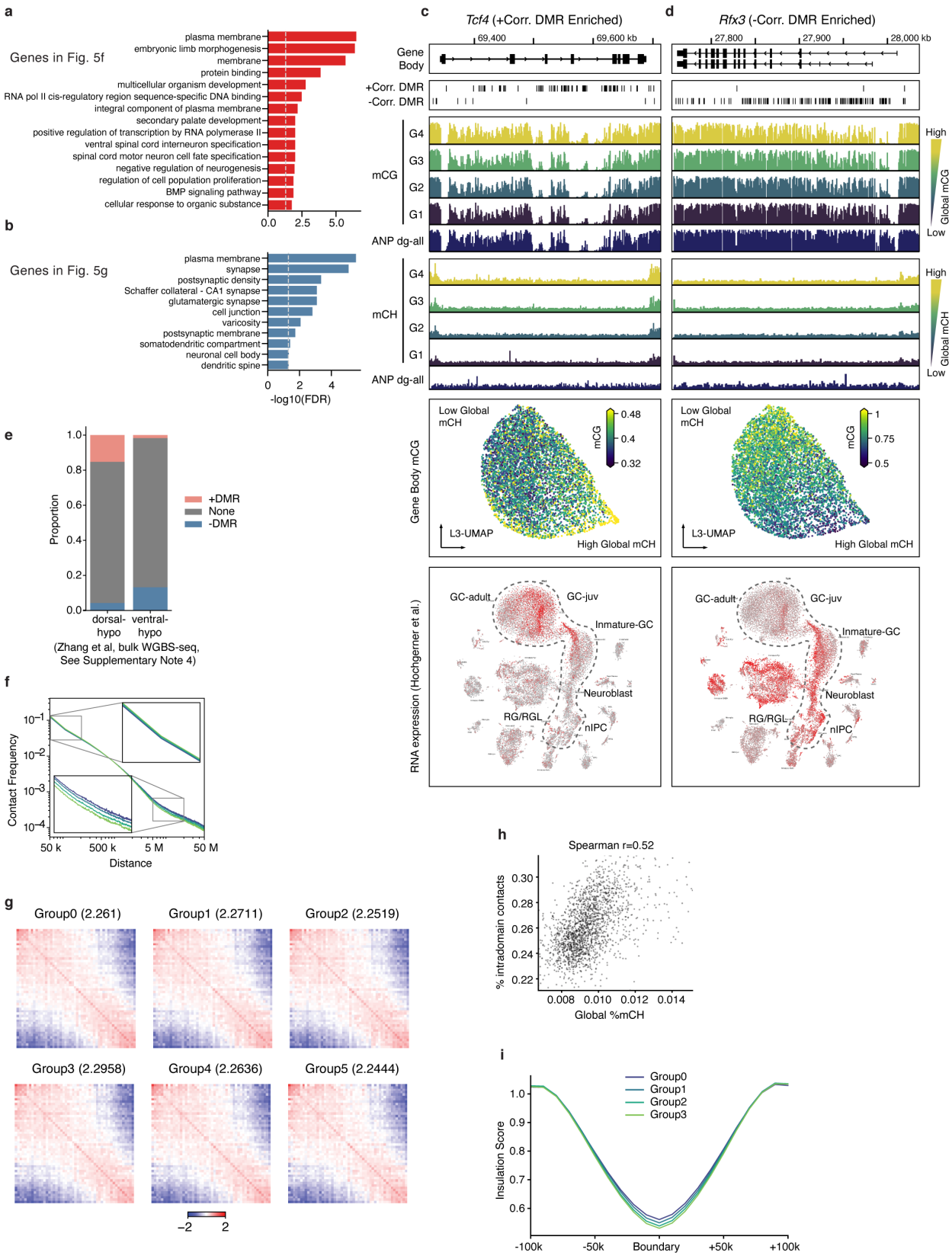


Supplementary Figure 2.9 DNA methylation gradient of IT neurons.

a, Representative marker genes for laminar layers separation. The same dissection region layout in Fig. 2.5b was used here. **b**, Layer-dissection-region cell group taxonomy. **c**, Dot plot sized by the number of cells in each layer-dissection-region combination in excitatory IT neurons. Each group needs at least 50 cells to be included in the analysis. **d**, **e**, The top layer (**d**) and dissection region (**e**) for related TFs and JASPAR motifs ranked by total impact score.

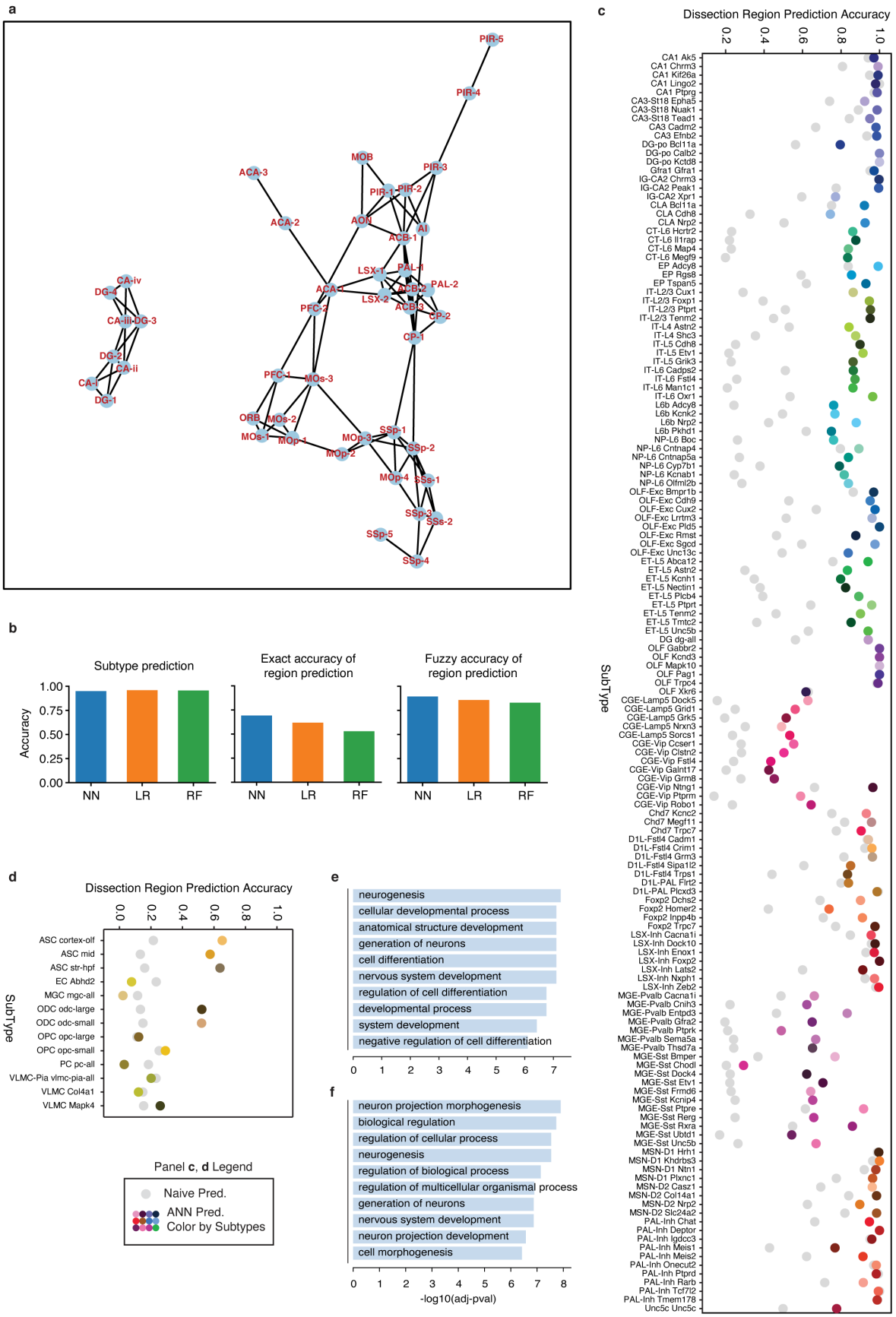
Supplementary Figure 2.10 DNA methylation gradient of DG granule cells.

a, **b**, Top enriched Gene Ontology (GO) terms for +DMRgenes (**a**) and −DMRgenes (**b**). Significant +DMRgenes and −DMRgenes are coloured in red and blue, respectively. **c**, **d**, For the +DMRgene *Tcf4* (**c**) or the −DMRgene *Rfx3* (**d**), the browser view of +DMRs or −DMRs, mCG or mCH in each DG cells groups and adult neural progenitors (ANP), L3 UMAP coloured by gene body mCG, and scRNA-Seq UMAP coloured by gene expression are shown. **e**, The proportion of dorsal or ventral DMRs that overlap with +DMRs and −DMRs. **f**, Interaction frequency decays with increasing genome distances in different groups. **g**, Saddle plots for different groups of DG cells separated by global mCH. Values in the title represent the compartment's strengths. **h**, Correlation between global mCH and proportion of intra-domain contacts across 1,904 DG cells. **i**, Insulation scores of 9,160 domain boundaries and flanking 100-kb regions.


a
Genes in Fig. 5f
plasma membrane
embryonic limb morphogenesis
membrane
protein binding
multicellular organism development
RNA pol II cis-regulatory region sequence-specific DNA binding
integral component of plasma membrane
secondary palate development
positive regulation of transcription by RNA polymerase II
ventral spinal cord interneuron specification
spinal cord motor neuron cell fate specification
negative regulation of neurogenesis
regulation of cell population proliferation
BMP signaling pathway
cellular response to organic substance
0.0 2.5 5.0
b
Genes in Fig. 5g
plasma membrane
synapse
postsynaptic density
Schaffer collateral - CA1 synapse
glutamatergic synapse
cell junction
varicosity
postsynaptic membrane
somatodendritic compartment
neuronal cell body
dendritic spine
0 2 4
-log10(FDR)
c
Tcf4 (+Corr. DMR Enriched)
69,400
69,600 kb
Gene Body
+Corr. DMR
-Corr. DMR
mCG
G4
G3
G2
G1
ANP dg-all
mCH
High
Global mCG
Low
Global mCH
Gene Body mCG
Low Global mCH
mCG
0.48
0.4
0.32
L3-UMAP
High Global mCH
RNA expression (Hochgerner et al.)
GC-adult
GC-juv
Inmature-GC
Neuroblast
RG/RGL
nIPC
d
Rfx3 (-Corr. DMR Enriched)
27,800
27,900
28,000 kb
1
0.75
0.5
e
Proportion
+DMR
None
-DMR
dorsal-hypo
ventral-hypo
(Zhang et al, bulk WGBS-seq,
See Supplementary Note 4)
f
Contact Frequency
50 k
500 k
5 M
50 M
Distance
g
Group0 (2.261)
Group1 (2.2711)
Group2 (2.2519)
Group3 (2.2958)
Group4 (2.2636)
Group5 (2.2444)
−2
2
h
Spearman r=0.52
% intradomain contacts
Global %mCH
i
Insulation Score
Group0
Group1
Group2
Group3
-100k
-50k
Boundary
+50k
+100k

Supplementary Figure 2.11 Evaluation of the predictive model.

a, The neighbour relation among the potential overlapping dissection regions. The network is constructed based on information of the dissection scheme and the 'Potential overlap' column in Supplementary Table 2 and is used to compute the fuzzy accuracy. **b**, The exact accuracy of subtype prediction (top), dissected region prediction (middle), and fuzzy accuracy of dissected region prediction (bottom) of neural network (NN, blue), logistic regression (LR, orange) and random forest (RF, green). **c**, **d**, Prediction accuracy of dissection region at cell subtype level of neurons (**c**) and non-neuronal cells (**d**). Coloured points denote the prediction accuracy of the model, whereas grey points denote the random guess accuracy when cell subtypes and corresponding spatial distributions are given. **g**, **h**, GO-term enrichment of top-loading genes of features that are important for predicting the spatial location of CT-L6 (**g**) and L6b (**h**).


a
PIR-5
PIR-4
MOB
PIR-3
ACA-3
PIR-1
PIR-2
ACA-2
AON
AI
ACB-1
LSX-1
PAL-1
ACA-1
ACB-2
PAL-2
LSX-2
PFC-2
ACB-3
CP-2
CP-1
CA-iv
DG-4
CA-ii-DG-3
DG-2
CA-ii
CA-i
DG-1
PFC-1
MOs-3
ORB
MOs-2
MOs-1
MOp-1
MOp-3
SSp-1
SSp-2
MOp-2
MOp-4
SSs-1
SSp-3
SSs-2
SSp-5
SSp-4
b
Subtype prediction
Exact accuracy of region prediction
Fuzzy accuracy of region prediction
Accuracy
1.00
0.75
0.50
0.25
0.00
NN
LR
RF
c
Dissection Region Prediction Accuracy
0.2
0.4
0.6
0.8
1.0
SubType
CA1 Ak5
CA1 Chrm3
CA1 Kif26a
CA1 Lingo2
CA1 Ptprg
CA3-St18 Epha5
CA3-St18 Nuak1
CA3-St18 Tead1
CA3 Cadm2
CA3 Efnb2
DG-po Bcl11a
DG-po Calb2
DG-po Kctd8
Gfra1 Gfra1
IG-CA2 Chrm3
IG-CA2 Peak1
IG-CA2 Xpr1
CLA Bcl11a
CLA Cdh8
CLA Nrp2
CT-L6 Hcrtr2
CT-L6 Il1rap
CT-L6 Map4
CT-L6 Megf9
EP Adcy8
EP Rgs8
EP Tspan5
IT-L2/3 Cux1
IT-L2/3 Foxp1
IT-L2/3 Ptprt
IT-L2/3 Tenm2
IT-L4 Astn2
IT-L4 Shc3
IT-L5 Cdh8
IT-L5 Etv1
IT-L5 Grik3
IT-L6 Cadps2
IT-L6 Fstl4
IT-L6 Man1c1
IT-L6 Oxr1
L6b Adcy8
L6b Kcnk2
L6b Nrp2
L6b Pkhd1
NP-L6 Boc
NP-L6 Cntnap4
NP-L6 Cntnap5a
NP-L6 Cyp7b1
NP-L6 Kcnab1
NP-L6 Olfml2b
OLF-Exc Bmpr1b
OLF-Exc Cdh9
OLF-Exc Cux2
OLF-Exc Lrrtm3
OLF-Exc Pld5
OLF-Exc Rmst
OLF-Exc Sgcd
OLF-Exc Unc13c
ET-L5 Abca12
ET-L5 Astn2
ET-L5 Kcnh1
ET-L5 Nectin1
ET-L5 Plcb4
ET-L5 Ptprt
ET-L5 Tenm2
ET-L5 Tmtc2
ET-L5 Unc5b
DG dg-all
OLF Gabbr2
OLF Kcnd3
OLF Mapk10
OLF Pag1
OLF Trpc4
OLF Xkr6
CGE-Lamp5 Dock5
CGE-Lamp5 Grid1
CGE-Lamp5 Grk5
CGE-Lamp5 Nrxn3
CGE-Lamp5 Sorcs1
CGE-Vip Ccser1
CGE-Vip Clstn2
CGE-Vip Fstl4
CGE-Vip Galnt17
CGE-Vip Grm8
CGE-Vip Ntng1
CGE-Vip Ptprm
CGE-Vip Robo1
Chd7 Kcnc2
Chd7 Megf11
Chd7 Trpc7
D1L-Fstl4 Cadm1
D1L-Fstl4 Crim1
D1L-Fstl4 Grm3
D1L-Fstl4 Sipa1l2
D1L-Fstl4 Trps1
D1L-PAL Flrt2
D1L-PAL Plcxd3
Foxp2 Dchs2
Foxp2 Homer2
Foxp2 Inpp4b
Foxp2 Trpc7
LSX-Inh Cacna1i
LSX-Inh Dock10
LSX-Inh Enox1
LSX-Inh Foxp2
LSX-Inh Lats2
LSX-Inh Nxph1
LSX-Inh Zeb2
MGE-Pvalb Cacna1i
MGE-Pvalb Cnih3
MGE-Pvalb Entpd3
MGE-Pvalb Gfra2
MGE-Pvalb Ptprk
MGE-Pvalb Sema5a
MGE-Pvalb Thsd7a
MGE-Sst Bmper
MGE-Sst Chodl
MGE-Sst Dock4
MGE-Sst Etv1
MGE-Sst Frmd6
MGE-Sst Kcnip4
MGE-Sst Ptpre
MGE-Sst Rerg
MGE-Sst Rxra
MGE-Sst Ubtd1
MGE-Sst Unc5b
MSN-D1 Hrh1
MSN-D1 Khdrbs3
MSN-D1 Ntn1
MSN-D1 Plxnc1
MSN-D2 Casz1
MSN-D2 Col14a1
MSN-D2 Nmb2
MSN-D2 Slc24a2
PAL-Inh Chat
PAL-Inh Deptor
PAL-Inh Igdcc3
PAL-Inh Meis1
PAL-Inh Meis2
PAL-Inh Onecut2
PAL-Inh Ptprd
PAL-Inh Rarb
PAL-Inh Tcf7l2
PAL-Inh Tmem178
Unc5c Unc5c
d
Dissection Region Prediction Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
SubType
ASC cortex-olf
ASC mid
ASC str-hpf
EC Abhd2
MGC mgc-all
ODC odc-large
ODC odc-small
OPC opc-large
OPC opc-small
PC pc-all
VLMC-Pia vlmc-pia-all
VLMC Col4a1
VLMC Mapk4
Panel c, d Legend
Naive Pred.
ANN Pred.
Color by Subtypes
e
neurogenesis
cellular developmental process
anatomical structure development
generation of neurons
cell differentiation
nervous system development
regulation of cell differentiation
developmental process
system development
negative regulation of cell differentiation
0
1
2
3
4
5
6
7
f
neuron projection morphogenesis
biological regulation
regulation of cellular process
neurogenesis
regulation of biological process
regulation of multicellular organismal process
generation of neurons
nervous system development
neuron projection development
cell morphogenesis
0
1
2
3
4
5
6
7
8
-log10(adj-pval)

2.19 Supplementary Notes

2.19.1 Neuronal Subtype Vignettes

2.19.1.1 Methylome similarity between Indusium Griseum (IG) and Hippocampal region CA2 neuron subtypes

Nearly all hippocampal excitatory neurons formed their own clusters separately from cells in other brain regions (Extended Data Fig. 1g), except the IG-CA2 neurons (745 cells, three subtypes) (Extended Data Fig. 4a). Several markers of the hippocampal region CA2[97] (e.g., *Pcp4*[98], *Cacng5*[99], *Ntf3*[100]) were marked by low gene body mCH (Extended Data Fig. 4c) in IG-CA2 subtypes. However, one subtype "IG-CA2 Xpr1" (subtype 1 in Extended Data Fig. 2a, 152 cells) was located in ACA (72 cells) and Lateral Septal Complex (LSX, 71 cells), which are anatomically distinct from the hippocampus (Extended Data Fig. 1). In-situ hybridization (ISH) data from the Allen Brain Atlas (ABA) of *Ntf3* (Extended Data Fig. 4e) indicates that those cells potentially come from the IG region[101], which is a thin layer of gray matter lying dorsal to the corpus callosum at the base of the anterior half of the cingulate cortex, included in the ACA and LSX dissection regions (Extended Data Fig. 4b). With the power of single-cell epigenomic profiling, we are able to classify cells from this region without additional dissection. Moreover, the similarity of their DNA methylomes indicates the possible functional and/or developmental relationship between the CA2 and IG excitatory neurons. Finally, the hypo-mCG regions surrounding the marker genes like *Ntf3* (Extended Data Fig. 4d) identify candidate cell-type-specific enhancers which may further our understanding of how specific gene expression programs are regulated in these structures.

2.19.1.2 Methylation signatures of striatal medium spiny neurons located in patch compartments

A major GABAergic inhibitory cell type in the striatum, the *Drd1*+ medium spiny neuron (MSN-D1) from the caudoputamen (CP, dorsal), and nucleus accumbens (ACB, ventral), is further separated into four subtypes (Extended Data Fig. 4f, k, n). Two subtypes "MSN-D1 Plxnc1" (subtype 3 in Extended Data Fig. 4f) and "MSN-D1 Ntn1" (subtype 4 in Extended Data Fig. 4f), mainly (79%) from the ACB, are further separated by location along the anterior-posterior axis (Extended Data Fig. 4g, l, based on dissection), indicating spatial diversity may exist in addition to the canonical dorsal-ventral gradient of the striatum[102,103]. "MSN-D1 Khdrbs3" and "MSN-D1 Hrh1" subtypes are mostly (94%) from CP dissections, one of them marked by gene *Khdrbs3* and its potential regulatory elements (Extended Data Fig. 4h-j) corresponds to the neurochemically defined patch[104] compartments in CP. Here the methylome profiling data provides evidence of previously unseen spatial epigenetic diversity in the striatum D1 neurons, which is also observed in the other major type MSN-D2 (*Drd2*$^+$, Extended Data Fig. 4m, o) of the striatum.

2.19.2 Estimate subtype CG-DMR false discovery rate

To estimate the FDR for DMRs, we randomly partitioned the cells into the same number of groups as the number of clusters. We used methylpy DMRfind[41] (Methods) and the same filter as above to identify the DMRs between these random groups (shuffle-DMRs). 105,310 shuffle-DMRs were identified, compared to 3,947,795 subtype-DMRs, so the average FDR is about 2.7%. For each DMR, the effect size was calculated by subtracting the minimum mCG fraction across samples from the samples' robust mean. We then divided the DMRs into different groups based on the number of DMSs and the effect size of the DMR and computed the FDR within each group (Extended Data Fig. 6a, b). Most (93%) of the shuffle-DMRs only have a single DMS (no other DMS within ± 250 bp), while this proportion decreases to 35% for subtype-DMRs. When the effect

size is greater than 0.3, the FDR for DMS = 1 bins range from 0.071 to 0.093. The FDR for remaining bins having DMS > 1 is close to or well below 0.01 (Extended Data Fig. 6c).

To test whether the 1,645,355 (35%) single-DMS subtype-DMRs (sDMRs) are biologically meaningful, we used the shuffle-DMR as background regions and performed motif enrichment analysis on sDMRs and multi-DMS subtype-DMRs (mDMRs) respectively. We found 108 / 174 motifs enriched in mDMRs are also enriched in sDMRs, and the odds ratio of all mDMR-enriched motifs are highly correlated (Extended Data Fig. 6d, Pearson's r=0.86, P = 1e-53). These results indicate that although single-DMS DMRs are noisier than multi-DMS DMRs, they are likely biologically relevant. Removing single-DMS DMRs would improve the overall FDR from 3.0% to 0.3%, with the cost of reducing the power to identify true positives. We provided the number of DMRs or each subtype using different filtering criteria in Supplementary Table 12.

2.19.3 Total Impact score summarizes variation of each gene or motif

Beyond identifying specific cell subtype characteristics, we hypothesized that ranking of genes or motifs by methylation variation may provide a route toward understanding their relative importance in cell type diversification and/or function. Thus, for each gene or motif, we calculated a total impact (TI) score to summarize their variation among the subtypes, using all the branch-specific impact scores (Extended Data Fig. 7e bottom right, Methods). Genes (Excitatory: Extended Data Fig. 7f, g; Inhibitory: Extended Data Fig. 7i, j) or motifs (Excitatory: Extended Data Fig. 7h; Inhibitory: Extended Data Fig. 7k) with higher TI scores impact more branches of the phylogeny, thus are predicted to have a higher importance in distinguishing multiple subtypes. Intriguingly, by comparing the TI scores of genes and motifs calculated from the inhibitory and excitatory phylogenies, we found more TF genes and motifs having large TI scores in both cell classes than specific to either one (Extended Data Fig. 7l). For instance, *Bcl11b* distinguishes

"OLF-Exc" and Isocortex IT neurons in the excitatory lineage and distinguishes "CGE-Lamp5" and "CGE-Vip" in the inhibitory lineage. Similarly, *Satb1* separates IT-L4 from IT-L2/3, and MGE from CGE in excitatory and inhibitory cells, respectively. These findings indicate broad repurposing of TFs for cell-type specification among distinct developmental lineages.

In contrast, we also find TF genes and motifs only having large TI scores in one cell class (Extended Data Fig. 7l, left panel). For example, the *Tshz1* gene body mCH shows a striking difference of diversity between excitatory and inhibitory cells (Extended Data Fig. 7m), suggesting that it may function in shaping inhibitory subtypes, but not excitatory subtypes. Similarly, bHLH DNA binding motifs show much higher TI scores for excitatory subtypes compared with inhibitory (Extended Data Fig. 6l, right panel). While genes in this TF family such as *Neurod1/2* have long been known to participate in excitatory neuron development, they have not been reported to regulate GABAergic neuron differentiation[105].

2.19.4 Spatial axis of DG granule cells

Neurogenesis is differentially distributed across the dorsal-ventral axis of the DG, with newborn neurons enriched in the dorsal pole of the DG[106,107]. Bulk tissue samples dissected from the dorsal DG had ~2-fold lower global mCH compared to the ventral DG, suggesting that the gradient in global mCH we observed could be related to their dorsal-ventral location[71]. Consistent with this, dorsal-hypo-DMRs from the bulk tissue samples[71] overlapped more with +DMRs (15% overlap, $P < 0.001$, two-sided permutation test) than -DMRs (4%, $P < 0.001$) (Extended Data Fig. 10e). Similarly, ventral-hypo-DMRs overlapped more with -DMRs (13%, $P < 0.001$) than +DMRs (1.7%, $P = 0.004$). These moderate overlaps suggest that the axis of the global mCH gradient can be partially explained by the traditionally defined dorsal-ventral axis. Further understanding of the

relationship between these two axes will require spatially resolved transcriptome/epigenome data from cells with defined birthdates[82].

2.19.5 Artificial Neuron Network trained for predicting non-neuronal cells.

Using the same network structure and features, we achieved 42% accuracy to predict the cell body location of non-neuronal cells. The accuracy of the model was highly cell types dependent, with stronger predictability in astrocytes (ASC, 62%) and oligodendrocytes (ODC, 52%) that are comparable to cortical inhibitory neurons (Extended Data Fig. 11d). These findings are supported by evidence that astrocytes are derived from regionally patterned radial glia[108], while the signature of oligodendrocytes depends on the local environment[109]. It is worth noting that unsupervised clustering did not separate non-neuronal cells from different regions into different clusters, which further emphasizes the utility of supervised analyses to study the regional specificity of these cell types.

CHAPTER 3

SINGLE NUCLEUS MULTI-OMICS IDENTIFIES HUMAN CORTICAL CELL REGULATORY GENOME DIVERSITY

3.1 Abstract

Single-cell technologies measure unique cellular signatures but are typically limited to a single modality. Computational approaches allow the fusion of diverse single-cell data types, but their efficacy is difficult to validate in the absence of authentic multi-omic measurements. To comprehensively assess the molecular phenotypes of single cells, we devised single-nucleus methylCytosine, chromatin Accessibility and Transcriptome sequencing (snmCAT-seq) and applied it to post-mortem human frontal cortex tissue. We developed a cross-validation approach using multi-modal information to validate fine-grained cell types and assessed the effectiveness of computational data fusion methods. Correlation analysis in individual cells revealed distinct relations between methylation and gene expression. Our integrative approach enabled joint analyses of the methylome, transcriptome, chromatin accessibility and conformation for 63 human cortical cell types. We reconstructed regulatory lineages for cortical cell populations and found specific enrichment of genetic risk for neuropsychiatric traits, enabling the prediction of cell types that are associated with diseases.

3.2 Design

Simultaneous DNA methylcytosine and transcriptome sequencing using snmCAT-seq allows RNA and DNA molecules to be molecularly partitioned by incorporating 5'-methyl-dCTP instead of dCTP during reverse transcription of RNA (Fig. 3.1A). We treated single cells/nuclei with Smart-seq or Smart-seq2 reactions for *in situ* cDNA synthesis and amplification of full-length

cDNA (Table 3.1)[110,111]. Replacing dCTP by 5'-methyl-dCTP results in fully cytosine-methylated double-stranded cDNA amplicons. Following bisulfite treatment converting unmethylated cytosine to uracil, sequencing libraries containing both cDNA- and genomic DNA-derived molecules were generated using snmC-seq2 (Table 3.1)[3,28]. With this strategy, all sequencing reads initially derived from RNA are completely cytosine methylated and do not show C to U sequence changes during bisulfite conversion. By contrast, more than 95% of cytosines in mammalian genomic DNA are unmethylated and converted by sodium bisulfite to uracils that are read during sequencing as thymine[112]. In this way, sequencing reads originating from RNA and genomic DNA can be distinguished by their total mC density. Since 70-80% of CpG dinucleotides are methylated in mammalian genomes, we used the read-level non-CG methylation (mCH) to uniquely partition sequencing reads into RNA or DNA bins. Specifically, we expect the level of mCH for all RNA-derived reads to be greater than 90%, while for DNA-derived reads the level is no more than 50% even considering the enrichment of mCH in adult neurons[2]. Using this threshold, only 0.02% ± 0.01% of single-cell methylome reads (n=100 cells profiled with snmC-seq2[15]) were misclassified as transcriptome reads and only 0.23% ± 0.17% of single-cell RNA-seq reads (n=100 cells profiled with Smart-seq[34]) were misclassified as methylome reads (Supplementary Fig. 3.1A). For a snmCAT-seq profile containing 90% of methylome reads and 10% of transcriptome reads, the estimated specificity for classifying methylome and transcriptome reads is 99.997% and 99.97%, respectively. These results show that RNA- and DNA-derived snmCAT-seq reads can be effectively separated. We extended the multi-omic profiling to include a measure of chromatin accessibility by incorporating the Nucleosome Occupancy and Methylome-sequencing assay (NOMe-seq, Fig. 3.1A, Table 3.1)[20,23,113,114]. In the snmCAT-seq assay, regions of accessible

chromatin are marked by treating bulk nuclei with the GpC methyltransferase M.CviPI prior to fluorescence-activated sorting of single nuclei into the reverse transcription reaction (Fig. 3.1A).

A detailed bench protocol for snmCAT-seq and future updates to the method can be found at https://www.protocols.io/view/snmcat-v1-bwubpesn.

3.3 Joint analysis of RNA and DNA methylome in cultured human cells.

We first tested the efficacy of the joint profiling of RNA and DNA methylome by applying snmCAT-seq to either single whole cells or single nuclei of cultured human H1 embryonic stem cells and HEK293 cells (Supplementary Table 3.1-3.2), without the labeling of accessible chromatin using GpC methyltransferase. snmCAT-seq transcriptome profiling detected 4,220 ± 1,251 genes from single whole cells using exonic reads while detected 4,531 ± 1,888 genes using both exonic and intronic reads (Supplementary Fig. 3.1B). Similar to previously reported single-nuclei RNA-seq datasets, a minor fraction (17.3 ± 6.1%) of snmCAT-seq transcriptome reads generated from single nuclei were mapped to exons, whereas 68.1 ± 15.2% of snmCAT-seq reads generated from single cells were mapped to exons (Supplementary Fig. 3.1C). Transcriptome reads accounted for 22.2 ± 13.6% and 9.2 ± 6.5% of all mapped reads for snmCAT-seq data generated from single cells or nuclei, respectively (Supplementary Fig. 3.1D). The snmCAT-seq profiles could clearly separate H1 and HEK293 cells by their transcriptomic signatures[78] (Supplementary Fig. 3.1E-F) and recapitulate specific gene expression signatures (Supplementary Fig. 3.1G).

To assess whether the two cell types could be distinguished using mC signatures derived from snmCAT-seq, tSNE was performed using the average CG methylation (mCG) level of 100 kb non-overlapping genomic bins (Supplementary Fig. 3.1H-I). As exemplified by the NANOG and CRNDE loci (Supplementary Fig. 3.1J), snmCAT-seq produced mC profiles highly consistent with data generated from bulk methylomes[115]. snmCAT-seq data generated from both single cells

and single nuclei identified global mC differences between H1 and HEK293T cells, showing that H1 cells are more methylated in both CG (83.6%) and non-CG (1.3%) contexts compared with HEK293T cells (mCG: 60.1%, no significant mCH detected, Supplementary Fig. 3.1K-N)[112]. To examine whether local mC signatures can be recapitulated in snmCAT-seq, we identified differentially methylated regions (DMRs) from bulk H1 and HEK293 methylomes. Plotting mCG levels measured using snmCAT-seq profiles across DMRs showed highly consistent patterns compared to bulk cell methylomes (Supplementary Fig. 3.1O-P).

3.4 Multi-omic profiling of postmortem human brain tissue with snmCAT-seq

We generated snmCAT-seq profiles from 4,358 single nuclei isolated from postmortem human frontal cortex tissue from two young male donors (21 and 29 years old, Supplementary Table 3.3-3.4). The data quality was similar to datasets generated from nuclei isolated from cultured human cells with respect to the fraction of sequencing reads mapped to the transcriptome (Supplementary Fig. 3.2A), the fraction of transcriptome reads mapped to introns and exons (Supplementary Fig. 3.2B) and the number of genes detected (Fig. 3.1B). Compared with snmC-seq and snmC-seq2 data generated from human single nuclei[3,28], the DNA methylome component of snmCAT-seq had comparable genomic coverage (Fig. 3.1C), mapping efficiency (Fig. 3.1D), and showed only moderately reduced library complexity (Fig. 3.1E) with similar coverage uniformity (Fig. 3.1F-G).

To compare each data modality profiled by snmCAT-seq with their corresponding single modality assays, we first identified 20 cell types by multi-modal clustering analysis using transcriptome, methylome and chromatin accessibility. We used RNA abundance across the gene body for the transcriptome, mCH and mCG level of chromosome non-overlapping 100kb-bins, and binarized NOMe-seq signal of 5kb bins for chromatin accessibility (See methods). For

snmCAT-seq, the HCH context was counted for CH methylation and HCG is counted for CG methylation to exclude GCH and GCG sites that can be methylated by M.CviPI. We identified highly variable features and calculated principal components separately for each modality. We observed substantial differences across data modalities in their ability to resolve cell populations using the top 10 principal components (Supplementary Fig. 3.2G). Therefore, only informative principal components from each data modality were concatenated as the input features for multi-modal clustering and visualization using Uniform Manifold Approximation and Projection (UMAP) [56] of the three data types (Fig. 3.1H-I). The selection of informative principal components for multi-modal clustering is agnostic to the type of molecular profile being analyzed and could be generalized to other multi-omic approaches. We found non-CG methylation as the most distinguishing measurement explaining 63.7% of the total variance, while CG methylation, RNA abundance and NOMe-seq signal each explained 15.8%, 20.2% and 0.4% of the variance, respectively (Supplementary Fig. 3.2C). These cell types were effectively separated by performing dimensionality reduction using each data type (Supplementary Fig. 3.2D-F). The comparison of homologous clusters between snmCAT-seq transcriptome and snRNA-seq (Supplementary Table 3.5) shows a robust global correlation: Pearson r = 0.82 for both PV-expressing inhibitory neurons (MGE_PVALB, $p = 1x10^{-145}$) and superficial layer excitatory neurons (L1-3 CUX2, $p = 3x10^{-301}$) (Supplementary Fig. 3.2H-I). Moreover, highly consistent expression patterns of cell-type signature genes were observed (Fig. 3.1J).

To test whether snmCAT-seq transcriptome data can be integrated with snRNA-seq (Table 3.1), we integrated snRNA-seq and the transcriptome component of snmCAT-seq using a mutual nearest neighbor approach[116] (Fig. 3.1K-L). The integration confirmed that the cell types identified using the snmCAT-seq transcriptome are strongly correlated with the cell types found using

snRNA-seq. Similar to the transcriptome, both mCH and mCG profiles correlate strongly between methylomes generated with snmCAT-seq and snmC-seq2 either globally (Supplementary Fig. 3.2J-K) or at cell-type-specific signature genes (Fig. 3.1M).

The presence of high levels of mCH in the human brain confounds the analysis of chromatin accessibility using methylation at GpC sites (GmC). However, we found that in GCT and GCC sequence contexts, GmC introduced by M.CviPI greatly surpasses the levels of native methylation by 6.4 and 16-fold, respectively (Supplementary Fig. 3.2L). Thus for snmCAT-seq, we focused our analyses of chromatin accessibility on GmC at GCY (Y=C or T) sites in the genome. We further developed a computational strategy to first identify significantly methylated GCY (GmCY) sites using a Hidden Markov Model approach[117], followed by the calling of open chromatin regions using the frequency of GmCY sites. Chromatin accessibility measured by the frequency of GmCY sites correlates closely with snATAC-seq signal at cell-type-specific open chromatin sites both globally (Fig. 3.1N and Supplementary Fig. 3.2M-N, p-value < 2.2×10^{-308}) and at cell-population specific genes such as *BDNF*, *POU3F2*, *DLX2/3* and *SOX11* (Supplementary Fig. 3.2O). In addition, open chromatin regions identified with GmCY frequency overlapped substantially with regions found using snATAC-seq (Supplementary Fig. 3.2P-Q). In summary, snmCAT-seq can simultaneously profile transcriptome, methylome and chromatin accessibility in single nuclei, accurately recapitulating cell-type signatures for each data type.

3.5 Paired RNA and mC profiling enables cross-validation and quantification of over-/under-splitting for single-cell clusters

A fundamental challenge for single-cell genomics is to objectively determine the number of biologically meaningful clusters in a dataset[39]. Cross-dataset integration of the same data type, or fusion of distinct data types can be used to assess cluster robustness, but it may be limited by

systematic differences between the datasets or modalities used[118]. To address this, we devised a novel cross-validation procedure using matched transcriptome and DNA methylation information to estimate the number of reliable clusters supported by both modalities in snmCAT-seq data (3,898 neurons, Fig. 3.2A). We first clustered the cells with different resolutions using mC information, then tested how well each clustering is supported by the matched transcriptome profiles. We used the cross-validated mean squared error between the RNA expression profile of individual cells and the cluster centroid as a measure of cluster fidelity (Fig. 3.2B-C). Mean squared error for cells in the training set decreased monotonically with the number of clusters, whereas over-clustering leads to an increase in mean squared error for the test set. The U-shaped mean squared error curve shows that aggressively splitting cells into fine-scale clusters based on mC signatures is not supported by corresponding RNA signatures. The cluster resolution with the minimum mean squared error represents the finest subdivision of cells that is well supported across both modalities. In addition to directly evaluating error on a test set, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) of the training set were also applied to estimate test error (see methods, Fig. 3.2B-C and Supplementary Fig. 3.3A). Indeed, AIC curves largely overlapped with test errors and gave similar estimates of the optimum cluster numbers; BICs consistently reach smaller optimums than the other two metrics as they penalize model complexity more stringently. Using these approaches, we found a range of 20-50 clusters with strong multimodal support in the current snmCAT-seq dataset (Fig. 3.2B-C). The same approach can also be applied to each individual modality separately, to identify the number of clusters supported by DNA methylation features and by RNA features, respectively (Supplementary Fig. 3.3A).

The above approach objectively identified a range of appropriate cluster resolutions for the

whole dataset. To assess the quality of individual clusters, we further developed metrics to quantify over-splitting and under-splitting (Methods; Fig. 3.2D, Supplementary Fig. 3.3B). After jointly embedding mC and RNA data in a common low-dimensional space[119], we define a graph connecting each cell to *k* cells with the greatest cross-modality similarity (called *k*-partners). An over-splitting score was calculated as the fraction of each cell's *k*-partners that are not in the same cluster (Methods; Fig. 3.2D, E). We assessed the over-splitting of 17 major neuronal clusters and 52 neuronal sub-clusters (Fig. 3.4 and Supplementary Table 3.6) identified by single-cell methylomes and found major clusters resemble ideal, homogeneous clusters (simulated by shuffling gene features) with low over-splitting scores (Fig. 3.2E, Supplementary Fig. 3.3C, E), with only 1/17 major clusters have an over-splitting score ≥0.6. Most sub-clusters also had relatively little over-splitting; only 10/52 sub-clusters had an over-splitting score ≥0.6 (Supplementary Fig. 3.3C, E).

To assess under-splitting, we reasoned that if a cluster cannot be further split (no under-splitting) all its cells should be statistically equivalent. Therefore, each cell's mC profile should be no more correlated with its own RNA profile than with the RNA profile of any other cell of the same type. By contrast, an under-split cluster will contain some residual discrete or continuous variation that is correlated between modalities. We tested this by defining the self-radius (the distance between mC and RNA profiles of the same cell, see Methods) for each cell and comparing the distribution of self-radii for each cluster with that expected for homogeneous clusters using a permutation procedure. We found that major neuronal clusters had substantial within-cluster variation across cells, indicating that they are under-split (Fig. 3.2F, Supplementary Fig. 3.3D, F). By contrast, subtypes resembled ideal (shuffled) clusters to a greater degree. Combining both scores, we quantitatively mapped the lumper-splitter tradeoff in terms of the degree of over- and

under-splitting for each major type or subtype (Fig. 3.2G).

The fusion of single-cell genomic data across multiple data types has been a focus of recent computational studies, yet existing methods lack validation on ground truth from experimental single-cell multi-omic datasets[120]. By treating snmCAT-seq transcriptome and mC profiles as if they were generated from different single cells, we could test the performance of computational data fusion using Seurat[36], Harmony[52], Scanorama[37], LIGER[38] and Single Cell Fusion (Methods; Fig. 3.2H, Supplementary Fig. 3.3G-K, first row). To evaluate the fusion at the cell-level, we calculated the self-radius as mentioned above and determined mis-fused events by normalized self-radius > 0.3 (Supplementary Fig. 3.3G-K, second row). We also quantified the cluster level accuracy as the fraction of cells whose transcriptome and mC profiles were assigned to the same cluster (Fig. 3.2I, Supplementary Fig. 3.3G-K, third row). Overall, the Single Cell Fusion and Seurat outperform the other tools, with the Single Cell Fusion achieving the lowest mis-fusion ratio (5.7%) and highest overall major cell-type level accuracy (87.3%) (Fig. 3.2I-J, Supplementary Fig. 3.3G). We also tested the Single Cell Fusion accuracy at the subtype level. As expected, computational fusion of fine-grain clusters was less accurate (62.6%) and more variable across clusters (Fig. 3.2I), potentially because of the greater degree of over-clustering (Fig. 3.2E).

3.6 Diverse correlation between gene body mCH and gene expression

Using the paired profiling of transcriptome and mC by snmCAT-seq, we found diverse patterns of correlation between mCH and gene expression across thousands of single cells. Fig. 3.3A shows examples of three distinct types of correlations between gene body mCH and gene expression. *KCNIP4* shows an inverse correlation between mCH and RNA across a broad range of cell types. *ADARB2* is a marker gene for CGE-derived inhibitory cells and showed a strong inter-cluster correlation, but no intra-cluster correlation between mCH and RNA. Finally, *GPC5*

has a gradient of mCH across clusters (low in CGE VIP, high in L1-3 CUX2), but no corresponding pattern of differential gene expression across cell types. Applying this correlation analysis to all 13,637 sufficiently covered genes, we found that 38% (n=5,145) have a significant negative correlation between mCH and RNA (mCH-RNA coupled, FDR < 5%). The majority of genes (62%) had no apparent correlation that could be distinguished from noise (mCH-RNA uncoupled, Fig. 3.3B). The pattern of correlation was highly consistent between the specimens we profiled and robust with respect to normalization and data smoothing (Supplementary Fig. 3.4A-H). We found mCH-RNA correlation is correlated (r=0.63) with mCG-RNA correlation, consistent with previous findings[3,4] (Supplementary Fig. 3.4I). Genes with a significant correlation between mCH and gene expression are longer, more highly expressed, show greater chromatin accessibility, and are enriched in neuronal functions (Supplementary Fig. 3.4J-M).

We further investigated the factors that determine the degree of correlation between mCH and RNA for each gene. We reasoned that housekeeping genes with a strong expression and little variation across cell types would show weak mCH-RNA correlation, whereas mCH-RNA coupling is enriched in genes with cell-type-specific expression. We quantified the cell-type specificity of gene expression and DNA methylation by calculating the fraction of variance in gene expression explained by cell type ($RNA\ \eta^2$ and $mCH\ \eta^2$, Fig. 3.3C-E and Supplementary Fig. 3.4N). Consistent with our hypothesis, genes with greater $RNA\ \eta^2$ had a stronger inverse-correlation between mCH and RNA (Fig. 3.3D-E). Notably, we found a large number of genes (n=1,243) with strong gene body mCH diversity across cell types ($mCH\ \eta^2$> 0.25) but no apparent correlation between mCH and RNA (r < -0.03) (box in Fig. 3.3C). This suggests that the lack of correlation between mCH and gene expression is driven by variability in gene expression within cell types despite conserved DNA methylation signatures.

The accumulation of mCH in the frontal cortex starts from the second trimester of embryonic development and continues into adolescence[2,121]. The developmental dynamics of mCH motivated us to compare the developmental expression of mCH-RNA coupled and uncoupled genes. We found that mCH-RNA uncoupled genes, on average, are highly expressed during early fetal brain development (PCW 8-9) and are later repressed, whereas the expression of mCH-RNA coupled genes is moderately increased during development (Fig. 3.3F). Consistently, developmentally down-regulated genes are significantly enriched in the mCH-RNA uncoupled group (Fig. 3.3G). We speculated that the developmentally down-regulated genes may be repressive by alternative epigenomic marks such as histone H3K27 trimethylation (H3K27me3), which leads to the uncoupling of RNA and gene body mCH. By binning all the genes by their expression dynamics during brain development, we indeed found the promoter of both down- and up-regulated genes are enriched in H3K27me3 and depleted in active histone marks (Fig. 3.3H and Supplementary Fig. 3.4O). We directly compared mCH-RNA correlation and H3K27me3 in purified human cortical glutamatergic and GABAergic neurons[122], and found genes with strong H3K27me3 signal clearly show weak correlations between gene body mCH and gene expression (e.g. CDC27 Fig. 3.3I-K). In summary, although mCH and gene expression are clearly inversely correlated at a global scale, substantial variations can be observed from genes to genes at a single-cell level and can be partially explained by the presence of alternative epigenetic pathways such as polycomb repression.

3.7 Multi-omic integration of chromatin conformation, transcriptome, methylome and chromatin accessibility

The snmCAT-seq dataset for the human frontal cortex was combined with previously published human frontal cortex datasets (Supplementary Table 3.3): sn-m3C-seq which

simultaneously profiles mC and chromatin conformation[24] and snmC-seq methylomes for single neurons[3]. We additionally generated new snmC-seq and snmC-seq2 data for the frontal cortex from two independent donors (Supplementary Table 3.3). These datasets can be readily integrated using single-nucleus methylomes as the common modality (Fig. 3.4A). To identify both major cell types and subtypes of frontal cortex, we integrated 15,030 single-cell methylomes generated by snmC-seq (n=5,131), snmC-seq2 (n=1,304), snmCAT-seq (n=4,358) and sn-m3C-seq (n=4,238) prior to the clustering analysis (Supplementary Table 3.6). We used an iterative clustering approach to identify 20 major cell populations including 9 excitatory neuron types, 8 inhibitory neuron types and 3 non-neuronal cell types in the first round of clustering (Fig. 3.4B-C). A second round of iterative clustering of each major cell type identified 63 cell subtypes, including 19 excitatory neuronal subtypes, 33 inhibitory neuronal subtypes and 11 non-neuronal cell subtypes (Fig. 3.4B-C). Each fine-grained cell subtype can be distinguished from any other cell type by at least 10 mCH signature genes for neuronal clusters, or 10 mCG signature genes for non-neuronal clusters. Consistent with our previous results[3] as well as transcriptomic studies[84], we found greater diversity among human cortical inhibitory neurons than excitatory cells (Fig. 3.4C). The methylome data generated by these diverse multi-omic methods and from multiple donors were uniformly represented in major cell type and subtype clusters (Fig. 3.4D).

We next performed fusion of single-cell methylome and snATAC-seq (Fig. 3.4E, Supplementary Table 3.7) profiles by transferring the cluster labels defined by mC into ATAC-seq cells using a nearest neighbor approach[116] that was adapted for epigenomic data and implemented in a new software package (https://github.com/mukamel-lab/SingleCellFusion, see methods). For each cell population, we reconstructed four types of molecular profiles: transcriptome (from snmCAT-seq), methylome (from snmC-seq1/2 and sn-m3C-seq), chromatin

accessibility (from snmCAT-seq mGCY frequency or snATAC-seq) and chromatin conformation sn-m3C-seq[24,123] (Fig. 3.4F). This integrative analysis revealed extensive correlations across epigenomic marks at cell-type signature genes. For example, *ADARB2* is a signature gene of inhibitory neurons derived from the caudal ganglionic eminence (CGE). In CGE-derived VIP neurons, *ADARB2* was associated with abundant transcripts, reduced mCG and mCH, and distinct chromatin interactions compared with other neuron types (Fig. 3.4F). In contrast, in VIP neurons the *MEF2C* locus showed lower transcript abundance (TPM - L1-3 CUX2: 75.8, L4-5 FOXP2: 80.2, MGE PVALB: 77.5, CGE VIP: 49.0), reduced chromatin interaction, and more abundant gene body mCG (Fig. 3.4F). Although nearly identical open chromatin sites were identified at the promoter regions of *ADARB2* and *MEF2C* using snmCAT-seq GpC methylation and snATAC-seq, the two methods revealed distinct cell-type specificity of chromatin accessibility. At the *ADARB2* promoter, snATAC-seq but not the snmCAT-seq GpC methylation profile showed enriched chromatin accessibility in VIP neurons. However, at *the MEF2C* promoter, snmCAT-seq GpC methylation indicated a depletion of open chromatin in VIP neurons which is more consistent with the reduced gene expression and increased gene body mCG in this inhibitory cell population. The cause of these differences in measures of chromatin accessibility is not clear, and further work is needed to clarify their respective sensitivities and biases[117].

3.8 snmCAT-seq identifies RNA and mC signatures of neuronal subtypes.

The integration of 15,030 single-cell methylomes allowed the determination of fine-grained brain cell subtypes with a sensitivity comparable to snRNA-seq (Fig. 3.4B-C). For example, we identified 15 subtypes of CGE-derived inhibitory neurons using single-cell methylomes, whereas 26 subtypes were identified by snRNA-seq[84]. To ask whether snmCAT-seq can recapitulate the molecular signatures of neuronal subtypes, we integrated snmCAT-seq

transcriptome with snRNA-seq datasets for inhibitory neurons followed by joint clustering (Fig. 3.5A-B, S5A). Individual nuclei profiled with snmCAT-seq transcriptome and snRNA-seq were uniformly distributed across joint clusters except for cluster 13 (Fig. 3.5B-C), suggesting that in general, the snmCAT-seq transcriptome recapitulates the full range of inhibitory neuron diversity. Cluster 13 contained snmCAT-seq data with lower numbers of transcriptome reads (Supplementary Fig. 3.5B), but the methylome profiles of the same cells showed acceptable quality and were robustly co-clustered with other inhibitory neurons (Supplementary Fig. 3.5B-C). Similarly, integration of snmCAT-seq transcriptomes and snRNA-seq for excitatory neurons and non-neuronal cells showed that brain cell type diversity across all cell classes can be recapitulated from the snmCAT-seq transcriptome profiles (Supplementary Fig. 3.5D-K). We further compared the expression of a panel of signature genes for inhibitory neuron subpopulations and found that snmCAT-seq transcriptome and snRNA-seq identified highly consistent expression patterns (Fig. 3.5D). Lastly, we identified cell-type marker genes across inhibitory neuronal populations using transcriptome profiles generated with either snmCAT-seq or snRNA-seq (Supplementary Table 3.8). Analysis of the marker genes using a database curated for neuronal functions - SynGO[124] revealed consistent enrichment in ontological categories associated with synaptic signaling and synapse organization for inhibitory neuron marker genes identified with both snmCAT-seq transcriptome and snRNA-seq data (Fig. 3.5E-F).

3.9 DNA methylation signatures of hierarchical transcription factor regulation in neural lineages

Temporally regulated expression of transcription factors (TF) during specific developmental stages is critical for neuronal differentiation[125,126]. We hypothesized that the cell-type hierarchy reconstructed from mC information reflects the developmental lineage of human

cortical neurons. If so, then key transcription factors that specify neuronal lineage can be identified for each branch of the hierarchy. We separately constructed hierarchies for inhibitory and excitatory neurons based on the concatenated principal components of mCH and mCG (Fig. 3.6A and S6A). The inhibitory neuron hierarchy comprises two major branches corresponding to medial ganglionic eminence (MGE) and caudal ganglionic eminence (CGE) derived cells (Fig. 3.6A). These major populations contain intermediate neuronal populations such as PVALB-expressing Basket Cell (BC) and Chandelier Cell (ChC), or the recently reported LAMP5-expressing Rosehip neurons (Fig. 3.6A)[127]. At the finest level, the hierarchy contains 33 neuronal subtypes (Fig. 3.6A). To identify TFs involved in the specification of neuronal lineages, we compared three levels of molecular information for each of 1639 human TFs[128] between the daughter branches (Fig. 3.6B). To assess the genome-wide DNA binding activity of the TF at regulatory elements, we used enrichment of DNA binding sequence binding motifs in differentially methylated regions (DMRs). To assess TF gene expression, we used both mRNA expression and TF gene body mCH level.

Our integrated strategy taking advantage of matched information for TF motif enrichment, transcript abundance and TF gene body mCH level allowed us to distinguish the relative importance of closely related TFs sharing a common binding motif based on their cell-type-specific expression[129] (Fig. 3.6C). For example, we predicted that NFIB and NFIX contribute to CGE lineage specification since they show greater RNA abundance and stronger gene body mCH depletion than closely related TFs NFIA and NFIC. We systematically applied this approach across the inhibitory neuron hierarchy, using 579 curated motifs from the JASPAR 2018 CORE vertebrates database (Supplementary Fig. 3.6E-H)[85]. Many predicted lineage regulators were homologous to cell-type lineage regulators in mouse cortical development, such as NFIX, NFIB for CGE-derived neurons (Supplementary Fig. 3.6E), or LHX6, SOX6 and SATB1 for MGE-

derived neurons (Supplementary Fig. 3.6E)[126,130]. The motifs of some TFs were also recurrently enriched in multiple lineages. For example, the NFIB gene[131] is not only specific to CGE neurons but also highly expressed and hypo-methylated in PV-expressing chandelier cells (ChC) but not basket cells (BC) (Fig. 3.6D). The same expression pattern of NFIB was found in a comparison of mouse ChC - BC[130]. These findings provide cogent evidence that the conserved major cell types of human and mouse[84] also have shared basic rules of TF regulation. The same TF gene may perform multiple roles in different cell type lineages.

Previous studies including ours have found that discrete genomic regions with reduced mCG (hypomethylated DMRs) mark active regulatory elements[4,41,132,133]. We expected that TF binding motifs would be enriched in hypomethylated DMRs for cell types where the TF gene is actively expressed and has low gene-body mCH. However, we identified several TFs with an opposite pattern: their binding motif was enriched in the hypomethylated DMRs of the alternative lineage showing low TF expression and high gene body mCH. For example, the motifs of NR2F1 and PBX1 were enriched in the hypomethylated DMRs of ChC, but both TFs were actively expressed in BC and not ChC (Fig. 3.6D). Similarly, the PKNOX2 motif was enriched in hypomethylated DMRs of VIP cells, yet *PKNOX2* is preferentially expressed in NDNF neurons (Fig. 3.6E). These data suggest that certain TFs can preferentially bind to hypermethylated regions (i.e. hypomethylated regions in the alternative lineage). This non-classical preference for methylated binding sites has been extensively demonstrated in *in vitro* studies[6,134]. In particular, Yin et al. used an *in vitro* assay to bind each recombinant TF protein to a pool of synthetic DNA (methyl-SELEX). They identified hundreds of TFs whose binding is inhibited (MethylMinus) or promoted (MethylPlus) by the presence of methylated CpG sites in their binding motifs. We analyzed the *in vivo* binding of MethylPlus TFs to hypermethylated DNA by analyzing chromatin

accessibility measured by the snmCAT-seq NOMe-seq profile (Fig. 3.6F) as well as snATAC-seq (Supplementary Fig. 3.6I). We quantified the average chromatin accessibility at TF binding motifs that are lowly methylated (overlapping with hypomethylated DMRs) or highly methylated (overlapping with hypermethylated DMRs) (Fig. 3.6F and Supplementary Fig. 3.6I), and used the difference in chromatin accessibility to determine the *in vivo* sensitivity of each TF to cytosine methylation. Using both chromatin accessibility assays (NOMe-seq and ATAC-seq), we found a general agreement between our *in vivo* approach and the *in vitro* methyl-SELEX results with MethylMinus TFs showing enrichment in the upper part of Fig. 3.6F and S6I (e.g. ETV1 in Fig. 3.6J), which showed a greater difference in chromatin accessibility between lowly and highly methylated TF motifs (Fig. 3.6H and Supplementary Fig. 3.6K). Consistently, MethylPlus TFs are strongly depleted in the upper part of Fig. 3.6F and Supplementary Fig. 3.6I (Fig. 3.6G and Supplementary Fig. 3.6J). Therefore, our joint analysis of mC and chromatin accessibility using snmCAT-seq provided *in vivo* evidence for the modulation of TF binding by cytosine methylation.

Lastly, we examined the correlation between chromatin accessibility and the presence of CA dinucleotide in the TF binding motifs, since CA is the predominant sequence context of mCH in the human brain[2]. Intriguingly, we found a significant enrichment of TF binding motifs containing CA dinucleotides in the lower part of Fig. 3.6F and Supplementary Fig. 3.6I, using either NOME-seq and ATAC-seq to quantify chromatin accessibility (Fig. 3.6I and Supplementary Fig. 3.6L), suggesting the accessibility of TF binding motifs containing CA is less affected by mC. Across all TF binding motifs examined, the accessibility of motifs containing both CA and CG dinucleotides (CA+ CG+, p-value = 1×10^{-4}, e.g. ATF4, Fig. 3.6L) or only CA (CA+ CG-, p-value = 5.7×10^{-6}, e.g. RARB, Fig. 3.6K) show significantly less sensitivity to mC than motifs containing CG dinucleotides only (CA- CG+) (Fig. 3.6M). The results suggest certain TFs may be able to

bind hypermethylated regions through the interaction with mCA sites. The modulation of TF binding by mCA has not been systematically explored since previous studies have focused on the effect of mCG sites[6,134].

3.10 Cortical cell regulatory genomes predict developmental and adult cell types associated with neuropsychiatric diseases

The strong enrichment of disease heritability in gene regulatory elements has allowed the prediction of disease-associated cell types using epigenomic signatures[135], including neuropsychiatric disorders[122]. By reconstructing mC and open chromatin maps from single-cell profiles, we used LD score regression partitioned heritability to infer the relevant cell types for a set of neuropsychiatric traits using DMRs and ATAC-seq peaks (Supplementary Table 3.9-3.10)[135]. To capture regulatory elements active during early development which may be implicated in psychiatric disease, we further included lowly methylated regions identified from bulk fetal (PCW 19) human cortex methylome[121] and DNase-seq peaks identified from fetal brain samples[136]. We first compared the set of DMRs in each brain cell type individually to a baseline containing DMRs identified across non-brain human tissues[41]. Using a statistical threshold of FDR < $1x10^{-5}$, we identified 72 disease-cell type associations across 21 cortical cell types or bulk samples for 16 neuropsychiatric traits (Supplementary Fig. 3.7A). Each association corresponds to a significant enrichment of disease heritability within the corresponding cell type's active regulatory regions. By contrast, no association was found in DMRs identified from 18 bulk non-brain tissues (Supplementary Fig. 3.7A)[41]. This result strongly suggests our partitioned heritability analysis has correctly identified the brain as the relevant tissue types for neuropsychiatric traits.

To discern the relative enrichment of disease risk between brain cell types, we further constructed multiple regression models including all adult brain cell types and the fetal brain (Fig.

3.7A-D). In most cases, our partitioned heritability analyses enhanced the cell-type resolution compared to previous efforts. For example, using single-cell RNA-seq datasets, the genetic risk of schizophrenia was previously mapped to broad cortical neuronal populations including neocortical somatosensory pyramidal cells, and cortical interneurons[122,137]. Our analysis further identified the enrichment of schizophrenia heritability in multiple types of intratelencephalic (IT) neuron types (L1-3 CUX2, L4-5 FOXP2 and L5-6 PDZEN4), in addition to a medial ganglionic eminence derived inhibitory cell type (MGE CALB1) (Fig. 3.7A). Intriguingly, the heritability of bipolar disorder was specifically enriched in a deep layer neuron type L5-6 PDZEN4 (Fig. 3.7B). We also found a specific enrichment of autism spectrum disorder risk in a deep-layer thalamic-projecting neuronal population L6 TLE4 (Fig. 3.7C). By contrast, the heritability of educational attainment was broadly distributed across multiple types of neurons including excitatory cells (L1-3 CUX2, L4 PLCH1 and L6 TLE4) and inhibitory neurons derived from both caudal (CGE LAMP5) and medial ganglionic eminence (MGE CALB1) (Fig. 3.7D). Consistent with the neurodevelopmental hypothesis that gene misregulation during brain development underlies certain psychiatric disorders[138], lowly methylated regions in fetal cortex DMRs are enriched in the heritability for schizophrenia and educational attainment (Supplementary Fig. 3.7A). However, the partitioned heritability analysis using the fetal cortex sample is likely underpowered due to the cell-type heterogeneity. To corroborate our results generated using LD score regression partitioned heritability, we applied RolyPoly[139] to prioritize trait-relevant cell types using GWAS SNP effect sizes and cell-type-specific mCG levels at DMRs (Supplementary Fig. 3.7E-F). The analysis using RolyPoly recapitulated a number of predictions, such as the association between schizophrenia and L5-6 PDZRN4 cells and MGE derived inhibitory cells; bipolar disorder with L5-6 PDZRN4

cells; autism spectrum disorder with L6 TLE4 cells, as well as educational attainment with the L1-3 CUX2 population (Supplementary Fig. 3.7E-F).

We performed partitioned heritability analyses using three complementary types of molecular signatures (Supplementary Fig. 3.7B-D): genes with cell-type-specific expression (Supplementary Fig. 3.7B,D), DMRs and open chromatin regions identified with both snATAC-seq and NOMe-seq (Supplementary Fig. 3.7B,C) (or DNase-seq peaks for the prenatal brain sample). To our surprise, the results obtained using DMRs and ATAC-seq peaks were substantially different. For example, the partition of schizophrenia heritability across DMRs identified enrichment in four adult cell types in addition to the fetal cortex (Fig. 3.7A), whereas the analysis using open chromatin regions only found enrichment in L1-3 CUX2 cells and the fetal brain (Fig. 3.7E). To understand this discrepancy, we stratified DMR regions into two groups [DMR (ATAC-pos) and DMR (ATAC-neg)] by their overlap with open chromatin regions. Partitioned heritability across the stratified DMR regions revealed that in adult cells, DMR regions without open chromatin signature are more strongly enriched in heritability for the neuropsychiatric traits (Fig. 3.7I-L). In the fetal cortex, however, a stronger enrichment of schizophrenia and educational attainment heritability was found in DMRs associated with open chromatin.

We speculate that DMRs without open chromatin contain vestigial enhancers[7], which contribute to the enrichment of disease heritability. Vestigial enhancers are active regulatory elements during embryonic development but become dormant in adult tissues[7]. However, vestigial enhancers remain lowly methylated in adult tissues and can be identified as DMRs. Thus, vestigial enhancers can be strongly enriched in the genetic risk of neuropsychiatric traits since these regions are active regulatory elements during brain development. We identified the fraction of adult brain DMRs that correspond to vestigial enhancers, i.e. overlapping with open chromatin regions in the

embryonic, but not the adult brain (Supplementary Fig. 3.7G-J). Consistent with our speculation, in many cases, vestigial enhancers show stronger enrichment of disease heritability (Supplementary Fig. 3.7K-N). In particular, the enrichment of autism spectrum disorder genetic risk in L4 PLCH1 and L6 TLE4 cells can only be identified in vestigial enhancers (Supplementary Fig. 3.7M). In summary, we found that single cell-type DMRs integrate regulatory information during brain development and in the adult brain and can be used to predict cell types involved in neuropsychiatric disorders. However, our predictions should be considered in light of important limitations. Statistical approaches such as LD score regression partitioned heritability[135] and RolyPoly[139] have been validated for the prioritization of trait-associated tissues, but their application to fine-grained cell types remain preliminary. In addition, experimental validation of the association between disease and cell types is challenging due to the difficulty in accurately recapitulating disease phenotypes and modeling diverse cell populations in cell cultures[140]. Together, the investigation of disease-associated cell types is still in its infancy and will require further methodological breakthroughs in cell culture and gene editing approaches.

3.11 Discussion

Epigenomic studies often incorporate multiple molecular profiles from the same sample to explore possible correlations between gene regulatory elements and expression. The need for multi-omic comparison poses a challenge for single-cell analysis, since most existing single-cell techniques terminally consume the cell, precluding multi-dimensional analysis. To address this challenge, we have developed a single-nucleus multi-omic assay snmCAT-seq to jointly profile the transcriptome, DNA methylome and chromatin accessibility and can be applied to either single cells or nuclei isolated from frozen human tissues. snmCAT-seq requires no physical separation of DNA and RNA and is designed to be a “single-tube” reaction for steps before bisulfite

conversion to minimize material loss. snmCAT-seq is fully compatible with high-throughput single-cell methylome techniques such as snmC-seq2[28] and can be readily scaled to analyze thousands of cells/nuclei.

The continuous development of multi-omic profiling techniques such as scNMT-seq[20] and snmCAT-seq, and several methods for joint RNA and chromatin accessibility profiling sci-CAR[141], SNARE-seq[142], Paired-seq[143] and SHARE-seq[144] provide the opportunity to classify cell types with multiple molecular signatures. Our study developed computational methods to cross-validate clustering-based cell-type classifications using multi-modal data. Through cross-validation between matched single-cell mC and RNA profiles, we found that between 20-50 human cortical cell types can be identified from our moderate size snmCAT-seq dataset (4,358 cells) with sound cluster robustness. This is consistent with the number of human frontal cortex cell types we reported in our previous (21 major types[3]) and current (20 major types and 63 subtypes) studies. Determining the optimal number of clusters for any dataset should consider statistical robustness, the need of the biological questions, and the cell-type resolution of companion data modalities. Practical factors could also impact the choice of clusters, such as the requirement of certain minimum coverage for the pseudobulk methylome for DMR analysis. Together, although statistical robustness is essential for any cell type classification using clustering methods, the optimal number of clusters is to some extent an investigator-driven choice depending on the context of the study. Using snmCAT-seq as a "ground-truth", we determined that computational multi-modal data fusion tools perform well at the major cell-type level but show variable accuracy for the fusion of fine-grain subtypes. The computational strategies developed in this study can be applied to other types of multi-omic profiling including methods involving physiological measurement such as Patch-seq[145,146].

Epigenomic studies at both bulk and single-cell levels have established both mC and open chromatin as reliable markers for regulatory elements[31]. However, the difference between the information provided by the two epigenomic marks has been less clear in the context of normal development and diseases. Our study found that DMRs contain disease-related regulatory information of both adult and embryonic tissues, with vestigial enhancers [7] as a possible mechanism that informs developmental gene regulation. The strong enrichment of genetic risks for neuropsychiatric disorders in vestigial enhancers enabled the prediction of cellular lineages associated with diseases using DMRs for partitioned heritability analyses and identified more diverse disease-associated brain cell populations than similar analyses using open chromatin regions. The abundance of developmental information in DNA methylome suggests the possibility to study developmental processes and gene regulation in cell lineages using methylome profiling of adult tissues, especially given the practical and ethical challenges for obtaining primary human tissues from developmental stages.

3.12 Limitations of the Study

The transcriptome assay of snmCAT-seq was based on the Smart-seq2 method[111] published more than 7 years ago. The incorporation of further optimized single-cell RNA approaches such as Smart-seq3[147] may enhance the performance of transcriptome profiling for snmCAT-seq. Similar to other bisulfite sequencing-based approaches, the relatively high cost of resequencing the bisulfite-converted genome limits the number of cells that can be profiled with snmCAT-seq. However, with the continuous reduction of sequencing cost, it will become feasible to routinely profile hundreds of thousands of snmCAT-seq libraries. Although the current plate-based library preparation method of snmCAT-seq has a maximum throughput of approximately 10,000 cells per week[28], the molecular partitioning design of snmCAT-seq is a simple “single-tube” reaction and

can be readily combined with combinatorial indexing based methylome preparation methods such as sci-MET[14]. In snmCAT-seq, the ratio between transcriptome and methylome reads is determined by the absolute quantity of mRNA and pre-mRNA in a single nucleus, since the amount of genomic DNA is a constant ~5pg per nucleus in diploid human cells. Therefore, the application of snmCAT-seq to a new tissue type requires testing of the number of cycles of cDNA amplification necessary to achieve an optimized representation of transcriptome reads in the sequencing library.

Although we have successfully incorporated NOMe-seq in snmCAT-seq for the profiling of chromatin accessibility, the single-nucleus NOMe-seq profiles have moderate signal-to-noise ratio and may be better suited for identifying open chromatin regions using pseudo bulk profiles, rather than for the *de novo* clustering of single-cell using chromatin accessibility information (Supplementary Fig. 3.2H, K). This could be due to our use of frozen tissue providing an intrinsically lower signal-to-noise ratio than experiments using freshly harvested cells[17]. Nevertheless, we have demonstrated that following the robust identification of cell types using the methylome and transcriptome components of snmCAT-seq, the quantitative analysis of pseudo bulk NOMe-seq profiles has generated insights about the modulation of TF binding by methylcytosines (Fig. 3.6F-I), suggesting the unique applications of single-cell multi-modal datasets.

3.13 Methods

3.13.1 Cell cultures

HEK293T cells were cultured in DMEM with 15% FBS and 1% Penicillin-Streptomycin and dissociated with 1X TrypLE. H1 human ESCs (WA01, WiCell Research Institute) were

maintained in a feeder-free mTesR1 medium (Stemcell Technologies). hESCs (passage 26) were dispersed with 1U/ml Dispase and collected for single-cell sorting or nuclei isolation. For the sorting of single H1 and HEK293T cells, equal amounts of H1 and HEK293T cells were mixed and stained with anti-TRA-1-60 (Biolegend, Cat#330610) antibody.

3.13.2 Human brain tissues

Postmortem human brain biospecimens GUID: NDARKD326LNK and NDARKJ183CYT were obtained from NIH NeuroBioBank at the University of Miami Brain Endowment Bank. Postmortem human brain biospecimens UMB4540, UMB5577 and UMB5580 were obtained from NIH NeuroBioBank at the University of Maryland Brain and Tissue Bank. Published snmC-seq was generated from frontal cortex (medial frontal gyrus) tissue obtained from a 25-year-old Caucasian male (UMB4540, labeled as M_25yr_1 in this study) with a postmortem interval (PMI) = 23 h. The snATAC-seq dataset was generated from specimen UMB4540. Additional snmC-seq data was generated in frontal cortex (superior frontal gyrus, Brodmann area 10) tissues obtained from a 58-year-old Caucasian male (GUID: NDARKD326LNK, labeled as M_58yr in this study) with a postmortem interval (PMI) = 23.4 h. snmC-seq2 data was generated from frontal cortex (Brodmann area 10) tissue from a 25-year-old Caucasian male (GUID: NDARKJ183CYT, labeled as M_25yr_2 in this study) with a PMI = 20.8 h. snmCAT-seq and sn-m3C-seq data were generated from a 21-year-old Caucasian male (UMB5577, labeled as M_21yr in this study) with a PMI = 19h, and a 29-year-old Caucasian male (UMB5580, labeled as M_29yr in this study) with a PMI = 8h. The samples were taken from unaffected control subjects who died from accidental causes. The snRNA-seq dataset was generated from postmortem brain specimen H18.30.002 from the Allen Institute for Brain Science. The frontal cortex (BA44-45, 46) from this donor was used for

the generation of single nucleus RNA-seq data. The donor was a 50 year old male with a PMI = 12 h.

3.13.3 Nuclei isolation from cultured cells for snmCAT-seq

Cell pellets containing 1 million cells were resuspended in 600 µl NIBT [250 mM Sucrose, 10 mM Tris-Cl pH=8, 25 mM KCl, 5mM $MgCl_2$, 0.1% Triton X-100, 1mM DTT, 1:100 Proteinase inhibitor (Sigma-Aldrich P8340), 1:1000 SUPERaseIn RNase Inhibitor (ThermoFisher Scientific AM2694), 1:1000 RNaseOUT RNase Inhibitor (ThermoFisher Scientific 10777019)]. The lysate was transferred to a pre-chilled 2 ml Dounce homogenizer (Sigma-Aldrich D8938) and Dounced using loose and tight pestles for 20 times each. The lysate was then mixed with 400 µl of 50% Iodixanol (Sigma-Aldrich D1556) and gently pipetted on top of 500 µl 25% Iodixanol cushion. Nuclei were pelleted by centrifugation at 10,000 x g at 4°C for 20 min using a swing rotor. The pellet was resuspended in 2 ml of DPBS supplemented with 1:1000 SUPERaseIn RNase Inhibitor and 1:1000 RNaseOUT RNase Inhibitor. Hoechst 33342 was added to the sample to a final concentration of 1.25 nM and incubated on ice for 5 min for nuclei staining. Nuclei were pelleted by 1,000 x g at 4°C for 10 min and resuspended in 1 ml of DPBS supplemented with RNase inhibitors.

3.13.4 Nuclei isolation from human brain tissues and GpC methyltransferase treatment for snmCAT-seq

Brain tissue samples were ground in liquid nitrogen with cold mortar and pestle, and then aliquoted and store at -80°C. Approximately 100mg of ground tissue was resuspended in 3 ml NIBT (250 mM Sucrose, 10 mM Tris-Cl pH=8, 25 mM KCl, 5mM $MgCl_2$, 0.2% IGEPAL CA-630, 1mM DTT, 1:100 Proteinase inhibitor (Sigma-Aldrich P8340), 1:1000 SUPERaseIn RNase Inhibitor (ThermoFisher Scientific AM2694), 1:1000 RNaseOUT RNase Inhibitor (ThermoFisher

Scientific 10777019)). The lysate was transferred to a pre-chilled 7 ml Dounce homogenizer (Sigma-Aldrich D9063) and Dounced using loose and tight pestles for 40 times each. The lysate was then mixed with 2 ml of 50% Iodixanol (Sigma-Aldrich D1556) to generate a nuclei suspension with 20% Iodixanol. Gently pipet 1 ml of the nuclei suspension on top of 500 µl 25% Iodixanol cushion in each of the 5 freshly prepared 2ml microcentrifuge tubes. Nuclei were pelleted by centrifugation at 10,000 x g at 4°C for 20 min using a swing rotor. The pellet was resuspended in 1ml of DPBS supplemented with 1:1000 SUPERaseIn RNase Inhibitor and 1:1000 RNaseOUT RNase Inhibitor. A 10 µl aliquot of the suspension was taken for nuclei counting using a Biorad TC20 Automated Cell Counter. One million nuclei aliquots were pelleted by 1,000 x g at 4°C for 10 min and resuspended in 200 µl of GpC methyltransferase M.CviPI (NEB M0227L) reaction containing 1X GC Reaction Buffer, 0.32 nM S-Adenoslylmethionime, 80U 4U/µl M.CviPI, 1:100 SUPERaseIn RNase Inhibitor and 1:100 RNaseOUT RNase Inhibitor and incubated at 37°C for 8 min. The reaction was stopped by adding 800 µl of ice-cold DPBS with 1:1000 RNase inhibitors and mixing. Hoechst 33342 was added to the sample to a final concentration of 1.25 nM and incubated on ice for 5 min for nuclei staining. Nuclei were pelleted by 1,000 x g at 4°C for 10 min, resuspended in 900 µl of DPBS supplemented with 1:1000 RNase inhibitors and 100 µl of 50mg/ml Ultrapure™ BSA (Ambion AM2618) and incubated on ice for 5 min for blocking. Neuronal nuclei were labeled by adding 1 µl of AlexaFluor488-conjugated anti-NeuN antibody (clone A60, MilliporeSigma MAB377XMI) for 20 min.

3.13.5 Reverse transcription for snmCAT-seq

Single cells or single nuclei were sorted into 384-well PCR plates (ThermoFisher 4483285) containing 1 µl snmCAT-seq reverse transcription reaction per well. The snmCAT-seq reverse transcription reaction contained 1X Superscript II First-Strand Buffer, 5mM DTT, 0.1% Triton X-

100, 2.5 mM $MgCl_2$, 500 μM each of 5'-methyl-dCTP (NEB N0356S), dATP, dTTP and dGTP, 1.2 μM dT30VN_4 oligo-dT primer (5'-AAGCAGUGGUAUCAACGCAGAGUACUTTTTTUTTTTTUTTTTTUTTTTTUTTTTTVN-3' was used the cultured cell snmCAT-seq experiments; 5'-/5SpC3/AAGCAGUGGUAUCAACGCAGAGUACUTTTTTUTTTTTUTTTTTUTTTTTUTTTTTVN-3' was used for human brain snmCAT-seq experiments), 2.4 μM TSO_3 template switching oligo (5'-/5SpC3/AAGCAGUGGUAUCAACGCAGAGUGAAUrGrG+G-3'), 1U RNaseOUT RNase inhibitor, 0.5 U SUPERaseIn RNase inhibitor, 10U Superscript II Reverse Transcriptase (ThermoFisher 18064-071). For snmCAT-seq performed with nuclei samples, the reaction further included 2 μM N6_2 random primer (5'-/5SpC3/AAGCAGUGGUAUCAACGCAGAGUACNNNNNN-3'). After sorting, the PCR plates were vortexed and centrifuged at 2000 x g. The plates were placed in a thermocycler and incubated using the following program: 25°C for 5 min, 42°C for 90min, 70°C 15min followed by 4°C.

3.13.6 cDNA amplification for snmCAT-seq

3 μl of cDNA amplification mix was added into each snmCAT-seq reverse transcription reaction. Each cDNA amplification reaction containing 1X KAPA 2G Buffer A, 600 nM ISPCR23_2 PCR primer (5'-/5SpC3/AAGCAGUGGUAUCAACGCAGAGU-3'), 0.08U KAPA2G Robust HotStart DNA Polymerase (5 U/μL, Roche KK5517). PCR reactions were performed using a thermocycler with the following conditions: 95°C 3min -> [95°C 15 sec -> 60°C 30 sec -> 72°C 2min] -> 72°C 5min -> 4°C. The cycling steps were repeated for 12 cycles for snmCAT-seq using H1 or HEK293 whole cells, 15 cycles for snmCAT-seq using H1 or HEK293 nuclei and 14 cycles for snmCAT-seq using human brain tissue nuclei.

3.13.7 Digestion of unincorporated DNA oligos for snmCAT-seq

For snmCAT-seq using H1 and HEK293 cells, 1 µl uracil cleavage mix was added into cDNA amplification reaction. Each 1 µl uracil cleavage mix contains 0.25 µl Uracil DNA Glycosylase (Enzymatics G5010) and 0.25 µl Endonuclease VIII (Enzymatics Y9080) and 0.5 µl Elution Buffer (Qiagen 19086). Unincorporated DNA oligos were digested at 37°C for 30 min using a thermocycler. We have found that Endonuclease VIII is dispensable for the digestion of unincorporated DNA oligos since the alkaline condition during the desulfonation step of bisulfite conversion can effectively cleave abasic sites created by Uracil DNA Glycosylase [148]. Therefore for snmCAT-seq using human brain tissues, each cDNA amplification reaction was treated with 1µl uracil cleavage mix containing 0.5 µl Uracil DNA Glycosylase (Enzymatics G5010-1140) and 0.5 µl Elution Buffer (Qiagen 19086).

3.13.8 Bisulfite conversion and library preparation

Detailed methods for bisulfite conversion and library preparation are previously described for snmC-seq2 [3,28]. The following modifications were made to accommodate the increased reaction volume of snmCAT-seq: Following the digestion of unused DNA oligos, 25 µl instead of 15 µl of CT conversion reagent was added to each well of a 384-well plate. 90 µl instead of 80 µl M-binding buffer was added to each well of 384-well DNA binding plates. snmCAT-seq libraries performed using whole H1 or HEK293 cells were generated using the snmC-seq method as described in Luo et al., 2017 [3]. The rest of the snmCAT-seq libraries were generated using the snmC-seq2 method as described in Luo et al., 2018 [28]. The snmCAT-seq libraries generated from H1 and HEK293 cells were sequenced using an Illumina HiSeq 4000 instrument with 150 bp paired-end reads. The snmCAT-seq libraries generated from human brain specimens were sequenced using an Illumina Novaseq 6000 instrument with S4 flowcells and 150 bp paired-end

mode.

3.13.9 The mapping pipeline for snmC-seq, snmC-seq2, and snmCAT-seq

We implemented a versatile mapping pipeline (cemba-data.rtfd.io) for all the methylome based technologies developed by our group [3,28]. The main steps of this pipeline include: 1) Demultiplexing FASTQ files into single-cell; 2) Reads level QC; 3) Mapping; 4) BAM file processing and QC; 5) final molecular profile generation.

For snmC-seq and snmC-seq2, the details of the five steps are described previously [3,28]. For snmCAT-seq, steps 1 and 2 are identical as snmC-seq2, steps 3 to 5 are split into "a" for methylome and "b" for transcriptome as following:

Step 3a (methylome). To map methylome reads, reads from step 2 were mapped onto the human hg19 genome using Bismark [149] with the same setting as snmC-seq2.

Step 3b (transcriptome). To map transcriptome reads, reads from step 2 were mapped to GENCODE human v28 indexed hg19 genome using STAR 2.7.2b [150] with the following parameters: *--alignEndsType EndToEnd --outSAMstrandField intronMotif --outSAMtype BAM Unsorted --outSAMunmapped Within --outSAMattributes NH HI AS NM MD --sjdbOverhang 100 --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --outSAMattrRGline ID:4 PL:Illumina*.

Step 4a (methylome). PCR duplicates were removed from mapped reads using Picard MarkDuplicates. The non-redundant reads were then filtered by MAPQ > 10. To select genomic reads from the filtered BAM, we used the "XM-tag" generated by Bismark to calculate reads methylation level and keep reads with mCH ratio < 0.5 and the number of cytosines ≥ 3.

Step 4b (transcriptome), the STAR mapped reads were first filtered by MAPQ > 10. To select RNA reads from the filtered BAM, we used the "MD" tag to calculate reads methylation level and keep reads with mCH ratio > 0.9 and the number of cytosines ≥ 3. The stringency of read partitioning was determined by applying the criteria for identifying snmCAT-seq transcriptome reads to snmC-seq2 data (SRR6911760, SRR6911772, SRR6911776) [28], which contains no transcriptomic reads. Similarly, the criteria for identifying snmCAT-seq methylome reads were applied to Smart-seq data (SRR944317, SRR944318, SRR944319, SRR944320) [111], which contains no methylome reads.

Step 5a (methylome), Tab-delimited (ALLC) files containing methylation level for every cytosine position was generated using methylpy *call_methylated_sites* function [41] on the BAM file from the step 4a. For snmCAT-seq, an additional base was added before the cytosine in the context column of the ALLC file using the parameter "--num_upstr_bases 1", to distinguish GpC sites from HpC sites for the NOMe-seq modality.

Step 5b (transcriptome), BAM file from step 4b were counted across gene annotations using featureCount 1.6.4 [151] with the default parameters. Gene expression was quantified using either only exonic reads with "-t exon" or both exonic and intronic reads with "-t gene".

3.13.10 Methylome feature generation

After allc files were generated, the methylcytosine (*mc*) and total cytosine basecalls (*cov*) were summed up for each 100kb bin across the hg19 genome. For snmC-seq and snmC-seq2, cytosine and methylcytosine basecalls in CH (H=A, T, C) and CG context were counted separately. For snmCAT-seq, the HCH context was counted for CH methylation and HCG is counted for CG methylation. The GCY (Y=T, C) context was counted as the chromatin accessibility signal (NOMe-seq in snmCAT-seq) and the HCY context was counted as the endogenous mCH

background. In addition to the 100kb feature set, we also counted gene body methylation levels using gene annotation from GENCODE v28. The 100kb feature set was used in methylation-based clustering analysis and data integration; the gene body feature set was used in methyl-marker identification, cluster annotation and data fusion between methylome and transcriptome.

3.13.11 Preprocessing of snmC-seq and snmC-seq2 data for clustering analyses

Cell filtering. We filtered the cells based on these main mapping metrics: 1) mCCC rate < 0.03. mCCC rate reliably estimates the upper bound of bisulfite non-conversion rate [3], 2) overall mCG rate > 0.5, 3) overall mCH rate < 0.2, 4) total final reads > 500,000, 5) Bismark mapping rate > 0.5. Other metrics such as genome coverage, PCR duplicates rate, index ratio were also generated and evaluated during filtering. However, after removing outliers with the main metrics 1-5, few additional outliers can be found.

Feature filtering. 100kb genomic bin features were filtered by removing bins with mean total cytosine base calls < 300 or > 3000. Regions that overlap with the ENCODE blacklist [73] were also removed from further analysis.

Computation and normalization of the methylation rate. For CG and CH methylation, the computation of methylation rate from the methylcytosine and total cytosine matrices contains two steps: 1) prior estimation for the beta-binomial distribution and 2) posterior rate calculation and normalization per cell.

Step 1, for each cell we calculated the sample mean, m, and variance, v, of the raw mc rate *(mc / cov)* for each sequence context (CG, CH). The shape parameters (α, β) of the beta distribution were then estimated using the method of moments:

$$\alpha = m(m(1-m)/v - 1)$$

$$\beta = (1-m)(m(1-m)/v - 1)$$

This approach used different priors for different methylation types for each cell, and used weaker prior to cells with more information (higher raw variance).

Step 2, We then calculated the posterior: $\widehat{mc} = \frac{\alpha+mc}{\alpha+\beta+cov}$., We normalized this rate by the cell's global mean methylation, $m = \alpha/(\alpha+\beta)$. Thus, all the posterior $\widehat{mc}$ with 0 *cov* will be constant 1 after normalization. The resulting normalized *mc* rate matrix contains no NA (not available) value and features with less *cov* tend to have a mean value close to 1.

Selection of highly variable features. Highly variable methylation features were selected based on a modified approach using the scanpy package *scanpy.pp.highly_variable_genes* function [44]. In brief, the *scanpy.pp.highly_variable_genes* function normalized the dispersion of a gene by scaling with the mean and standard deviation of the dispersions for genes falling into a given bin for mean expression of genes. In our modified approach, we reasoned that both the mean methylation level and the mean *cov* of a feature (100kb bin or gene) can impact *mc* rate dispersion. We grouped features that fall into a combined bin of mean and *cov*, and then normalized the dispersion within each mean-*cov* group. After dispersion normalization, we selected the top 3000 features based on normalized dispersion for clustering analysis.

Dimension reduction and combination of different mC types. For each selected feature, *mc* rates were scaled to unit variance and zero mean. PCA was then performed on the scaled *mc* rate matrix. The number of significant PCs was selected by inspecting the variance ratio of each PC using the elbow method. The CH and CG PCs were then concatenated together for further analysis in clustering and manifold learning.

3.13.12 Preprocessing of snmCAT-seq data for clustering analysis

Methylome preprocessing. The methylome modality preprocessing is similar to snmC-seq2 with one major modification: non-CG methylation is quantified using the HCH context; CG

methylation is quantified using the HCG context. Chromosome 100kb bin features with mean total cytosine base calls between 250 and 2500 were included in downstream analyses.

Transcriptome preprocessing. The whole gene RNA read count matrix is used for snmCAT-seq transcriptome analysis. Cells are filtered by the number of genes expressed > 200 and genes are filtered by the number of cells expressed > 10. The count matrix X is then normalized per cell and transformed by ln(X + 1). After log transformation, we use the *scanpy.pp.highly_variable_genes* to select the top 3000 genes based on normalized dispersion, using a process similar to the selection of highly variable methylation features. The selected feature matrix is scaled to unit variance and zero mean per feature followed by PCA calculation.

Chromatin accessibility (NOMe-seq) preprocessing. For clustering analysis, cytosine methylation in the GCY context (GmCY) is counted as the open chromatin signal from NOMe-seq. For each 5 kb bin, we modeled its GmCY basecall in a single cell using a binomial distribution Bi(*cov*, *global*), where *cov* represents the total GCY basecall of the bin in the cell, and *global* represents the global GmCY level of the cell. We then computed the probability of observing equal or greater GmCY basecall than observed as the survival function of the binomial distribution. The bins with this probability smaller than 0.05 were marked as 1, and otherwise 0, by which we generated a $\#cell \times \#bin$ binarized matrix as the open chromatin signals. Latent semantic analysis with log term frequency was used to compute the embedding. Specifically, we selected the bins that are open in > 10 cells, then computed the column sum of the matrix and kept only the bins with Z-scored column sum < 2. The filtered matrix A was row normalized to B by dividing the row sum, and $C_{ij} = log(B_{ij} + 1) \times log(1 + \frac{\#cells}{\sum_{i\prime=1}^{\#cells} A_{i\prime j}})$ was used for dimension reduction by singular value decomposition. We used the first 15 dimensions of the left singular vector matrix as the input of UMAP for visualization.

3.13.13 General strategies for clustering and manifold learning

Consensus clustering on concatenated PCs. We used a consensus clustering approach based on multiple Leiden-clustering [74] over K-Nearest Neighbor (KNN) graph to account for the randomness of the Leiden clustering algorithms. After selecting dominant PCs from PCA in all available modalities of different technologies (mCH, mCG for snmC-seq and snmC-seq2; mCH, mCG, RNA, NOMe-seq for snmCAT-seq, etc.), we concatenated the PCs together to construct KNN graph using scanpy.pp.neighbors. Given fixed resolution parameters, we repeated the Leiden clustering 200 times on the KNN graph with different random starts and combined these cluster assignments as a new feature matrix, where each single Leiden result is a feature. We then used the outlier-aware DBSCAN algorithm from the scikit-learn package to perform consensus clustering over the Leiden feature matrix using the hamming distance. Different epsilon parameters of DBSCAN are traversed to generate consensus cluster versions with the number of clusters that range from minimum to the maximum number of clusters observed in the 200x Leiden runs. Each version contains a few outliers that usually fall into three categories: 1. Cells located between two clusters that have gradient differences instead of clear borders, e.g. L2-3 IT to L4 IT; 2. Cells with a low number of reads that potentially lack information in important features to determine the exact cluster. 3. Cells with a high number of reads that are potential doublets. The number of type 1 and 2 outliers depends on the resolution parameter and is discussed in the choice of the resolution parameter section, the type 3 outliers are very rare after cell filtering. The final consensus cluster version is then determined by the supervised model evaluation.

Supervised model evaluation on the clustering assignment. For each consensus clustering version, we performed a Recursive Feature Elimination with Cross-Validation (RFECV) [75] process from the scikit-learn package to evaluate clustering reproducibility. We first removed

the outliers from this process, then we held out 10% of the cells as the final testing dataset. For the remaining 90% of the cells, we used tenfold cross-validation to train a multiclass prediction model using the input PCs as features and *sklearn.metrics.balanced_accuracy_score* [76] as an evaluation score. The multiclass prediction model is based on BalancedRandomForestClassifier from the imblearn package that accounts for imbalanced classification problems [77]. After training, we used the 10% testing dataset to test the model performance using the balanced_accuracy_score score. We kept the best model and corresponding clustering assignments as the final clustering version. Finally, we used this prediction model to predict outliers' cluster assignments, we rescued the outlier with prediction probability > 0.5, otherwise labeling them as outliers.

Choice of resolution parameter. Choosing the resolution parameter of the Leiden algorithm is critical for determining the final number of clusters. We selected the resolution parameter by three criteria: 1. The portion of outliers < 0.05 in the final consensus clustering version. 2. The final prediction model performance > 0.95. 3. The average cell per cluster ≥ 30, which controls the cluster size in order to reach the minimum coverage required for further epigenome analysis such as DMR calling. All three criteria prevent the over-splitting of clusters thus we selected the maximum resolution parameter under meeting the criteria using grid search in each specific clustering analysis below.

Cluster marker gene identification and cluster trimming. After clustering, we used a one-vs-rest strategy to calculate methylation (methyl-marker) and RNA (rna-marker, for snmCAT-seq only) marker genes for each cluster. We used all the protein-coding and long non-coding RNA genes with evidence level 1 or 2 from gencode v28. For the rna-marker, we used the *scanpy.tl.rank_genes_group* function with the Wilcoxon test and Benjamini-Hochberg multi-test correction, and filtered the resulting marker gene by adjusted P-value < 0.01 and log2(fold-

change) > 1, we also used AUROC score as a measure of marker gene's predictability of corresponding cluster, and filtered genes by AUROC > 0.8. For the methyl-marker, we used the normalized gene body mCH rate matrix to calculate markers for neuronal clusters and the normalized gene body mCG rate matrix for non-neuronal clusters, and we modified the original Wilcoxon test function to used a reverse score to select genes that have significant decrease (hypomethylation). Marker gene is chosen based on adjusted P-value < 0.01, delta methylation level change < -0.3 (hypo-methylation), AUROC > 0.8. The delta methylation level is calculated as the normalized methylation rate change between the cluster and the mean value of the rest clusters. For the ensemble methylome clustering, if a cluster with the number of methyl-markers < 10 is detected, the cluster with the minimum total marker genes are merged to the closest clusters based on cluster centroids euclidean distance in the PC space, then the marker identification process is repeated until all clusters found enough marker genes.

Manifold learning. The T-SNE and UMAP embedding are run on the PC matrix the same as the clustering input using the scanpy package.

3.13.14 Identification of open chromatin regions using snmCAT-seq GCY methylation profiles

Methylated GCY sites were identified using a Hidden Markov Model (HMM) method gNOMePeaks [117], with the methylation level of GCY sites modeled using binomial distribution. The accessibility state of each GCY site was modeled with a three-state HMM model with state 3 indicating accessible chromatin. To tune the HMM model for the different background mCH levels between neuronal and non-neuronal cell types, we skipped the expectation–maximization (EM) algorithm for estimating the p parameter of binomial distribution. Instead, for neuronal cell types, p parameters were specified as 0.1, 0.2 and 0.5 for states 1-3, respectively; for non-neuronal cell

types, p parameters were specified as 0.1, 0.25 and 0.4 for states 1-3, respectively. The density of methylated GCY sites across the genome was modeled using Poisson distribution by MACS2 [152] and regions with a significant enrichment of methylated GCY sites were identified with MACS2 *callpeak* with a p-value < 0.01. Peaks with q-value < 0.01 were selected for downstream analyses.

3.13.15 snATAC-seq data generation

Combinatorial barcoding single nucleus ATAC-seq was performed as described previously in Fang et al. [153]. Isolated brain 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 cOmplete™, 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. For tagmentation, 1 µL barcoded Tn5 transposomes [153] 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 was 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) was 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 Fluorescence-activated cell sorter (Sony), 40 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 the addition of 1 µL 0.2% SDS, samples were incubated at 55 °C for 7 min with shaking (500 rpm). 1 µL 12.5% Triton-X was added to each well to quench the SDS. Next, 12.5 µL NEBNext High-Fidelity 2× PCR Master Mix (NEB) were added and 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 (Beckman Coulter, 0.55x and 1.5x). Libraries were purified one more time with SPRI beads (Beckman Coulter, 1.5x). Libraries were quantified using a Qubit fluorometer (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 + 37 + 50 (Read1 + Index1 + Index2 + Read2) [30].

3.13.16 snATAC-seq data processing

Using a custom python script, we first demulticomplexed FASTQ files by integrating the cell barcode (concatenate reads pair in I1.fastq and I2.fastq) into the read name (R1.fastq and R2 fastq) in the following format: "@"+"barcode"+":"+"original_read_name". Demulticomplexed reads were aligned to the corresponding reference genome (hg19) using bwa (0.7.13-r1126) [154] in pair-end mode with default parameter settings. Alignments were then sorted based on the read name using samtools (v1.9) [155]. Pair-end reads were converted into fragments and only those that are 1) properly paired (according to SMA flag value); 2) uniquely mapped (MAPQ > 30); 3) with length less than 1000bp were kept. Since fragments were sorted by barcode (integrated into the read name), fragments belonging to the same cell (or barcode) were automatically grouped together which allowed for removing PCR duplicates for each cell separately. Using the remaining

fragments, a snap-format (Single-Nucleus Accessibility Profiles) file was generated. snap file is 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 metadata. BM session contains cell-by-bin matrices of different resolutions (or bin sizes). PM session contains a cell-by-peak count matrix. PM session contains a cell-by-gene count matrix. FM session contains all usable fragments for each cell. Fragments are indexed for fast search. A detailed documentation of snap file can be found here (https://docs.google.com/document/d/1AGyn_WJTr0A1SKcfrEgum-jvAJjWd84jZ6Dwbwisencing Fragments (>1,000); 2) Mapping Ratio (>0.8); 3) Properly Paired Ratio (>0.9); 4) Duplicate Ratio (<0.5); 5) Mitochondrial Ratio (<0.1). [153].

3.13.17 Clustering analysis of snATAC-seq data

We used the snapATAC package for the clustering analysis of snATAC-seq data, the detail steps were described in [153]. Briefly, we used the binarized cell-by-bin matrix of the whole genome 5kb non-overlapping bins as input (1 means open, 0 means close or missing data). We first determined the coverage of each bin and converted the coverage distribution to log-normal distribution and converted the bin coverage to z-score. Bins with extremely high (zscore > 1.5) or low coverage (zscore < -1.5), or overlap with ENCODE blacklist [73] are removed. We then converted the cell-by-bin matrix into a cell-by-cell similarity matrix by calculating the Jaccard index between cells. To normalize the cell coverage impact on the Jaccard index, we used the observed over expected (OVE) method from snapATAC, which calculates the residual of the linear regression model between the expected Jaccard matrix given cell coverage and the overserved matrix. We then performed PCA on a standardized residual matrix and used the top 25 PCs for

Leiden clustering (resolution = 1) and UMAP visualization.

3.13.18 Open chromatin peak calling using snATAC-seq data

Open chromatin peaks were identified using snATAC-seq reads combined for each cell type using MACS *callpeak* with the following parameters -f BED --nomodel --shift 37 --ext 73 --pvalue 1e-2. Peaks with q-value < 0.01 were further selected for downstream analyses.

3.13.19 snRNA-seq data generation

Nuclei were isolated from human postmortem brain tissues and sorted based on NeuN fluorescence as previously described [84]. Each sample contained approximately 80% NeuN-positive and 20% NeuN-negative nuclei. snRNA-seq data was generated using 10x Genomics v3 single-cell chemistry per the manufacturer's protocol. RNA-seq reads were aligned with Cell Ranger v3 using the human GRCh38.p2 reference genome, and intronic and exonic mapped reads were included in gene expression quantification.

3.13.20 snRNA-seq clustering and annotation

Nuclei were included in downstream analysis if they passed the following QC thresholds: > 500 genes detected (UMI > 0) in non-neuronal nuclei or > 1000 genes detected (UMI > 0) in neuronal nuclei; and doublet score < 0.3. Cells were grouped into transcriptomic cell types using the iterative clustering procedure described in [35]. Briefly, genes from the mitochondrial and sex chromosomes were excluded, and expression was normalized to UMI per million and log2-transformed. Nuclei were clustered using the following steps: high variance gene selection, dimensionality reduction, dimension filtering, Jaccard–Louvain or hierarchical (Ward) clustering, and cluster merging. Differential gene expression (DGE) was computed for every pair of clusters, and pairs that did not meet the DGE criteria were merged. Differentially expressed genes were defined using two criteria: 1) significant differential expression (> 2-fold; Benjamini-Hochberg

false discovery rate < 0.01) using the R package limma and 2) binary expression (CPM > 1 in more than half of the cells in one cluster and < 30% of this proportion in the other cluster). We define the deScore as the sum of the −log10(false discovery rate) of all differentially expressed genes (each gene contributes to no more than 20), and pairs of clusters with deScore < 150 were merged. This process was repeated within each resulting cluster until no more child clusters met DGE or cluster size criteria (minimum of 10 cells). The entire clustering procedure was repeated 100 times using 80% of all cells sampled at random, and the frequency with which nuclei co-cluster was used to generate a final set of clusters, again subject to differential gene expression and cluster size termination criteria. Clusters were identified as outliers if more than 40% of nuclei co-expressed markers of inhibitory (GAD1, GAD2) and excitatory (SLC17A7) neurons or were NeuN+ but did not express the pan-neuronal marker SNAP25. Median values of total UMI counts and gene counts were calculated for each cluster and used to compute the median and inter-quartile range (IQR) of all cluster medians. Clusters were also identified as outliers if the cluster median QC metrics deviated by more than three times the IQRs from the median of all clusters. In total, 23,379 nuclei passed QC criteria and were split into three broad classes of cells (13,997 excitatory neurons, 7,094 inhibitory neurons, and 1,914 non-neuronal cells) based on NeuN staining and cell class marker-gene expression. A final merge step required at least 4 marker genes to be more highly expressed in each pair of clusters. The clustering pipeline is implemented in an R package publicly available at github (https://github.com/AllenInstitute/scrattch.hicat). The clustering method is provided by the run_consensus_clust function.

3.14 FIGURE-SPECIFIC METHODS

3.14.1 Cell line dataset analysis (Supplementary Fig. 3.1)

Clustering. For snmCAT-seq dataset generated from the whole cell and nucleus of H1 and HEK293 cells, PCA was used for the dimension reduction of the mCG and RNA matrices. Since only two cell types (H1 and HEK293T) need to be separated, only the first 5 PCs from each matrix were selected to construct K-Nearest Neighbor (KNN) graphs (K=25). On each KNN graph for mCG and RNA, Leiden clustering (r=0.5) is used to determine the two clusters and tSNE was used to visualize the PCs. Clusters were annotated by examining the genome-wide methylation levels and marker gene expression. Data acquired from single cells or nuclei were then merged for each cluster for comparisons with bulk methylome and transcriptome data.

Comparison to bulk H1 and HEK293 Methylome. The bulk HEK293 cell whole-genome bisulfite sequencing (WGBS-seq) data were downloaded from Libertini et. al. (GSM1254259) [156]. The bulk WGBS-seq data of the H1 cell was downloaded from Schultz et al (GSE16256) [41]. Methylpy was used to call CG-DMRs between these two cell lines[41]. DMRs were filtered by DMS (differentially methylated sites) ≥ 5 and methylation level difference ≥ 0.6.

Bulk H1 and HEK293 RNA Data Analysis. The bulk HEK293 cell RNA-seq data was downloaded from Aktas et. al. (GSE85161) [157], the bulk H1 cell RNA-seq data was downloaded from encodeproject.org (ENCLB271KFE, generated by Roadmap Epigenome). Gene count tables and bigwig tracks were generated using human GENCODE v19 gene annotation.

3.14.2 snmCAT-seq baseline clustering (Fig. 3.1)

To perform clustering analysis on the human frontal cortex snmCAT-seq dataset only, we first preprocessed three modalities separately as described in the preprocessing section above. We then concatenate all the dominant PCs together to run the consensus clustering identification

(resolution = 1). We annotated the clusters based on marker genes reported in the previous studies [3,84]. We also calculated the UMAP coordinates based on concatenated PCs (Fig. 3.1D) and PCs from every single modality separately (Supplementary Fig. 3.2I-K).

3.14.3 Methylome ensemble clustering (Fig. 3.4)

To generate an ensemble cell type taxonomy for the human frontal cortex, we combine four methylome-based technologies (Fig. 3.4A, snmCAT-seq, snmC-seq, snmC-seq2, sn-m3C-seq) in this study. Due to the high cell-type diversity, we performed a two-level iterative clustering analysis.

Level 1 clustering to identify major cell types. We first preprocessed the methylation matrix as described above for each technology separately to obtain the corresponding highly variable feature matrix. We then used Scanorama to integrate all cells using the union of highly variable features from all technologies, with K=25 and default values for other parameters. After the integration, we performed PCA on the integrated matrix and used the dominant PCs for the subsequent consensus clustering analysis (resolution = 0.5) as described above. We also calculated UMAP coordinates using the ensemble PCs (Fig. 3.4C).

Level 2 clustering to identify subtypes for each major cell type. After level 1 clustering, we selected cells from each major cell type and repeated all the steps from highly variable feature selection to final clustering (K=20, resolution = 0.8) including Scanorama integration. The highly variable features selected in this step are more specific to the intracluster diversity of each major type, which helps to better separate the subtype. The subcluster UMAP coordinates are calculated from PCs in each subtype analysis (Fig. 3.4C insets).

3.14.4 Cross-validation of cell clusters (Fig. 3.2)

The analysis starts with 2 cell-by-gene data matrices: one for gene-body non-CG DNA

methylation (mCH) and the other for RNA expression. We first filter out low-quality cells and low-coverage genes. After removing glia and outliers in the snmCAT-seq dataset, we get 3,898 high-quality neuronal cells. By selecting genes expressed in >1% of cells and with >20 cytosines coverage at gene body in >95% of cells, we get 13,637 sufficiently covered genes. Then we normalize the mCH matrix by dividing the raw mCH level by the global mean mCH level of each cell; and we normalize the RNA matrix by ($\log_{10}$(TPM+1)).

The goal of cluster cross-validation is to cluster cells with one part of the features, and to validate clustering results with the other part of features. We first generate clusterings with different granularity, ranging from coarse to very fine, using DNA methylation features. Clusterings are generated by the Leiden method applied to the top 20 principal components with different settings of the resolution parameter controlling granularity. Following clustering, we randomly split cells into training and test sets. Using the training set, we estimate the cluster centroids of RNA expression. Using the test set, we calculated the mean squared error between the RNA expression profile of individual cells and that of cluster centroids. This procedure can be reversed by clustering with RNA features and evaluation with DNA methylation features.

To summarize the results, we plotted a curve of the number of clusters versus the mean squared error. To ensure robustness, clustering is repeated with five different random seeds, with each of the 5 clusters followed by 5 repetitions of 5-fold cross-validation on different random splits of training and test sets.

3.14.5 AIC and BIC metrics in the cluster cross-validation analysis (Fig. 3.2)

Akaike information criterion (AIC) and Bayesian information criterion (BIC) are metrics to estimate (in-sample) prediction error without a test set. The general definition of the two metrics are as follows,

$$AIC = -2 \cdot loglik + 2d$$

$$BIC = -2 \cdot loglik + (log\ N)\ d$$

where $loglik$is the log-likelihood of the model trained on a specific data set, d is the model dimension, $N$is the sample size. For both metrics, the first term evaluates the quality of fitting, whereas the second term penalizes model complexity.

In our case, we assume gene features of a single cell follows a Gaussian distribution around its cluster centroid:

$$y_{cell} = f(x)\ + \epsilon = y_{centroid} + \epsilon\ with\ \epsilon \sim N(0, \sigma I)$$

and with σbeing the standard deviation in the gaussian distribution that is the same across all dimensions (genes). Combining the model with the definitions of AIC and BIC, we get

$$AIC \sim \frac{1}{N}(y_{cell} - y_{centroid})^2\ +\ 2d \cdot \frac{\sigma^2}{N}$$

$$BIC \sim \frac{1}{N}(y_{cell} - y_{centroid})^2\ +\ (log\ N)\ d \cdot \frac{\sigma^2}{N}$$

where the first term is the mean squared error of the model fit, i.e., the training error, $N$is the number of cells, d is the number of cell clusters, and σ^2 is the variance of the dataset (assuming all cells are from the same cell clusters).

We applied this to 3,898 neuronal cells from the snmCAT-seq dataset. As for gene features, we include genes that have at least 1 RNA count in >1% cells, and with at least 20 methylation coverage in 95% of cells. This leaves 13,651 genes with both DNA methylation and RNA features. The DNA methylation features are calculated as the gene body non-CG methylation level (mHCH) normalized by the global mCHC level of each cell. The RNA features are $\log_{10}$(CPM+1) normalized. Using Leiden clustering with different resolutions, we generated clusters with different granularities. As a result, we report AIC, BIC, train and test error as functions of the

number of clusters. Errorbars are estimated from running the same settings repeatedly: [clustering with 10 different random seeds] x [10-time, 3-fold cross validations].

3.14.6 Quantification of over-splitting and under-splitting of cell clusters (Fig. 3.2)

Clustering of cell types requires a balance between over-splitting and under-splitting; this is the perennial tension between so-called lumpers and splitters as described by Darwin [158]. Over-splitting occurs when the noise in the data, for example due to random sampling of RNA or DNA molecules, drives the separation of cells which are not distinct. Under-splitting occurs when coarse-grained clusters fail to capture a meaningful biological distinction among subpopulations. The previous section described a cross-validation method to objectively pick a good clustering granularity for a given dataset. Here, we extend this to provide more detailed metrics of the degree of potential over- or under-splitting for particular cell clusters.

Our approach proceeds from the assumption that an ideal cluster should satisfy two requirements. First, all the cells within a cluster should be similar, with no clear discrete subdivisions that would indicate under-splitting. Second, the cells in one cluster should not resemble too closely the cells in any other cluster, which would indicate over-splitting. Unfortunately, no general methods for quantifying over- and under-splitting are available [159]. Taking advantage of the multimodal (RNA + DNA methylation) data, we defined metrics for over-splitting (S_{over}) and under-splitting (S_{under}), based on cross-validation analysis of the two data modalities. We have also added a supplementary tutorial (https://github.com/FangmingXie/mctseq_over_under_splitting/blob/master/over-under-splitting-analysis.ipynb) of the over- and under-splitting analysis to allow users to reproduce our results.

Cross-modality k-partner graph

First, we treat the two data modalities (mC and RNA) as independent measurements, as if

they came from separate DNA methylation and transcriptome assays performed on independent groups of cells. We embed cells from the two modalities into the same low-dimensional space using canonical correlation analysis [119]:

$$X \quad Y^T \approx USV^T,$$

where X and Y are cell-by-gene feature matrices for mC and RNA, respectively. For mC, the gene features are normalized mCH levels at the gene bodies; For RNA, the gene features are normalized RNA expression levels ($\log_{10}$(TPM+1)). U and V are cell-by-component matrices (number of components = 20). Mathematically this procedure is equivalent to a singular value decomposition of XY^T, where U and V are orthogonal and S is diagonal. One can interpret U and V as the coordinates of cells from the 2 data modalities in the shared low dimensional space.

After co-embedding, we calculated cell-cell distances between cells in the two modalities and defined k-nearest neighbors between cells. If we denote all the cells in the mC modality as I, and all the cells in the RNA modality as J, the distance between a cell $i \in I$ and a cell $j \in J$ is given by their Euclidean distance in the shared low dimensional space:

$$d_{ij} = \sqrt{(u_i - v_j)^T (u_i - v_j)},$$

where u_i and v_j are the i'th column of U and the j'th column of V, respectively. We build a bipartite graph, connecting each cell's profile in one modality with its k-nearest neighbors in the other modality. We refer to the these cross-modality neighbors as "*k-partners*", $P_i^{(k)} = \{j \mid d_{ij}$ are the k smallest distances for $j \in J\}$.

Over-splitting score

The over-splitting score for a cluster is the fraction of the *k-partners* of cells in that cluster that are *not* from the same cluster. This metric captures the intuition that clusters should include all of the cells with a similar molecular profile, and not divide cells with similar profiles into

distinct clusters. the over-splitting score is:

$$S_{over}(C_i) = 1 - \frac{1}{|C_i|^2}\sum_{i=1}^{|C_i|} \quad \sum_{j \in P_i^{(k)}} \quad I[C_i = C_j],$$

where i, j are indices of individual cells, C_i is the cluster containing cell i, and $|C_i|$ to represent the cluster size, $I[]$ is the indicator function, and $P_i^{(k)}$ are the k-partners of cell i (with $k = |C_i|$ = cluster size). In other words, the over-splitting score is one minus the mean fraction of a cell's k-partners (with $k = |C_i|$ = cluster size) that are also from the same cluster (C_i). Therefore, the over-splitting score is bounded between zero and one. $S_{over} = 0$ indicates no over-splitting, while larger values of S_{over} indicate less cross-modality stability for a cluster (i.e. more over-splitting).

Under-splitting score

If a cluster cannot be further split, its cells should be biologically equivalent to each other and differ only in terms of measurement noise. Otherwise, the cluster may be under-split. To quantify the equivalence of the cells within a cluster, we define the *self-radius* of a cell as the number of cells which appear equivalent to it in terms of consistent multimodal features. We first measured the distance, d_{ij}, between the mC and RNA profiles of all cell pairs (i, j) after embedding in the common CCA space (see above). We reasoned that any cell pair whose distance is smaller than the distance between the mC and RNA profiles of cell i (i.e. $d_{ij} < d_{ii}$) can be considered equivalent; these cells are as similar to each other as they are to themselves. We thus define a cell's self-radius, r_i, as the number of equivalent cells; in terms of k-partners, this can be expressed as:

$$r_i = arg\ \max_k \{d_{ij} < d_{ii}\ \forall\ j \in P_i^{(k)}\}.$$

The distribution of the self-radii for cells in a cluster will inform us the extent to which the cluster under-split (Fig. 3.2F). For example, if a cluster is not under-split at all, its cells' self-radii should be uniformly distributed between zero and the cluster size. We verified this empirically

with simulation: if we take a group of cells and randomly shuffle their gene-level profiles, we create a homogeneous cluster with no under-splitting. When we do this to all 17 major neuronal clusters, they all behave like ideal clusters without under-splitting (pink line in Fig. 3.2F). Compared to the uniform distribution in the ideal case, an under-split cluster should have an overall much smaller self-radii, indicating it can be potentially further split into several sub-clusters (yellow line in Fig. 3.2F). Therefore the slope of the cumulative distribution of self-radius informs us to what extent a cluster under-split. For an ideal cluster, its cumulative distribution of self-radii is a straight-line, therefore its slope is one. For an under-split cluster, the slope should be greater than one (Fig. 3.2F). We, therefore, defined as the slope of the cumulative distribution of self-radius:

$$S_{under}(C_i) = \frac{Cumulative\ fraction\ of\ cells\ with\ r<= |C_i|/4}{|C_i|/4}$$

where the slope is evaluated at $r = |C_i|/4$, as indicated in the above equation. For an ideal cluster, this score should be one; for an under-split cluster, it should be greater than one.

3.14.7 Computational data fusion with SingleCellFusion (Fig. 3.2)

Several computational methods have been proposed for integrating multiple single-cell sequencing datasets across batches, sequencing technologies, and modalities [37,38,52,116,119]. Many of these methods share a basic strategy of identifying neighbor cells across datasets. However, existing methods have not been optimized to integrate single cells from multiple transcriptomic and epigenomic data modalities, with potentially large systematic differences in the features measured for each dataset. Here, we fused the transcriptomes and DNA methylomes of the snmCAT-Seq dataset, treating the two data modalities as if they were acquired by two independent single-modality experiments in different cells. We developed a new data fusion method, *SingleCellFusion*, for this task (available at: https://github.com/mukamel-lab/SingleCellFusion),

which is based on finding k-partners, i.e. nearest neighbors across data modalities (see the previous section). Nearest neighbor based data integration has been successfully applied to combine multiple RNA-Seq datasets [37,116], while other approaches including canonical correlation analysis (CCA) and non-negative matrix factorization (NMF) have previously been used for the fusion of transcriptomic and epigenomic data [38,119]. Single Cell Fusion is designed to robustly fuse DNA methylation, ATAC-Seq and/or RNA-Seq data. The procedure comprises 4 major steps: preprocessing: within-modality smoothing, cross-modality imputation, and clustering and visualization.

1. **Preprocessing.** We defined a gene-by-cell feature matrix for both transcriptomes and epigenomes. Transcriptomic features are $\log_{10}$(TPM+1) normalized. DNA methylation data is represented by the mean gene body mCH level, normalized by the global (genome-wide) mean mCH level for each cell. We selected genes with significantly correlated gene body mCH and RNA expression (FDR < 0.05) across neuronal cells as features (n=5,107 genes).
2. **Within-modality smoothing.** To reduce the sparsity and noise of feature matrices, we share information among cells with similar profiles using data diffusion [160]. First, we generate a kNN graph of cells based on Euclidean distances in PC space [ndim = 50, k=30]. We next construct a sparse weighted adjacency matrix A. We first apply a Gaussian kernel on the distance between cell i and cell j: ${A^{(1)}}_{ij} \propto exp\ (-{d_{ij}}^2/{\sigma_i}^2)$, where σ_i is the distance to the k_a-th [k_a=5] nearest neighbor of cell i. We set diagonal elements to zero, ${A^{(1)}}_{ii} = 0$, and also set all elements to zero if they are not part of the kNN. We then symmetrize the matrix, $A^{(2)} = A^{(1)} + {A^{(1)}}^T$, and normalize each row: ${A^{(3)}}_{ij} = A^{(2)}/a_i$, where $a_i = \sum_j {A^{(2)}}_{ij}$. Finally, we reweight the adjacency matrix with a parameter, p, that explicitly

controls the relative contribution of diagonal and non-diagonal elements: $A = p\,I + (1-p)\,A^{(3)}$, where I is the identity matrix. We chose p=0.9 for DNA methylation; p=0.7 for RNA. Finally, we smooth the raw feature matrix by matrix multiplication with the adjacency matrix.

3. **Cross-modality imputation by Restricted k-Partners (RKP)**. Each cell has a set of measured features in one data modality (RNA or mC), which we call the "source modality." The goal of this step of the analysis is to impute the missing features from the other data type, called the "target modality." For each cell in the source modality, we select a set of k-partners in the target modality and use the average of the k-partners' features to estimate the missing modality for the original cell. However, care must be taken to avoid hub cells in the target modality which form k-partner relationships with a large fraction of all cells in the source modality. One way to avoid hub cells is by including only mutual nearest neighbors (MNN) [116]. We developed an alternative approach, restricted k-partners (RKP), that efficiently finds a set of k-partners for every source modality cell, while ensuring that every target-modality cell is connected with a roughly equal number of source modality cells.

As above, we first reduce the dimensionality of both source and target data matrices by canonical correlation analysis, retaining the top 50 canonical components. We then iterate over all cells in the source modality (in random order) k times, connecting each with its most similar partner cell in the target modality. Whenever a target modality cell is partnered with more than k' source modality cells, we remove it from the pool of eligible target cells so that it will not be the partner of additional source cells. We set $k' = [\,z\,k\,N_{source}/N_{target}\,]_+$, where and $z \geq 1$ is a relaxation parameter that determines how

much variability in the number of partners is allowed across target modality cells and $[]_+$ is the ceiling function. If $z = 1$ then every target cell will be connected to exactly k' or $k' - 1$ cells. We set $z = 3$, meaning that any individual target modality cell can have at most 3 times as many partners as the average. This algorithm is efficient and, in our analyses, provides robust k-partner graphs for cross-modality data imputation.

Having determined each source cell's restricted k-partners, we next impute the target features by averaging over the smoothed feature vectors of each cell's k-partners.

4. **Clustering and visualization.** After imputation, we cluster and visualize cells from the 2 data modalities as if they are from the same dataset. We reduce dimensionality for all cells by performing PCA, keeping the top 50 PCs of the (measured and imputed) DNA methylation features. This cell-by-PCs matrix is further used for downstream embedding and clustering. Next, we perform UMAP embedding [116] on the PC matrix [n_neighbors=30, min_dist=0.5]. Finally, we perform Leiden clustering (Traag [116] on the kNN graph (symmetrized, unweighted) generated from the final PC matrix [Euclidean distance, k=30, resolution=0.3, 1, 2, 4].

3.14.8 Evaluation of Computational Data Fusion Methods (Supplementary Fig. 3.3)

We tested five data fusion tools: 1) Scanorama [37]; 2) Harmony [52]; 3) Seurat [36]; 4) LIGER [38,161] and 5) SingleCellFusion (the present study). For tools 1 to 3, we used the same set of highly variable genes (HVG, Top 2000 genes identified by Seurat FindVariableGenes function) identified from the transcriptome matrix as starting features; for algorithms 4 and 5, we used top 5000 genes having the highest correlation between their RNA and mCH level. These genes were chosen based on their overall accuracy (see below). We reversed the methylation values (i.e. max(X) - X, where X denotes the cell-by-gene mCH fraction matrix) before data fusion to account for the negative

correlation of mCH fraction and RNA expression.

Below we describe the data fusion process of each tool, starting from the per cell normalized RNA-HVG matrix and reversed mCH-HVG matrix. After obtaining the decomposed matrix (PCs from 1,2,3,5 or H matrix from 4), we then evaluate the data fusion performance using metrics described below. For reproducibility, we uploaded all the steps and input files here: https://github.com/lhqing/snmCAT-seq_integration

1) For Scanorama, we used these parameters (sigma=100, alpha=0.1, knn=30) to perform the data fusion and dimension reduction using Scanorama V1.7 on the scaled (via scanpy.pp.scale) mC and RNA matrix. We used the top 20 fused PCs (n_components = 20) for data fusion evaluation.

2) Unlike Scanorama, Harmony directly takes dimension reduction matrices as input. Therefore, we first run PCA separately (n_components = 20) on the scaled mCH and RNA matrix first, and run Harmony (pyharmony from https://github.com/jterrace/pyharmony) with default parameters on the concatenated PCs. Fused PCs generated by Harmony were then used for evaluation.

3) For Seurat, we followed the Seurat (v4.0.0) vignette steps to perform data fusion (https://satijalab.org/seurat/articles/integration_introduction.html). When calculating anchors for data fusion (FindTransferAnchors), we use the RNA matrix as the reference matrix and mCH matrix as the query matrix and using CCA as the dimension reduction method. We then transfer the mCH matrix to the RNA space using the anchors and run PCA (n_components = 20) on the concatenated (mCH and RNA) matrix after the transfer.

4) For LIGER, we followed the tutorial from developers (http://htmlpreview.github.io/?https://github.com/welch-

lab/liger/blob/master/vignettes/online_iNMF_tutorial.html) and used the online_iNMF algorithm (Gao et al., 2020) with default parameters to perform data fusion and used the normalized matrix H (the cells' decomposed matrix from the online iNMF algorithm) for fusion evaluation.
Finally, the SingleCellFusion analysis was described in the manuscript, we used the fused PCs for evaluation.

3.14.9 Metrics for evaluation of data fusion

We used three different approaches to evaluate the single-cell data fusion results. First, We ran UMAP on the decomposed matrix from each tool to provide an overview of the fused dataset.

Second, We utilize the ground-truth information from the snmCAT-seq to calculate a self-radius at the single-cell level. Specifically, we first construct a nearest-neighbor index using Annoy (v1.17.0) on the decomposed matrix (euclidean distance). For the same cell, if its RNA vector is the mCH vector's Kth neighbor, we then use d=K as the self-radius. The quality of the data fusion can be normalized by d/2N, where N is the total number of snmCAT-seq cells involved in the analysis. The value of d/2N ranges from 0 to 1, with smaller values indicating good fusion and larger values indicating inadequate fusion of mC and RNA profiles of the same cell.

Finally, we performed Leiden co-clustering on the decomposed matrix (with different resolution parameters to obtain 17 co-clusters in all tools, which is the number of major neuronal cell types) and calculated the co-cluster accuracy as the fraction of cells whose RNA and mC profiles were assigned to the same cluster. This accuracy can be calculated for each co-cluster or the whole dataset. Higher accuracy means good fusion, and a low accuracy indicates inadequate fusion.

3.14.10 Correlation analysis of RNA expression and gene body DNA methylation (Fig. 3.3)

For each gene, we compute the Spearman correlation coefficient between RNA expression

($\log_{10}$(TPM+1)) and gene body mCH (normalized to global mCH of each cell). To determine if a correlation is statistically significant, we randomly shuffled cell labels to generate an empirical null distribution. Significantly correlated genes are defined with empirical FDR < 0.05. Applying this method to 3,898 neurons in the snmCAT-seq dataset, we get 5,145 genes with a significant negative correlation between RNA and mCH (RNA-mCH coupled).

3.14.11 Eta Squared of Genes Across Clusters (Fig. 3.3)

For each gene used for correlation analysis we compute the η^2 across neuronal sub clusters (n=52) generated from ensemble methylomes (Fig. 3.4) for both RNA ($\log_{10}$(TPM+1)) and gene body mCH (normalized by global mCH of each cell) signals. We also compute η^2 across 10X RNA-seq clusters for the same genes.

3.14.12 H3K27me3 ChIP-Seq data processing (Fig. 3.3)

We downloaded published H3K27me3 ChIP-Seq data of purified excitatory and inhibitory neurons from the human prefrontal cortex [122]. We calculated the average ChIP-Seq signal intensity (RPKM) across the gene body for excitatory and inhibitory neurons.

3.14.13 Fusion of DNA methylome and snATAC-Seq data (Fig. 3.4)

Ensemble methylomes and snATAC-seq data from neurons and glia were fused separately using our recently developed Single Cell Fusion method (see section **Computational data fusion with SingleCellFusion**). The top 4000 variable genes across clusters in the snmCAT-seq and snATAC-seq data were identified using a Kruskal-Wallis test; 1,652 genes were identified as being variable in both datasets and were used for the subsequent data fusion. For snATAC-seq the gene body was extended to include the promoter region (2kb upstream TSS). Prior to the fusion of mCH and open chromatin levels at gene bodies were smoothed to reduce sparseness (k = 20, ka = 4, epsilon = 1, p = 0.9; see section **Within-modality smoothing**) using a diffusion-based smoothing

method adapted from MAGIC ([160]). A constrained k-nearest neighbors graph was generated among cells across 2 datasets (k=20, z=10; see section **Cross-modality imputation by Restricted k-Partners**). Instead of calculating Euclidean distance in reduced dimensions, here we simply used Spearman correlation across 1,652 genes as the distance measure between cells. We used the kNN graph to impute the gene body mCH profile for each ATAC-Seq nucleus. The observed (ensemble methylomes) and imputed (snATAC-Seq nuclei) gene body mCH levels were then jointly used for Leiden clustering and UMAP embedding. Each snATAC-seq nucleus was assigned to a major cell type if at least half of its restricted k-Partners belonged to that cell type, the remaining cells were removed from subsequent analysis (n=499, 3.98%).

3.14.14 snmCAT-seq - snRNA-seq integration (Fig. 3.5, excitatory, inhibitory, non-neuron separately)

To perform the integration analysis of snmCAT-seq transcriptome and snRNA-seq, we separate the cells into three broad classes: excitatory neurons, inhibitory neurons, and non-neuronal cells. The RNA features used for the integration by Scanorama come from two sources for each cell class: 1) highly variable genes across individual cells; 2) cluster level RNA marker genes. To validate that the cluster level RNA marker genes are relevant for neuronal processes, we performed a synapse-specific GO enrichment test using the SynGO terms and all brain expressed genes as background [124]. The -log(adjusted P-value) of SynGO biological process enrichment in each selected gene set is color-coded on the sunburst chart of the hierarchical SynGO terms (Fig. 3.5E-F).

We then used the union of RNA features found in snmCAT-seq transcriptome and snRNA-seq for Scanorama integration and PCA calculation. The dominant PCs were then used to perform a co-clustering analysis on the cells profiled by snmCAT-seq cell or snRNA-seq. Instead of

directly using the co-clustering results, we used this intermediate clustering assignment to calculate the overlap score between the original methylome ensemble clusters and the snRNA-seq clusters. The overlap score range from 0 to 1 is defined as the sum of the minimum proportion of samples in each cluster that overlapped within each co-cluster [84], a higher score between one methylome cluster and one snRNA-seq cluster indicate they consistently co-clustered.

3.14.15 Cell type dendrogram and sub-cluster merge along the lineage (Fig. 3.6)

The cell-type hierarchy of inhibitory and excitatory cells was calculated separately using the concatenated PCs from mCG and mCH as the features used for computing cluster centroids. We used *scipy.cluster.hierarchy.linkage* function to calculate the ward linkage. Based on the linkage results, we merged the CpG sites from single-cell ALLC files in 2 steps: 1) we merged the single-cell ALLC files into each of the sub-clusters, 2) we then merge the sub-clusters into all nodes that appeared in the dendrograms. The merged CpG ALLC files are then used in the lineage-DMR analysis.

3.14.16 Neural lineage-specific DMR calling and motif enrichment analysis (Fig. 3.6)

We used the *methylpy findDMR* function [41] to identify mCG lineage-DMRs for each pair of lineages using merged ALLC files. The DMRs identified by methylpy in each branch comparison are further filtered by mCG rate difference > 0.3 and the number of differentially methylated sites (DMS) >= 2. Lineage pairs with $>10^4$ DMRs identified were used for motif enrichment analysis and TF marker identification. For each of these DMR sets, we use AME [162] to perform motif enrichment (fisher's exact test) analysis with the motifs' Position Weight Matrix (PWM) from the JASPAR database (JASPAR2018 CORE Vertebrates) [163]. The DMRs are length standardized into ±250bp of region center before motif scanning. Tissue-specific DMRs (without

brain tissue, and standardized in the same way) from the Roadmap Epigenomics project [41,136] were used as the background.

3.14.17 TF binding preference to methylated motifs (Fig. 3.6)

To further investigate the methylation level impact on the potential TF binding sites, we selected all the mCG DMSs ±25bp regions from the branch-DMRs and ran motif enrichment using motifs identified from the methyl-SELEX experiment [6]. In each branch pair, we used the left-DMSs as the background of right-DMS to find the right-branch-specific motif and *vice versa*. The significantly enriched "TF motif - branch" combinations were then intersected with the corresponding branch pair's DEG and DMG list to infer their gene mCH or RNA specificity.

3.14.18 Chromatin accessibility analysis of TF binding motifs (Fig. 3.6F-M and Supplementary Fig. 3.6I-L)

Genome-wide sites matching TF binding motifs (motif matches) identified by methyl-SELEX [6] were identified using FIMO 4.11.4 [86] with the following parameters --max-stored-scores 500000 --max-strand --thresh 1e-5. Methyl-SELEX only quantified the effect of CpG methylation on TF binding. Therefore only genomics sites containing CG dinucleotides were selected for further analyses. For each major cell type, the density of methylated GCY sites or ATAC-seq reads was quantified for motif matches that overlap with hypomethylated or hypermethylated DMRs. Fig. 3.6F and Supplementary Fig. 3.6I show the average chromatin accessibility at motif matches across major cell types. TF binding motifs were ranked by the difference of chromatin accessibility between motif matches located in hypomethylated and hypermethylated DMRs. To test the enrichment of MethylPlus and MethylMinus TFs, the ranked motif list was divided into 5 bins and the enrichment or depletion in each bin was tested using Matlab *hygecdf* function.

3.14.19 Partitioned heritability analysis (Fig. 3.7 and Supplementary Fig. 3.7)

Bulk human fetal frontal cortex methylomes from a PCW 20 donor [2] and a PCW 19 donor[121] were previously published. Fetal frontal cortex DMRs were identified using *methylpy findDMR* function[41] by comparing to adult bulk neuronal (NeuN+) and non-neuronal (NeuN-) methylomes[2]. Fetal brain Dnase-seq samples included fetal day 85d (GSM595922, GSM595923), 96d (GSM595926, GSM595928), 101d (GSM878650), 104d (GSM878651), 105d (GSM1027328), 109d (GSM878652), 112d (GSM665804), 117d (GSM595920) and 142d (GSM665819)[136]. Mapped reads files (BED format) were downloaded followed by DNase-seq peak calling using MACS2 2.0.10 with q-value < 0.01. Fetal brain DNase-seq peaks were defined as the union DNase-seq peaks of fetal brain DNase-seq datasets and were supported by at least two samples.

Summary statistics were downloaded from the Psychiatric Genomics Consortium portal (https://www.med.unc.edu/pgc/) for neuropsychiatric trait GWAS - ADHD[164], Aggression[165], Anorexia nervosa[166], Anxiety[167], ASD[168], Bipolar[169], Cognitive Performance[170], Educational Attainment[170], Alzheimer's[171], Internalizing, Loneliness[172], Major Depression[173], Neuroticism[174], OCD[175], Schizophrenia (PGC2)[176] and Schizophrenia (PGC1)[177].

The partitioned heritability analysis was performed using LD Score Regression (LDSC) Partitioned Heritability[135]. The partitioned heritability analysis was performed by constructing joint linear models by providing multiple regulatory element annotations in addition to the "baseline" annotation. Alternatively, we performed analyses by constructing individual models by comparing each annotation of regulatory elements individually against the "baseline". We built a "baseline" annotation using tissue-specific DMRs from non-brain human tissues[41] to control for generic gene regulation characteristics. Partitioned heritability analyses using cell-type specifically

expressed genes were performed as described in[137]. The reported q-values were derived from the "*Coefficient_z.score*" values reported by LDSC Partitioned Heritability.

3.14.20 Prioritization of trait-associated cell types using RolyPoly (Supplementary Fig. 3.7)

Although RolyPoly was originally developed to associate GWAS summary statistics with transcriptome data, we adapted the method to analyze epigenomic features such as DMRs and ATAC-seq peaks. RolyPoly analysis was performed using summary statistics for schizophrenia[176], bipolar disorder[169], ASD[168] and educational attainment[170]. DMRs or ATAC-seq peaks identified for each cell type that overlapped with the top 10,000 variants with the smallest p-value were provided as the feature list. As recommended by the RolyPoly tutorial (https://cran.r-project.org/web/packages/rolypoly/vignettes/intro.html), the absolute value of Z-scores computed for CG methylation level or ATAC-seq signal across samples were provided as in place of expression data. The analysis was performed with 100 times bootstrapping.

3.15 Data and Code Availability

Raw and processed data included in this study were deposited to NCBI GEO/SRA with accession number GSE140493. Methylome and transcriptomic profiles generated by snmCAT-seq from H1 and HEK293T cells can be visualized at [http://neomorph.salk.edu/Human_cells_snmCT-seq.php]. snmCAT-seq generated from brain tissues can be visualized at [http://neomorph.salk.edu/human_frontal_cortex_ensemble.php]. snRNA-seq data is available for download from the Neuroscience Multi-omics Archive (https://assets.nemoarchive.org/dat-s3creyz). The code for SingleCellFusion is available from https://github.com/mukamel-lab/SingleCellFusion. A detailed bench protocol for snmCAT-seq and future updates to the method can be found at https://www.protocols.io/view/snmcat-v1-bwubpesn.

3.16 Author Contributions

J.R.E, and C.L. conceived the study. J.R.E, E.A.M, M.M.B, B.R., E.L. S.L. and J.R.D supervised the study. C.L., B-A.W. and Z.Z. developed the snmCAT-seq method. C.L., B-A.W., R.C., A.B., A.R., and J.R.N. generated the snmCAT-seq data. C.L., R.C. and J.R.N. generated the snmC-seq data. K.S., T.E.B., R.D.H, L.H., S.L. and E.L. generated and analyzed the snRNA-seq data. R.F., S.P., X.W. and B.R. generated and analyzed the snATAC-seq data. D.A.D. and D.C.M. acquired human brain specimens. D-S.L. and J.R.D reanalyzed the sn-m3C-seq data. H.L., F.X., C.L., W.D., E.J.A., D-S.L., J.Z., S-Y.N. analyzed the data. C.L., H.L. and F.X. drafted the manuscript. J.R.E, E.A.M, T.E.B., R.D.H, D.A.D and D.C.M edited the manuscript.

3.17 Acknowledgements

Chapter 3, in full, is a reprint of the material as it appears in Cell Genomics, 2022. "Single nucleus multi-omics identifies human cortical cell regulatory genome diversity." Chongyuan Luo, Hanqing Liu, Fangming Xie, Ethan J. Armand, Kimberly Siletti, Trygve E. Bakken, Rongxin Fang, Wayne I. Doyle, Tim Stuart, Rebecca D. Hodge, Lijuan Hu, Bang-An Wang, Zhuzhu Zhang, Sebastian Preissl, Dong-Sung Lee, Jingtian Zhou, Sheng-Yong Niu, Rosa Castanon, Anna Bartlett, Angeline Rivkin, Xinxin Wang, Jacinta Lucero, Joseph R. Nery, David A. Davis, Deborah C. Mash, Rahul Satija, Jesse R. Dixon, Sten Linnarsson, Ed Lein, M. Margarita Behrens, Bing Ren, Eran A. Mukamel, and Joseph R. Ecker. This work was supported by NIH grants: 5R21HG009274, 5R21MH112161, and 5U19MH114831 to Joseph Ecker; R01MH125252 and U01HG012079 to Chongyuan Luo; R01HG010634 to Joseph Ecker and Jesse Dixon; and U01MH114812 to Ed Lein. Joseph Ecker is an investigator of the Howard Hughes Medical Institute. Wayne Doyle is supported by an NIH training award 5T32MH020002. Postmortem human brain tissues were obtained from the NIH NeuroBioBank at the University of Maryland Brain and Tissue Bank and

the University of Miami Brain Endowment Bank. The authors thank the tissue donors and their families for their invaluable contributions to the advancement of science. The authors thank the QB3 Macrolab at UC Berkeley for the purification of Tn5 transposase. Work at the Center for Epigenomics was supported in part by the UC San Diego School of Medicine. The dissertation author was the primary investigator and author of this paper.

3.18 Tables

Table 3.1. Genomic profiling methods discussed in chapter 3.

Method	Description	Reference
snmC-seq	Multiplexed single-nucleus DNA methylome profiling	[3]
snmC-seq2	Improved single-nucleus DNA methylome profiling methods with increased read mapping and enhanced throughput	[28]
snmCAT-seq	Single-nucleus joint profiling of DNA methylome, chromatin accessibility (based on NOMe-seq) and transcriptome	This Study
sn-m3C-seq	Single-nucleus joint profiling of chromatin conformation and DNA methylome	[24]
NOMe-seq	Profiling of nucleosome footprint and chromatin accessibility using in vitro GpC methyltransferase labeling	[114]
Smart-seq2	The generation and amplification of full-length cDNA and sequencing libraries	[111]
snRNA-seq	Single-nucleus RNA-seq. The human brain snRNA-seq in this study was generated using the 10x Genomics Chromium platform.	This Study
ATAC-seq	Assay for chromatin accessibility using Tn5 transposon	[178]
snATAC-seq	Combinatorial indexing-assisted single-cell assay for transposase-accessible chromatin	[30]

Table 3.2 Key resources table for chapter 3.

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Cell Lines		
HEK293T cells	Salk Institute Stem Cell Core	N/A
H1 hESC cells	WiCell	WA01
Experimental Models: Human Brain Tissue		
Brodmann area 10 (M_21yr)	NIH NeuroBioBank at University of Maryland Brain and Tissue Bank	UMB5577
Brodmann area 10 (M_29yr)	NIH NeuroBioBank at University of Maryland Brain and Tissue Bank	UMB5580
Medial Frontal Gyrus (M_25yr_1)	NIH NeuroBioBank at University of Maryland Brain and Tissue Bank	UMB4540
Brodmann area 10 (M_58yr)	NIH NeuroBioBank at University of Miami Brain Endowment Bank	NDARKD326LNK
Brodmann area 10 (M_25yr_2)	NIH NeuroBioBank at University of Miami Brain Endowment Bank	NDARKJ183CYT
Brodmann area 44-45, Brodmann area 46	Allen Institute for Brain Science	H18.30.002
Recombinant Proteins		
Custom Tn5 Transposase	MacroLab, University of California Berkeley	Custom Protein Purification
Deposited Data		
snmCAT-seq data generated from HEK293T and H1 hESC cells	This Study	GSE140493

Table 3.2 Key resources table for chapter 3 continued.

snmCAT-seq data generated from UMB5577 and UMB5580	This Study	GSE140493
snmC-seq and snmC-seq2 data generated from NDARKD326LNK and NDARKJ183CYT	This Study	GSE140493
snATAC-seq data generated from UMB4540	This Study	GSE140493
scRNA-seq data generated from H18.30.002	This Study	
sn-m3C-seq data generated from UMB5577 and UMB5580	[24]	GSE130711
snmC-seq data generated from UMB4540	[3]	GSE97179
H3K27me3 ChIP-seq	[122]	Synapse (syn12034263)
Oligonucleotides		
dT30VN_4	Integrated DNA Technologies	5'-/5SpC3/AAGCAGUGGUAUCAACGCAGAGUACUTTTTTUTTTTTUTTTTTUTTTTTUTTTTVN-3' (HPLC purified)
N6_2	Integrated DNA Technologies	5'-/5SpC3/AAGCAGUGGUAUCAACGCAGAGUACNNNNNN-3' (HPLC purified)
TSO_3	Exiqon (now Qiagen)	5'-/5SpC3/AAGCAGUGGUAUCAACGCAGAGUGAAUrGrG+G-3' (HPLC purified)
ISPCR23_2	Integrated DNA Technologies	5'-/5SpC3/AAGCAGUGGUAUCAACGCAGAGU-3' (HPLC purified)

Table 3.2 Key resources table for chapter 3 continued.

Software and Algorithms		
SingleCellFusion	This Study	https://github.com/mukamel-lab/SingleCellFusion
LIGER	[38]	https://github.com/welch-lab/liger
Bismark v0.14.4	[149]	http://www.bioinformatics.babraham.ac.uk/projects/bismark/; RRID:SCR_005604
STAR 2.5.2b	[150]	https://github.com/alexdobin/STAR; RRID:SCR_015899
YAP	[15]	https://hq-1.gitbook.io/mc/
ALLCools	[15]	https://github.com/lhqing/ALLCools
methylpy	[41]	https://github.com/yupenghe/methylpy
Seurat v4.0.0	[36]	https://satijalab.org/seurat/; RRID:SCR_016341
Scanorama v1.7	[37]	https://github.com/brianhie/scanorama
Harmony (pyharmony)	[52]	https://github.com/iandday/pyharmony

3.19 Figures

Figure 3.1. snmCAT-seq generates single-nucleus multi-omic profiles of the human brain. (**A**) Schematic diagram of snmCAT-seq. (**B**) Boxplot comparing the number of genes detected in each cell/nucleus by different single-cell or single-nucleus RNA-seq technologies. (**C**) Boxplot comparing the genome coverage of single-nucleus methylome between snmCAT-seq and snmC-seq. (**D-G**) snmCAT-seq methylome was compared to other single-cell methylome methods with respect to mapping rate (**D**), library complexity (**E**), enrichment of CpG islands (**F**) and coverage uniformity (**G**). (**H**) UMAP embedding of human frontal cortex snmCAT-seq profiles. (**I**) UMAP embedding of transcriptome, methylome and chromatin accessibility profiled by snmCAT-seq for *ADARB2*. The cells are colored by gene expression (CPM, counts per million), chromatin accessibility (MAGIC imputed GmCY ratio, see methods), non-CG DNA methylation (HmCH ratio normalized per cell) and CG DNA methylation (HmCG ratio normalized per cell). (**J**) Comparison of marker gene expression between clusters identified using snmCAT-seq and matching clusters identified using snRNA-seq. The matching clusters were merged from original snRNA-seq clusters based on cell integration and label transfer (see methods). Dot sizes represent the fraction of cells with detected gene expression. Dot colors represent the mean expression level across the cells with detected gene expression. (**K**) UMAP embedding of snmCAT-seq transcriptome and snRNA-seq cells after integration. (**L**) Confusion matrix comparing snmCAT-seq clusters to snRNA-seq clusters. The plot is colored by overlapping scores between clusters. (**M**) Comparison of marker gene non-CG methylation (HmCH) between clusters identified using snmCAT-seq and matching clusters identified using snmC-seq. Dot sizes represent the mean cytosine coverage per cell. Dot colors represent the mean HmCH ratio. *For non-neuronal cell markers, gene body CG methylation (HmCG) levels were compared between snmCAT-seq and snmC-seq. (**N**) Comparison of chromatin accessibility profiled by snmCAT-seq and snATAC-seq at cell-type-specific open chromatin sites. The left and right heatmaps show the density of methylated GCY sites and the density of ATAC-seq reads, respectively.

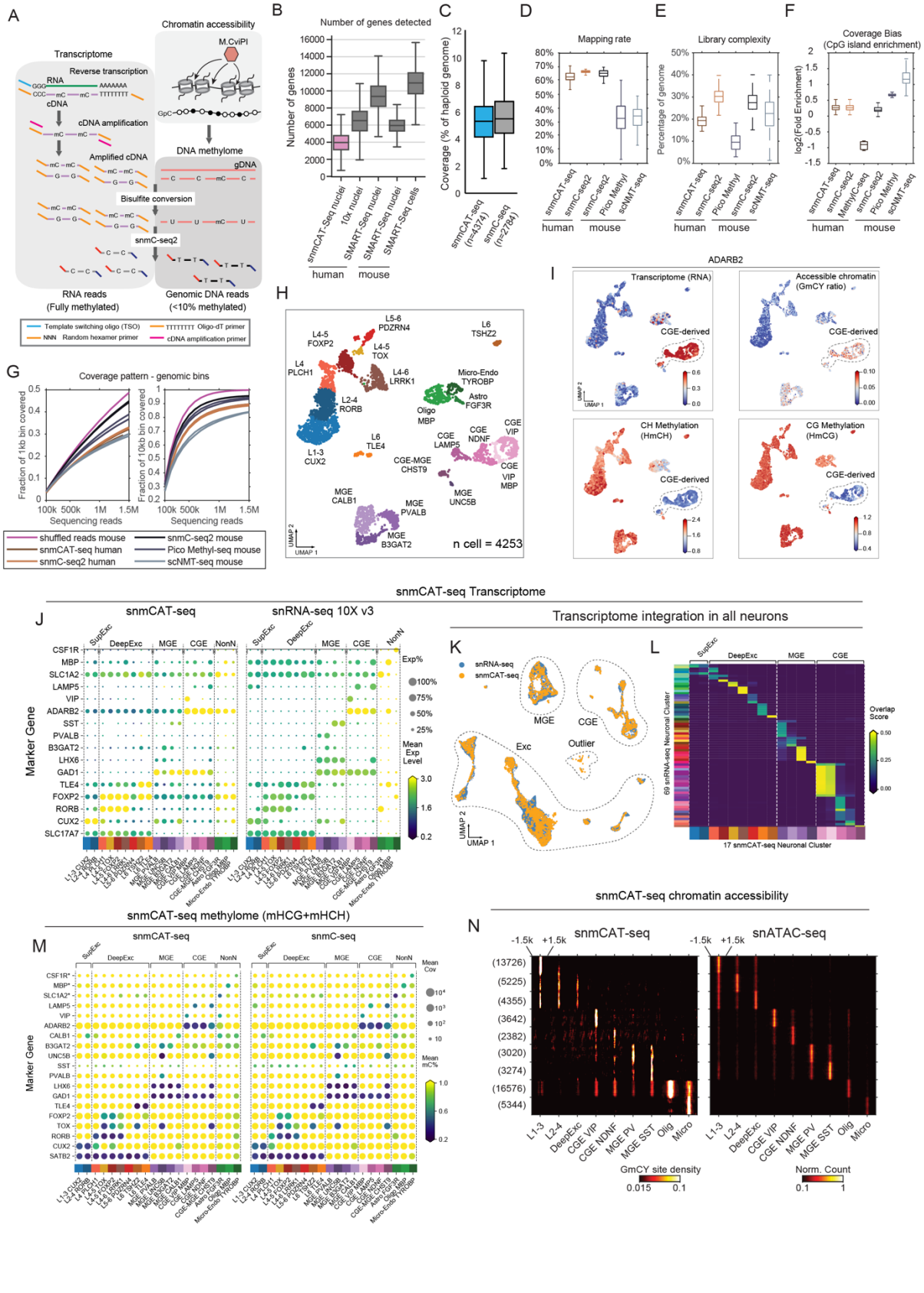

A
Chromatin accessibility
M.CviPI
Transcriptome
Reverse transcription
RNA
cDNA
cDNA amplification
Amplified cDNA
DNA methylome
gDNA
Bisulfite conversion
snmC-seq2
RNA reads
(Fully methylated)
Genomic DNA reads
(<10% methylated)
Template switching oligo (TSO)
TTTTTTT Oligo-dT primer
NNN Random hexamer primer
cDNA amplification primer
B
Number of genes detected
Number of genes
snmCAT-Seq nuclei
10x nuclei
SMART-Seq nuclei
SMART-Seq nuclei
SMART-Seq cells
human
mouse
C
Coverage (% of haploid genome)
snmCAT-seq (n=4374)
snmC-seq (n=2784)
D
Mapping rate
E
Library complexity
Percentage of genome
F
Coverage Bias (CpG island enrichment)
log2(Fold Enrichment)
snmCAT-seq
snmC-seq2
Pico Methyl
scNMT-seq
MethylC-seq
G
Coverage pattern - genomic bins
Fraction of 1kb bin covered
Fraction of 10kb bin covered
Sequencing reads
shuffled reads mouse
snmCAT-seq human
snmC-seq2 human
snmC-seq2 mouse
Pico Methyl-seq mouse
scNMT-seq mouse
H
L5-6 PDZRN4
L6 TSHZ2
L4-5 FOXP2
L4-5 TOX
L4 PLCH1
L4-6 LRRK1
Micro-Endo TYROBP
L2-4 RORB
Astro FGF3R
Oligo MBP
L6 TLE4
CGE LAMP5
CGE NDNF
CGE VIP
L1-3 CUX2
CGE-MGE CHST9
CGE VIP MBP
MGE CALB1
MGE UNC5B
MGE PVALB
MGE B3GAT2
n cell = 4253
I
ADARB2
Transcriptome (RNA)
Accessible chromatin (GmCY ratio)
CH Methylation (HmCH)
CG Methylation (HmCG)
CGE-derived
J
snmCAT-seq Transcriptome
snmCAT-seq
snRNA-seq 10X v3
SupExc
DeepExc
MGE
CGE
NonN
Marker Gene
Exp%
Mean Exp Level
K
Transcriptome integration in all neurons
snRNA-seq
snmCAT-seq
Exc
Outlier
L
69 snRNA-seq Neuronal Cluster
17 snmCAT-seq Neuronal Cluster
Overlap Score
M
snmCAT-seq methylome (mHCG+mHCH)
snmC-seq
Mean Cov
Mean mC%
N
snmCAT-seq chromatin accessibility
snATAC-seq
-1.5k
+1.5k
GmCY site density
Norm. Count

Figure 3.2. Integrative analysis of RNA and mC features cross-validates neuronal cell clusters.

(**A**) Schematic diagram of the cluster cross-validation strategy using matched single-cell methylome and transcriptome profiles. (**B-C**) Mean squared error, Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) between RNA expression profile (B) or mCH (C) of individual cells and cluster centroids were plotted as a function of the number of clusters. The shaded region in each plot highlights the range between the minimum and the minimum + standard error for the curve of test-set error. Cross-validation analysis was performed in reciprocal directions by performing Leiden clustering using mC (B) or RNA (C) profiles followed by cross-validation using the matched RNA (B) and mC (C) data, respectively. (**D**) Schematic diagram of the over- and under-splitting analysis using matched single-cell methylome and transcriptome profiles. (**E**) Over-splitting of mC-defined clusters was quantified by the fraction of cross-modal *k*-partners found in the same cluster defined by RNA. Shades indicate confidence intervals of the mean. (**F**) Under-splitting of clusters was quantified as the cumulative distribution function of normalized self-radius. (**G**) Scatter plot of over-splitting (S_{over}) and under-splitting (S_{under}) scores for all neuronal clusters. Dot sizes represent cluster size. The actual data trend shows a linearly regressed line on both major clusters and sub-clusters. (**H**) Joint UMAP visualization of snmCAT-seq transcriptome and methylome by computational fusion using the SingleCellFusion method, assuming snmCAT-seq transcriptomes and methylome were derived from independent datasets. (**I**) Accuracy of computational fusion determined by the fraction of cells with matched transcriptome and epigenome profile grouped in the same cluster. (**J**) Confusion matrix normalized by each row. Each row shows the fraction of cells from each joint cluster that are from each cluster defined in Fig. 3.4. Transcriptomes and DNA methylomes are quantified separately.

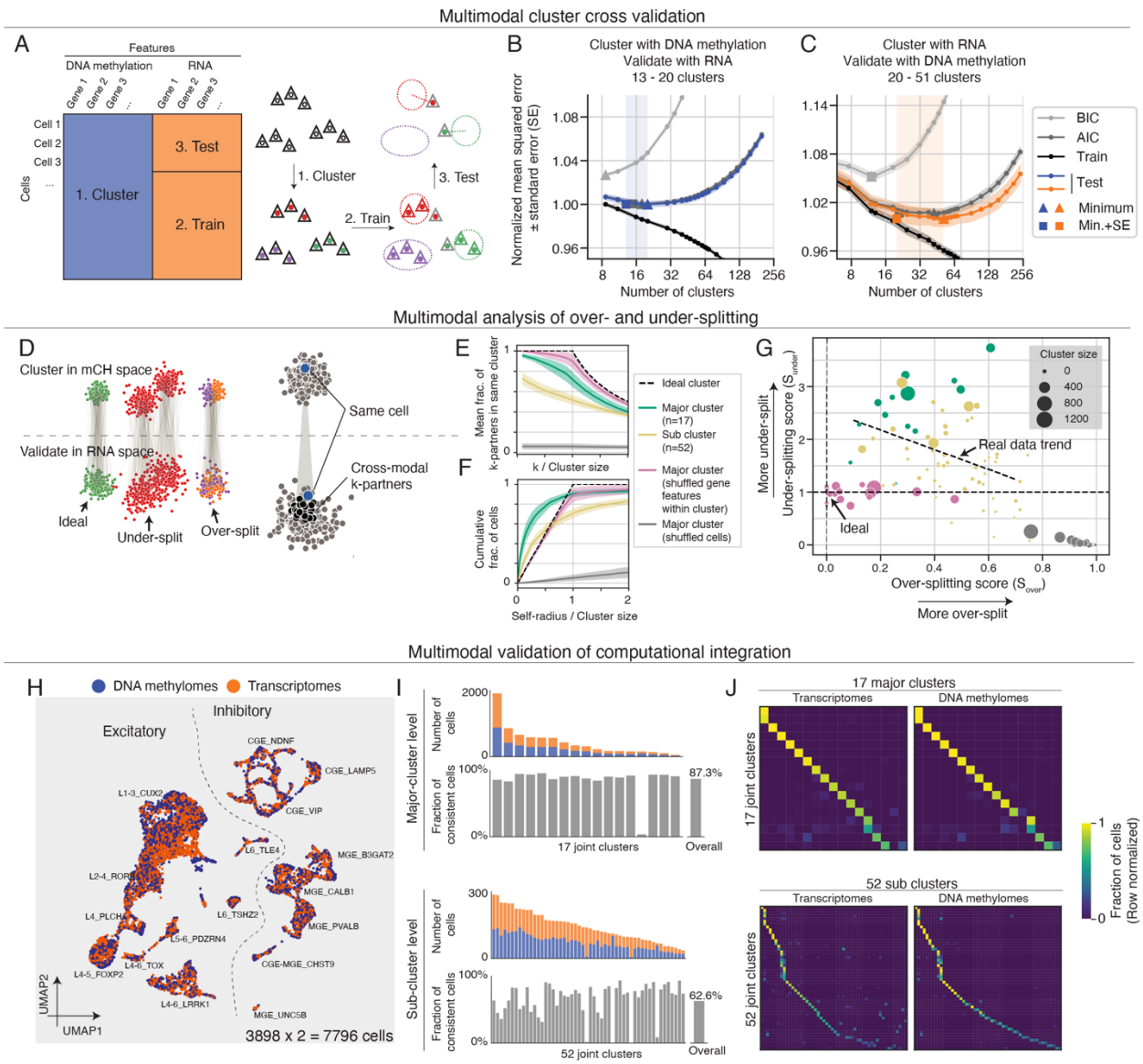

Multimodal cluster cross validation
A
Features
DNA methylation
RNA
Gene 1
Gene 2
Gene 3
Cell 1
Cell 2
Cell 3
Cells
1. Cluster
2. Train
3. Test
B
Cluster with DNA methylation
Validate with RNA
13 - 20 clusters
Normalized mean squared error ± standard error (SE)
Number of clusters
C
Cluster with RNA
Validate with DNA methylation
20 - 51 clusters
BIC
AIC
Train
Test
Minimum
Min.+SE
Multimodal analysis of over- and under-splitting
D
Cluster in mCH space
Validate in RNA space
Same cell
Cross-modal k-partners
Ideal
Under-split
Over-split
E
Mean frac. of k-partners in same cluster
k / Cluster size
Ideal cluster
Major cluster (n=17)
Sub cluster (n=52)
Major cluster (shuffled gene features within cluster)
Major cluster (shuffled cells)
F
Cumulative frac. of cells
Self-radius / Cluster size
G
Cluster size
0
400
800
1200
Real data trend
Ideal
More under-split
Under-splitting score (S_under)
Over-splitting score (S_over)
More over-split
Multimodal validation of computational integration
H
DNA methylomes
Transcriptomes
Inhibitory
Excitatory
CGE_NDNF
CGE_LAMP5
L1-3_CUX2
CGE_VIP
L6_TLE4
MGE_B3GAT2
L2-4_RORB
MGE_CALB1
L4_PLCH1
L6_TSHZ2
MGE_PVALB
L5-6_PDZRN4
L4-6_TOX
CGE-MGE_CHST9
L4-5_FOXP2
L4-6_LRRK1
MGE_UNC5B
UMAP2
UMAP1
3898 x 2 = 7796 cells
I
Major-cluster level
Number of cells
Fraction of consistent cells
87.3%
17 joint clusters
Overall
Sub-cluster level
62.6%
52 joint clusters
J
17 major clusters
Transcriptomes
DNA methylomes
17 joint clusters
52 sub clusters
52 joint clusters
Fraction of cells (Row normalized)

Figure 3.3. Single-cell correlation analysis of RNA expression and gene body non-CG methylation.

(**A**) Scatter plots of gene body mCH (normalized by the global mean mCH of each cell) and gene expression ($\log_{10}$(TPM+1)) of example genes (KCNIP4, ADARB2, GPC5) across all neuronal cells. Cells are colored by major cell types defined in Fig. 3.4. The Spearman correlation coefficient (r) is shown for each example gene. (**B**) Distribution of Spearman correlation coefficient between gene expression and gene body mCH. Blue represents the actual distribution; Gray represents the distribution with randomly shuffled cell labels. (**C-E**) Scatter plot of correlation coefficient of gene body mCH and RNA versus the fraction of variance explained by cell type (η^2) from 3 different datasets/features: snmCAT-seq mCH, snmCAT-seq RNA, and snRNA-Seq. (**F)** Line plot of mean relative expression over developmental time points with 2 different gene groups (mCH-RNA coupled in blue; mCH-RNA uncoupled in orange). Relative expression level is defined as the log2(RPKM) minus mean log2(RPKM) over all time points for each gene. (**G**) Barplot of the number of protein coding genes in each of the 4 categories according to whether it's developmentally up- or down-regulated, and whether its mCH-RNA is coupled or not. (**H**) Left: Line plots of mean relative expression level over developmental time points for 5 gene bins. Genes are binned by gene expression ratio between early fetal (PCW 8-9) and adult (>2 yrs). Right: boxplot of TSS H3K27me3 signals at each of the 5 gene bins. (**I**) Scatter plot of Spearman correlation of gene body mCH and gene expression versus the mean H3K27me3 signal in neurons at gene body level. The H3K27me3 ChIP-seq data is from purified Glutamatergic and GABAergic neurons from human frontal cortex (Kozlenkov et al, 2018). (**J**) Genome browser track visualization of CBLN2 (mCH-RNA coupled) and CDC27 (uncoupled). (**K**) Gene level signal of CBLN2 and CDC27: scatter plot of normalized gene body mCH versus gene expression for all neuronal cells. Raw mCH level is normalized by the global mean mCH level of each cell.

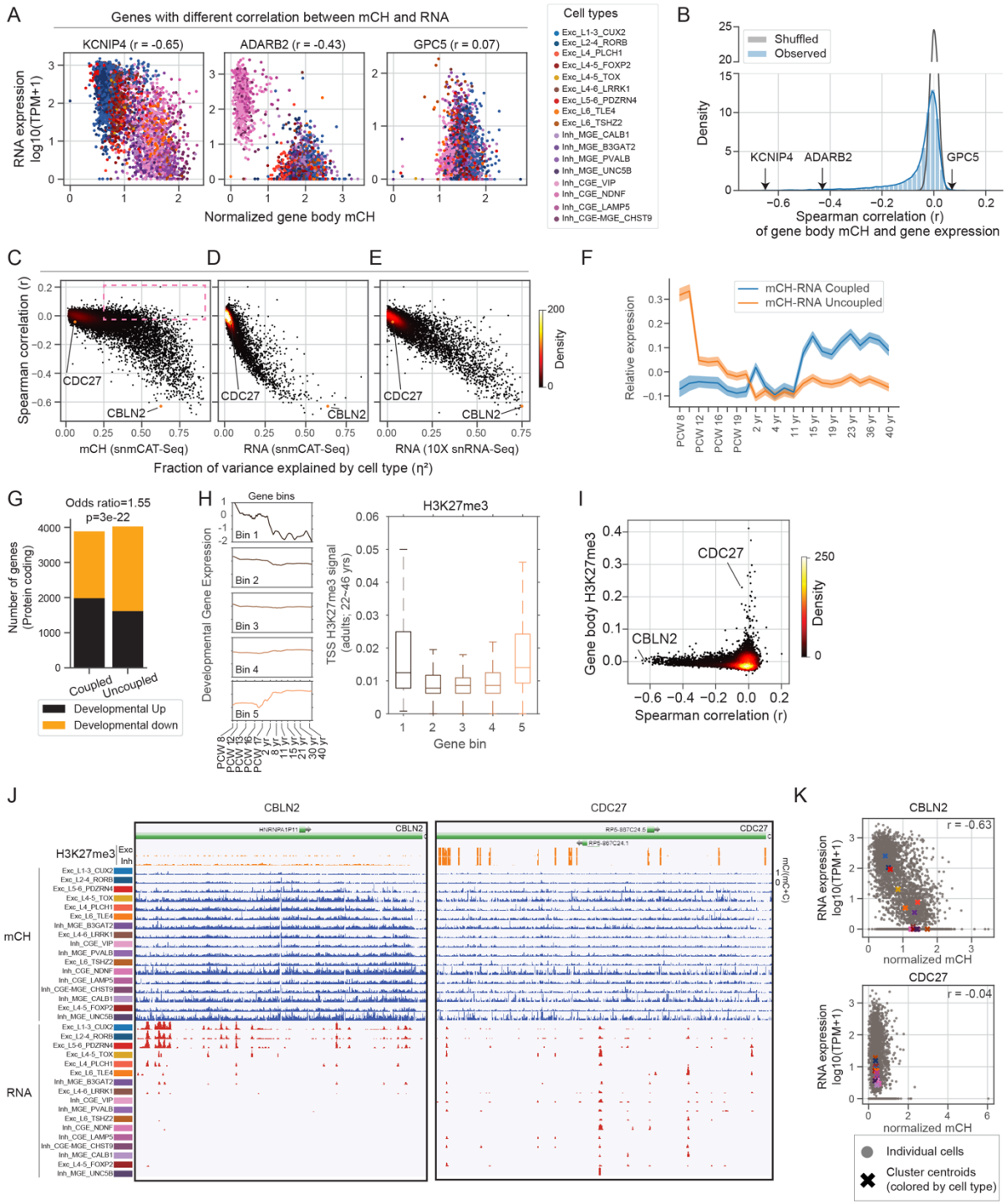
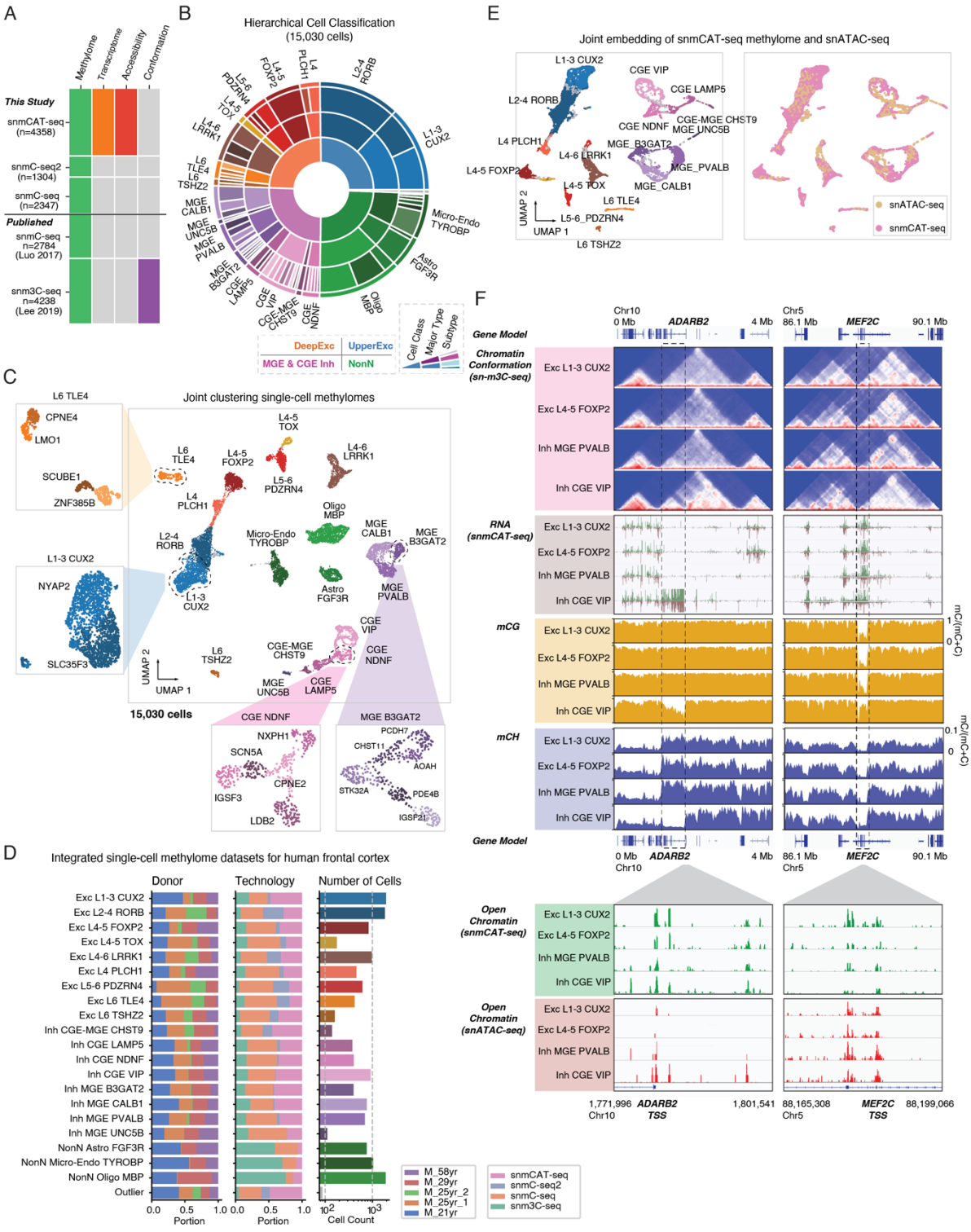

A
Genes with different correlation between mCH and RNA
KCNIP4 (r = -0.65)
ADARB2 (r = -0.43)
GPC5 (r = 0.07)
RNA expression log10(TPM+1)
Normalized gene body mCH
Cell types
Exc_L1-3_CUX2
Exc_L2-4_RORB
Exc_L4_PLCH1
Exc_L4-5_FOXP2
Exc_L4-5_TOX
Exc_L4-6_LRRK1
Exc_L5-6_PDZRN4
Exc_L6_TLE4
Exc_L6_TSHZ2
Inh_MGE_CALB1
Inh_MGE_B3GAT2
Inh_MGE_PVALB
Inh_MGE_UNC5B
Inh_CGE_VIP
Inh_CGE_NDNF
Inh_CGE_LAMP5
Inh_CGE-MGE_CHST9
B
Shuffled
Observed
Density
KCNIP4
ADARB2
GPC5
Spearman correlation (r)
of gene body mCH and gene expression
C
D
E
Spearman correlation (r)
CDC27
CBLN2
mCH (snmCAT-Seq)
RNA (snmCAT-Seq)
RNA (10X snRNA-Seq)
Density
Fraction of variance explained by cell type (η²)
F
mCH-RNA Coupled
mCH-RNA Uncoupled
Relative expression
PCW 8
PCW 12
PCW 16
PCW 19
2 yr
4 yr
11 yr
15 yr
19 yr
23 yr
36 yr
40 yr
G
Odds ratio=1.55
p=3e-22
Number of genes (Protein coding)
Coupled
Uncoupled
Developmental Up
Developmental down
H
Gene bins
Bin 1
Bin 2
Bin 3
Bin 4
Bin 5
Developmental Gene Expression
H3K27me3
TSS H3K27me3 signal (adults; 22-46 yrs)
Gene bin
I
Gene body H3K27me3
CDC27
CBLN2
Density
Spearman correlation (r)
J
CBLN2
CDC27
H3K27me3
Exc
Inh
mCH
RNA
K
CBLN2
r = -0.63
RNA expression log10(TPM+1)
normalized mCH
CDC27
r = -0.04
Individual cells
Cluster centroids (colored by cell type)

Figure 3.4. Integrated epigenomic atlas of the human frontal cortex.
(**A**) Methylome-based technologies and datasets included in the integrative analysis. (**B**) Sunburst visualization of the two-level methylome ensemble clustering analysis. The 4 cell classes (inmost ring) and 20 major cell types (middle ring and outer annotation) are identified in level 1 analysis, the 63 subtypes are identified in level 2 analysis. (**C**) UMAP embedding of 15,030 cells colored and labeled by major cell types from level 1 analysis. Several examples of level 2 analysis are shown in insets with UMAP colored and labeled by subtypes. (**D**) Donor (left) and technology (middle) composition and cell count (right) of each major cell type. (**E**) UMAP embedding of the cross-modality fusion of snmCAT-seq methylome and snATAC-seq profiles. The left panel is colored and labeled by level 1 major cell types; the right panel is color and labeled by the technologies. (**F**) Browser views of multi-modal data integration for *ADARB2* and *MEF2C* gene in four major cell types.


A
Methylome
Transcriptome
Accessibility
Conformation
This Study
snmCAT-seq (n=4358)
snmC-seq2 (n=1304)
snmC-seq (n=2347)
Published
snmC-seq n=2784 (Luo 2017)
snm3C-seq n=4238 (Lee 2019)
B
Hierarchical Cell Classification (15,030 cells)
L1-3 CUX2
L2-4 RORB
L4 PLCH1
L4-5 FOXP2
L5-6 PDZRN4
L4-5 TOX
L4-6 LRRK1
L6 TLE4
L6 TSHZ2
MGE CALB1
MGE UNC5B
MGE PVALB
MGE B3GAT2
CGE LAMP5
CGE VIP
CGE-MGE CHST9
CGE NDNF
Micro-Endo TYROBP
Astro FGF3R
Oligo MBP
DeepExc
UpperExc
MGE & CGE Inh
NonN
Cell Class
Major Type
Subtype
C
Joint clustering single-cell methylomes
L6 TLE4
CPNE4
LMO1
SCUBE1
ZNF385B
L1-3 CUX2
NYAP2
SLC35F3
L4-5 TOX
L6 TLE4
L4-5 FOXP2
L4-6 LRRK1
L5-6 PDZRN4
L4 PLCH1
Oligo MBP
L2-4 RORB
Micro-Endo TYROBP
MGE CALB1
MGE B3GAT2
Astro FGF3R
MGE PVALB
CGE VIP
CGE-MGE CHST9
CGE NDNF
L6 TSHZ2
MGE UNC5B
CGE LAMP5
UMAP 2
UMAP 1
15,030 cells
CGE NDNF
NXPH1
SCN5A
CPNE2
IGSF3
LDB2
MGE B3GAT2
PCDH7
CHST11
AOAH
STK32A
PDE4B
IGSF21
D
Integrated single-cell methylome datasets for human frontal cortex
Donor
Technology
Number of Cells
Exc L1-3 CUX2
Exc L2-4 RORB
Exc L4-5 FOXP2
Exc L4-5 TOX
Exc L4-6 LRRK1
Exc L4 PLCH1
Exc L5-6 PDZRN4
Exc L6 TLE4
Exc L6 TSHZ2
Inh CGE-MGE CHST9
Inh CGE LAMP5
Inh CGE NDNF
Inh CGE VIP
Inh MGE B3GAT2
Inh MGE CALB1
Inh MGE PVALB
Inh MGE UNC5B
NonN Astro FGF3R
NonN Micro-Endo TYROBP
NonN Oligo MBP
Outlier
0.0 0.5 1.0
Portion
Cell Count
M_58yr
M_29yr
M_25yr_2
M_25yr_1
M_21yr
snmCAT-seq
snmC-seq2
snmC-seq
snm3C-seq
E
Joint embedding of snmCAT-seq methylome and snATAC-seq
L1-3 CUX2
CGE VIP
CGE LAMP5
L2-4 RORB
CGE NDNF
CGE-MGE CHST9
MGE UNC5B
L4 PLCH1
MGE B3GAT2
L4-6 LRRK1
MGE PVALB
L4-5 FOXP2
MGE CALB1
L4-5 TOX
L6 TLE4
L5-6 PDZRN4
L6 TSHZ2
UMAP 2
UMAP 1
snATAC-seq
snmCAT-seq
F
Gene Model
Chr10
0 Mb
ADARB2
4 Mb
Chr5
86.1 Mb
MEF2C
90.1 Mb
Chromatin Conformation (sn-m3C-seq)
RNA (snmCAT-seq)
mCG
mCH
Exc L1-3 CUX2
Exc L4-5 FOXP2
Inh MGE PVALB
Inh CGE VIP
mC/(mC+C)
Open Chromatin (snmCAT-seq)
Open Chromatin (snATAC-seq)
1,771,996
ADARB2
TSS
1,801,541
88,165,308
MEF2C
TSS
88,199,066

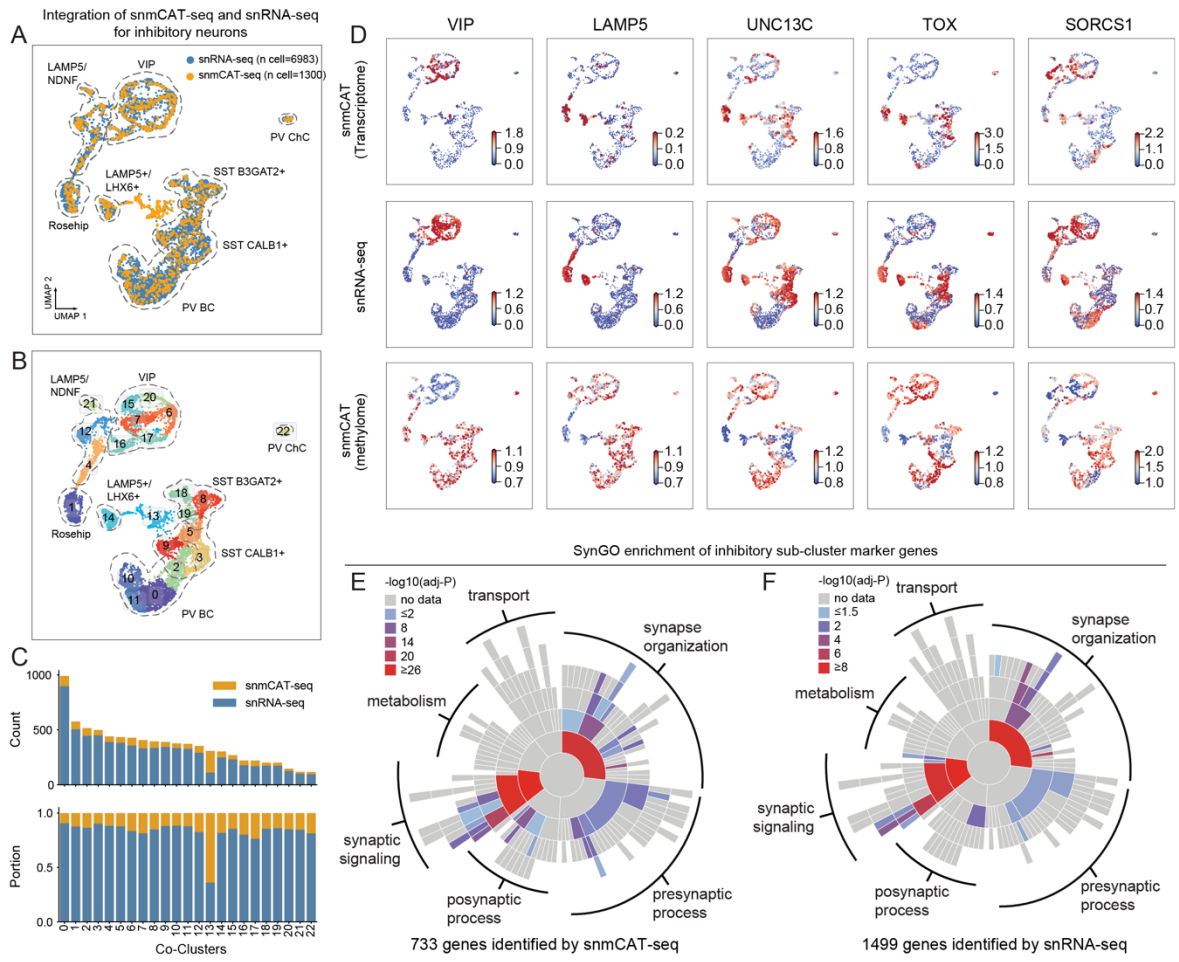
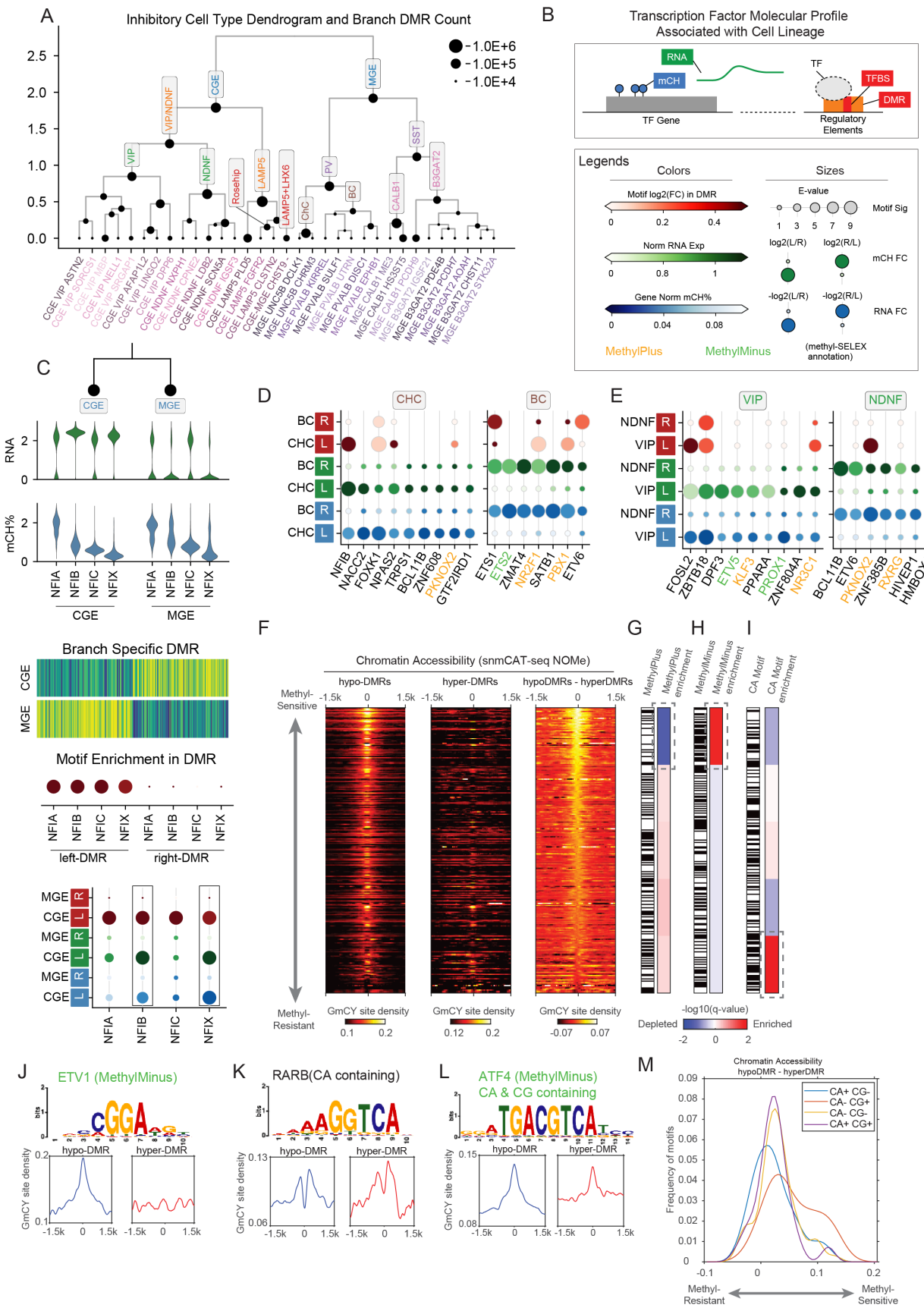


Figure 3.5. snmCAT-seq identifies RNA and mC signatures of neuronal subtypes.
(**A-B**) UMAP embedding of snmCAT-seq transcriptome and snRNA-seq for all the inhibitory neurons after MNN-based integration with the cells colored by technology (A) and joint clusters (B). (**C**) The composition of cells profiled by snmCAT-seq and snRNA-seq in inhibitory neurons joint clusters (same cluster IDs as shown in (B)). The upper and lower bar plots show the counts and portion of cells profiled by the two technologies in each joint cluster, respectively. (**D**) Normalized expression and gene body mCH rate of inhibitory neuron subtype marker genes quantified using snmCAT-seq and snRNA-seq. (**E-F**) Sunburst visualization of inhibitory cell type marker genes enrichment in SynGO biological process terms. Each sector is a SynGO term colored by -log10(adjusted P-value) of snmCAT-seq transcriptome marker genes (E) or snRNA-seq marker genes (F) enrichment.

Figure 3.6. DMR phylogeny and transcription factor hierarchy in the human cortex.
(**A**) Inhibitory neuron subtype dendrogram. The node size represents the number of DMRs detected between the left and right branches. Nodes corresponding to known inhibitory cell type groups are annotated in the dendrogram. (**B**) Schematics of the three levels of molecular information we use to identify candidate TF related to the specific lineage. (**C**) The workflow of TF analysis using the NFI family as an example. Three types of information are gathered for each of the TF genes: 1. RNA expression; 2. Gene body mCH level; 3. TF motif enrichment in the branch-specific DMR. We create a combined dot plot view for all three kinds of information, the genes show lineage specificity in both 1 and 2 are circled by black boxes. (**D-E**) The dot plot view for TFs showing ChC vs. BC (D) or VIP vs. NDNF (E) specificity in motif enrichment, RNA or mCH levels. Colors for every two rows from top to bottom: TF motif enrichment log2(fold change), branch mean expression log(1 + CPM), lineage mean gene body mCH level. Sizes for every two rows from top to bottom: E-value of the motif enrichment test, relative fold change of expression level, relative fold change of mCH level between the two branches. Colors for the motif names: TF motif methylation preference annotated by methyl-SELEX experiment [6], orange indicate MethylPlus, green indicate MethylMinus. (**F**) The binding of TFs to hypermethylated regions validated by chromatin accessibility measurement using the snmCAT-seq NOMe-seq profile. (**G-I**) Enrichment or depletion of MethylPlus TFs (G), MethyMinus TFs (H) and TFs whose binding motif contains CA dinucleotides (I). (**J-L**) Examples of chromatin accessibility profiles at the binding motifs of ETV1 (MethylMinus) (J), RARB (motif contains CA) (K) and ATF4 (motif contains CA and CG) (L). (**M**) Comparison of the chromatin accessibility at the binding motifs containing CA or CG dinucleotides.


A
Inhibitory Cell Type Dendrogram and Branch DMR Count
B
Transcription Factor Molecular Profile
Associated with Cell Lineage
RNA
mCH
TF
TFBS
DMR
TF Gene
Regulatory
Elements
Legends
Colors
Sizes
Motif log2(FC) in DMR
E-value
Motif Sig
Norm RNA Exp
log2(L/R)
log2(R/L)
mCH FC
Gene Norm mCH%
-log2(L/R)
-log2(R/L)
RNA FC
MethylPlus
MethylMinus
(methyl-SELEX
annotation)
C
CGE
MGE
RNA
mCH%
NFIA
NFIB
NFIC
NFIX
D
CHC
BC
E
VIP
NDNF
F
Chromatin Accessibility (snmCAT-seq NOMe)
hypo-DMRs
hyper-DMRs
hypoDMRs - hyperDMRs
Methyl-
Sensitive
Methyl-
Resistant
GmCY site density
G
H
I
MethylPlus
MethylPlus enrichment
MethylMinus
MethylMinus enrichment
CA Motif
CA Motif enrichment
-log10(q-value)
Depleted
Enriched
Branch Specific DMR
Motif Enrichment in DMR
left-DMR
right-DMR
J
ETV1 (MethylMinus)
K
RARB(CA containing)
L
ATF4 (MethylMinus)
CA & CG containing
M
Chromatin Accessibility
hypoDMR - hyperDMR
CA+ CG-
CA- CG+
CA- CG-
CA+ CG+
Frequency of motifs

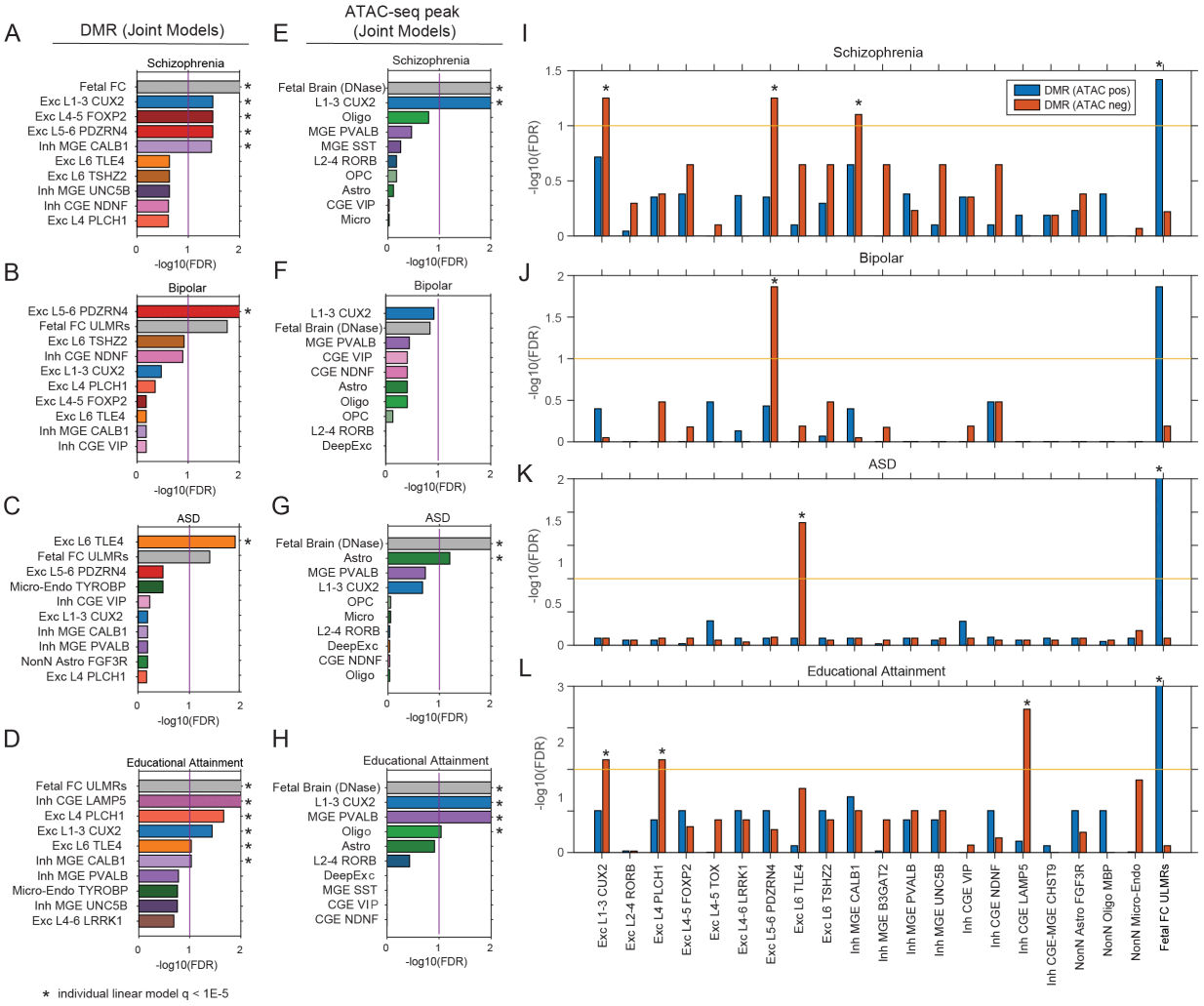
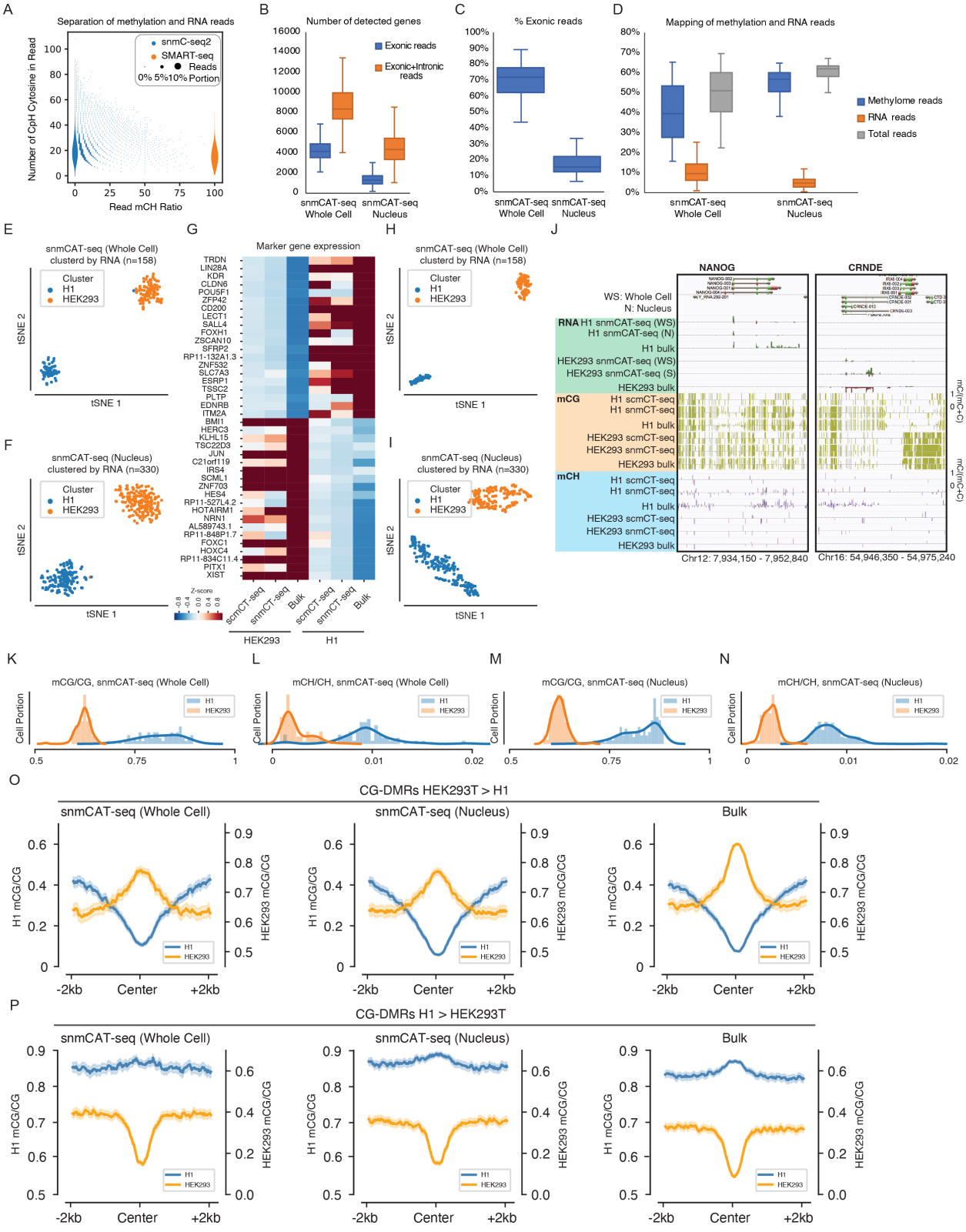


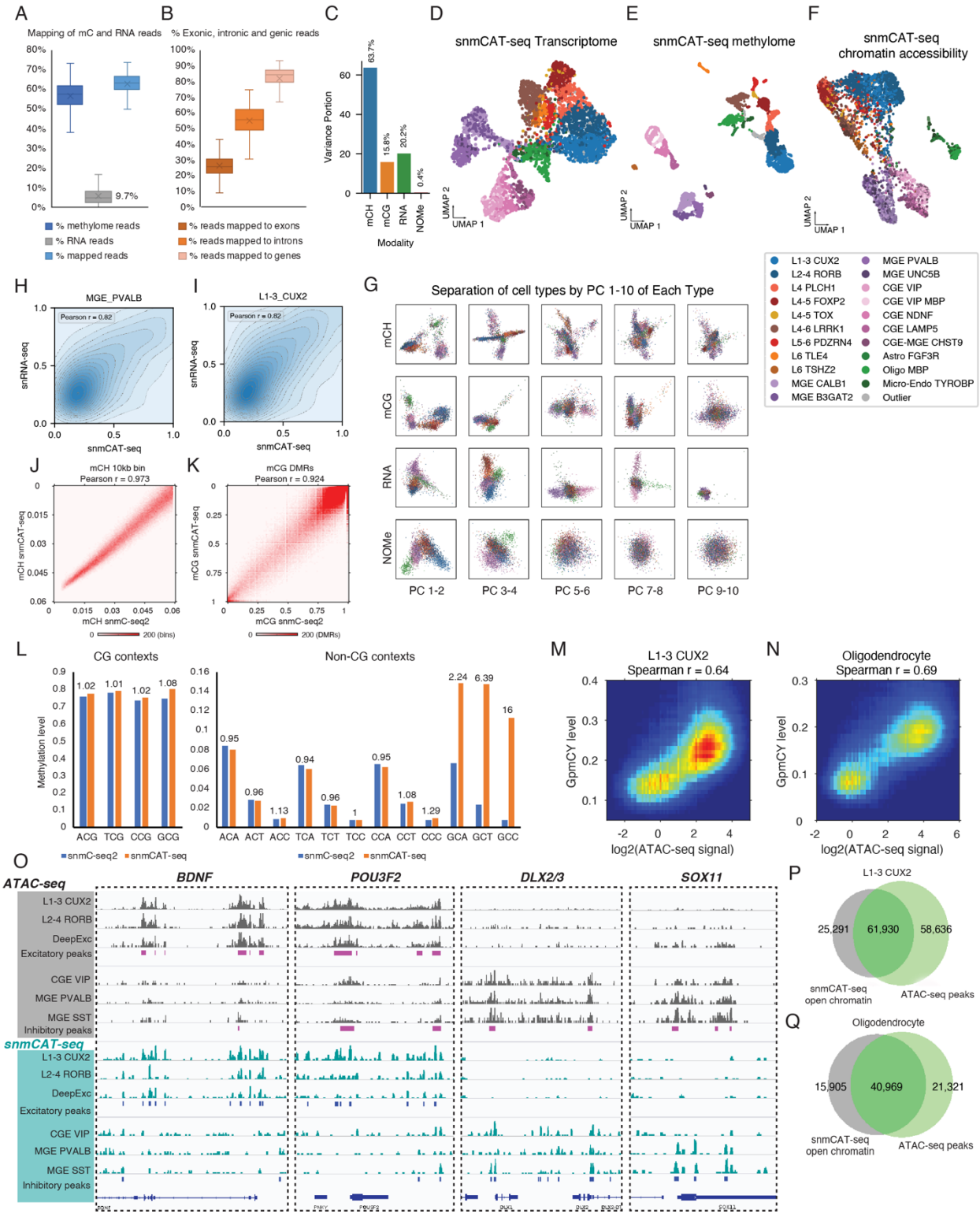
Figure 3.7. Identification of brain cell types involved in neuropsychiatric traits.
(**A-H**) Multiple regression partitioned heritability analysis using cell-type-specific DMRs (A-D) or ATAC-seq peaks (E-H). (**I-L**) Multiple regression partitioned heritability analysis using DMRs stratified for the overlap with open chromatin regions. Heritability enrichment with a p-value < 1E-5 compared to the baseline was indicated by asterisks.

3.20 Supplementary Figures

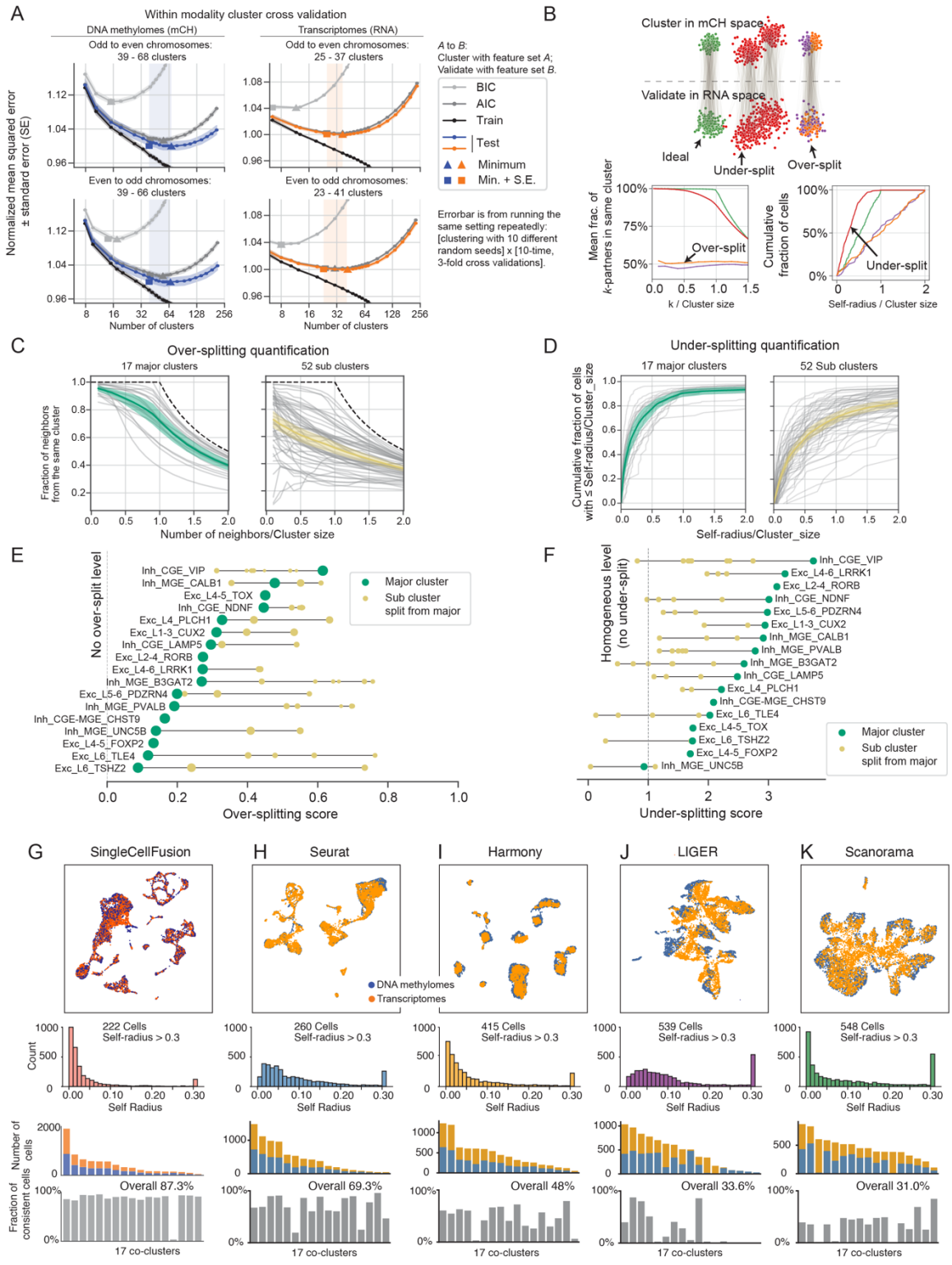
Supplementary Figure 3.1 snmCAT-seq captures transcriptome and DNA methylation signatures of H1 & HEK293 cells.
(A) The specificity for classifying methylation (snmC-seq2) and transcriptome (snRNA-seq) reads plotted as a function of the number of CpH cytosine in the reads. (B-D) The number of detected genes (B), percentage of mapped reads that are located in exons (C) and mapping rates of methylation and RNA reads (D) for snmCAT-seq (Whole Cell) and snmCAT-seq (Nucleus). (E-F) Separation of H1 and HEK293T cells by tSNE using transcriptome reads extracted from snmCAT-seq (Whole Cell) (E) or snmCAT-seq (Nucleus) (F) datasets. (G) snmCAT-seq (Whole Cell) and snmCAT-seq (Nucleus) detect genes specifically expressed in H1 or HEK293T cells. (H-I) Separation of H1 and HEK293T cells by tSNE using DNA methylation information extracted from snmCAT-seq (Whole Cell) (H) or snmCAT-seq (Nucleus) (I) datasets. (J) Browser view of NANOG and CRNDE loci. (K-N) Distribution of mCG and mCH levels for single H1 and HEK293 cells/nuclei as determined by snmCAT-seq (Whole Cell) and snmCAT-seq (Nucleus). (O-P) snmCAT-seq (Whole Cell) and snmCAT-seq (Nucleus) recapitulate bulk mCG patterns at CG-DMRs showing greater mCG levels in HEK293T (O) or H1 (P) cells.


A
Separation of methylation and RNA reads
snmC-seq2
SMART-seq
Reads
0% 5%10% Portion
Number of CpH Cytosine in Read
Read mCH Ratio
B
Number of detected genes
Exonic reads
Exonic+Intronic reads
snmCAT-seq Whole Cell
snmCAT-seq Nucleus
C
% Exonic reads
D
Mapping of methylation and RNA reads
Methylome reads
RNA reads
Total reads
E
snmCAT-seq (Whole Cell) clustered by RNA (n=158)
Cluster
H1
HEK293
tSNE 1
tSNE 2
F
snmCAT-seq (Nucleus) clustered by RNA (n=330)
G
Marker gene expression
TRDN
LIN28A
KDR
CLDN6
POU5F1
ZFP42
CD200
LECT1
SALL4
FOXH1
ZSCAN10
SFRP2
RP11-132A1.3
ZNF532
SLC7A3
ESRP1
TSSC2
PLTP
EDNRB
ITM2A
EMI1
HERC3
KLHL15
TSC22D3
JUN
C21orf119
IRS4
SCML1
ZNF703
HES4
RP11-527L4.2
HOTAIRM1
NRN1
AL589743.1
RP11-848P1.7
FOXC1
HOXC4
RP11-834C11.4
PITX1
XIST
Z-score
scmCT-seq
snmCT-seq
Bulk
HEK293
H1
H
I
J
NANOG
CRNDE
WS: Whole Cell
N: Nucleus
RNA H1 snmCAT-seq (WS)
H1 snmCAT-seq (N)
H1 bulk
HEK293 snmCAT-seq (WS)
HEK293 snmCAT-seq (S)
HEK293 bulk
mCG
H1 scmCT-seq
H1 snmCT-seq
HEK293 scmCT-seq
HEK293 snmCT-seq
mCH
mC/(mC+C)
Chr12: 7,934,150 - 7,952,840
Chr16: 54,946,350 - 54,975,240
K
mCG/CG, snmCAT-seq (Whole Cell)
L
mCH/CH, snmCAT-seq (Whole Cell)
M
mCG/CG, snmCAT-seq (Nucleus)
N
mCH/CH, snmCAT-seq (Nucleus)
Cell Portion
O
CG-DMRs HEK293T > H1
snmCAT-seq (Whole Cell)
snmCAT-seq (Nucleus)
Bulk
H1 mCG/CG
HEK293 mCG/CG
-2kb
Center
+2kb
P
CG-DMRs H1 > HEK293T

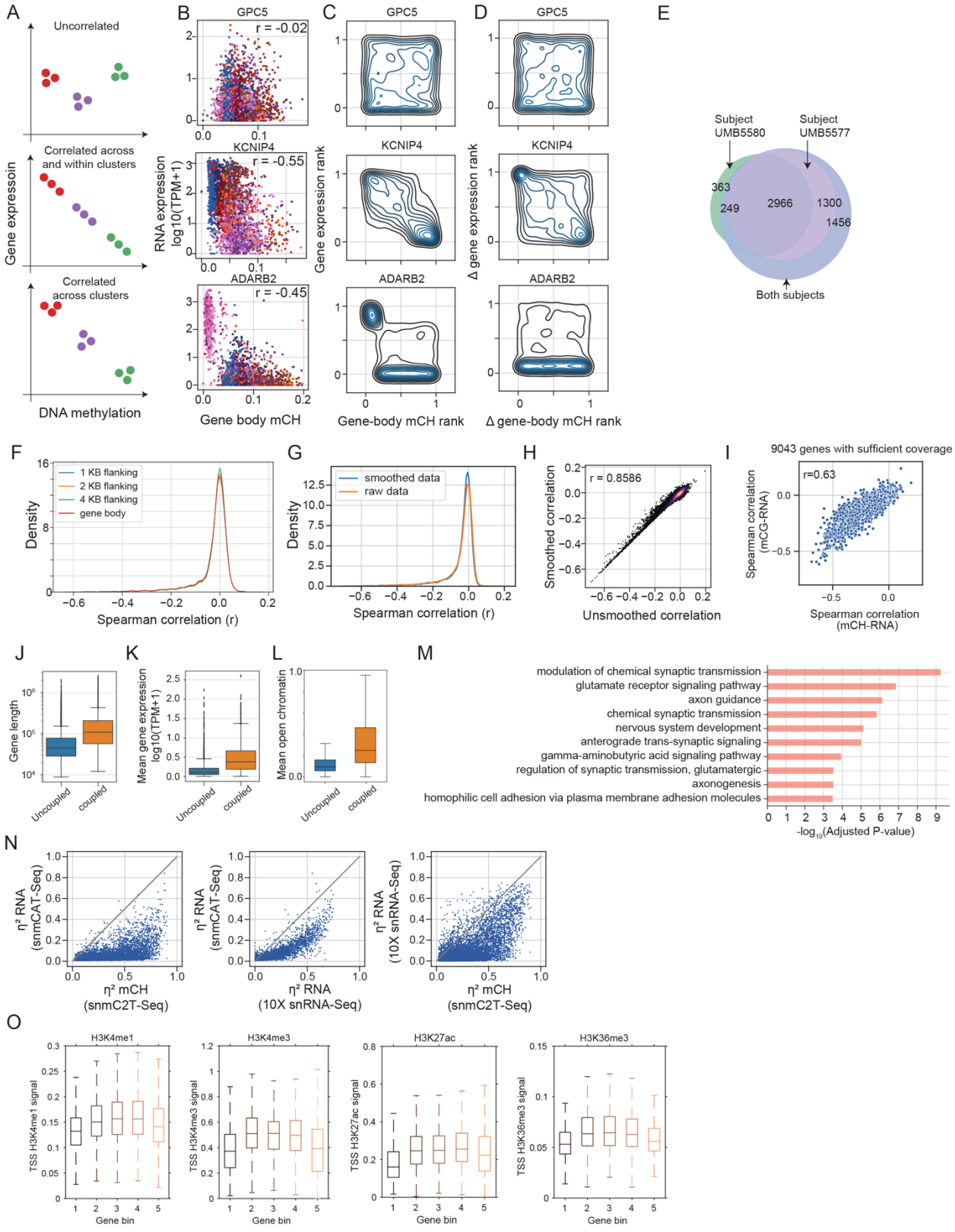
Supplementary Figure 3.2 snmCAT-seq generates single-nucleus multi-omic profiles from human brain tissues.
(A) The fraction of total snmCAT-seq reads derived from methylome and transcriptome. (B) The fraction of snmCAT-seq transcriptome reads mapped to exons, introns or gene bodies. (C) The portion of variance explained by each data modality. (D-F) UMAP embedding of 4253 snmCAT-seq cells using single modality information: transcriptome (D), methylome (mCH and mCG, E) and chromatin accessibility (F). (G) The separation of cell types by the top 10 principal components (PCs) of each data type. (H-I) Pearson correlation of gene expression quantified by snmCAT-seq transcriptome and snRNA-seq in MGE PVALB (H) and L1-3 CUX2 (I) cells. (J) Pearson correlation of gene body non-CG methylation quantified with snmCAT-seq methylome and snmC-seq for MGE PVALB cells. (K) Pearson correlation of CG methylation at DMRs quantified with snmCAT-seq methylome and snmC-seq for MGE PVALB cells. (L) Genome-wide methylation level for all tri-nucleotide context (-1 to +2 position) surrounding cytosines shows the sequence specificity of GpC methyltransferase M.CviPI. (M-N) Spearman correlation between GCY methylation level and ATAC-seq signal at open chromatin sites in L1-3 CUX2 (M) and Oligodendrocyte (N) cells. (Q) Browser views of chromatin accessibility profiles generated by snmCAT-seq and snATAC-seq at cell-type-specific genes. (P-Q) Overlap of open chromatin peaks identified by snmCAT-seq and snATAC-seq in L1-3 CUX2 (P) and oligodendrocyte (Q) cells.


A
Mapping of mC and RNA reads
% methylome reads
% RNA reads
% mapped reads
9.7%
B
% Exonic, intronic and genic reads
% reads mapped to exons
% reads mapped to introns
% reads mapped to genes
C
Variance Portion
63.7%
15.8%
20.2%
0.4%
mCH
mCG
RNA
NOMe
Modality
D
snmCAT-seq Transcriptome
UMAP 1
UMAP 2
E
snmCAT-seq methylome
F
snmCAT-seq
chromatin accessibility
L1-3 CUX2
L2-4 RORB
L4 PLCH1
L4-5 FOXP2
L4-5 TOX
L4-6 LRRK1
L5-6 PDZRN4
L6 TLE4
L6 TSHZ2
MGE CALB1
MGE B3GAT2
MGE PVALB
MGE UNC5B
CGE VIP
CGE VIP MBP
CGE NDNF
CGE LAMP5
CGE-MGE CHST9
Astro FGF3R
Oligo MBP
Micro-Endo TYROBP
Outlier
H
MGE_PVALB
Pearson r = 0.82
snRNA-seq
snmCAT-seq
I
L1-3_CUX2
Pearson r = 0.82
G
Separation of cell types by PC 1-10 of Each Type
mCH
mCG
RNA
NOMe
PC 1-2
PC 3-4
PC 5-6
PC 7-8
PC 9-10
J
mCH 10kb bin
Pearson r = 0.973
mCH snmCAT-seq
mCH snmC-seq2
200 (bins)
K
mCG DMRs
Pearson r = 0.924
mCG snmCAT-seq
mCG snmC-seq2
200 (DMRs)
L
CG contexts
Non-CG contexts
Methylation level
1.02
1.01
1.02
1.08
ACG
TCG
CCG
GCG
0.95
0.96
1.13
0.94
0.96
1
0.95
1.08
1.29
2.24
6.39
16
ACA
ACT
ACC
TCA
TCT
TCC
CCA
CCT
CCC
GCA
GCT
GCC
snmC-seq2
snmCAT-seq
M
L1-3 CUX2
Spearman r = 0.64
GpmCY level
log2(ATAC-seq signal)
N
Oligodendrocyte
Spearman r = 0.69
GpmCY level
log2(ATAC-seq signal)
O
ATAC-seq
BDNF
POU3F2
DLX2/3
SOX11
L1-3 CUX2
L2-4 RORB
DeepExc
Excitatory peaks
CGE VIP
MGE PVALB
MGE SST
Inhibitory peaks
snmCAT-seq
L1-3 CUX2
L2-4 RORB
DeepExc
Excitatory peaks
CGE VIP
MGE PVALB
MGE SST
Inhibitory peaks
P
L1-3 CUX2
25,291
61,930
58,636
snmCAT-seq
open chromatin
ATAC-seq peaks
Q
Oligodendrocyte
15,905
40,969
21,321
snmCAT-seq
open chromatin
ATAC-seq peaks

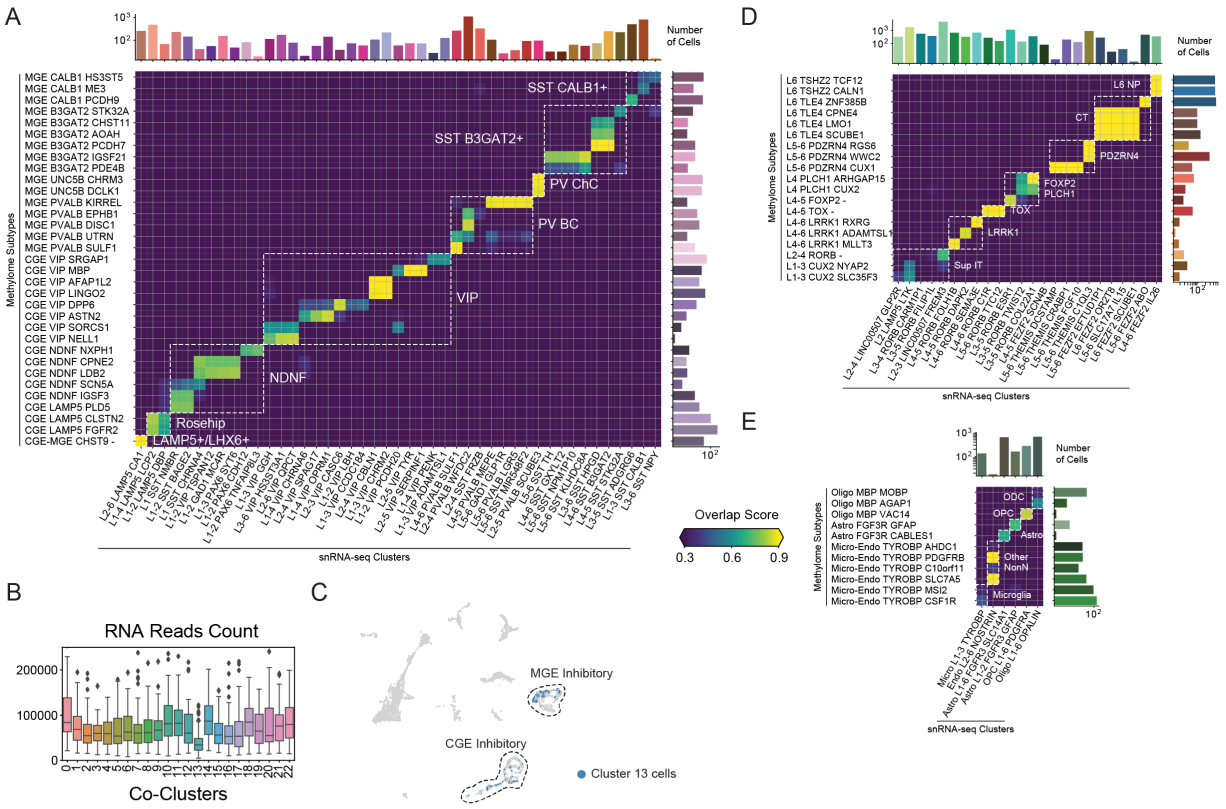
Supplementary Figure 3.3 Evaluation of cluster quality with paired transcriptome and methylome profiles.
(A) Intra-modality cross-validation of mCH- or RNA- defined clusters. Line plots show mean squared error between the single-cell profiles and cluster centroid, Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as a function of the number of clusters. The shaded region in each sub-plot highlights the range between the minimum and the minimum + standard error for the curve of test-set error. For the analysis using snmCAT-seq mC information (left panels), gene body mCH profiles of odd (even) chromosomes were used for clustering whereas even (odd) chromosomes were used for testing. A similar analysis was performed using snmCAT-seq transcriptome information (right panels). (B) Schematic diagram of the over- and under-splitting analysis using matched single-cell methylome and transcriptome profiles, complementing Fig. 3.2D. (C) The over-splitting quantification of mC-defined major clusters (n=17) and subclusters (n=52) was quantified by the fraction of cross-modal neighbors found in the same cluster defined by RNA. (D) The under-splitting of clusters was quantified as the cumulative distribution function of normalized self-radius for mC-define major clusters and subclusters. For (C-D), gray lines represent individual clusters while colored lines represent means and confidence intervals. (E) Over-splitting score for each major cluster (in green) and associated sub-clusters (in yellow). Dot size of sub-clusters represents cluster size normalized by the size of their "mother" major cluster. (F) Under-splitting score for each major cluster (in green) and associated sub-clusters (in yellow). (G-K) Fusion of snmCAT-seq transcriptome and mC profiles using the Single Cell Fusion (G), Seurat (H), Harmony (I), LIGER (J), and Scanorama (K). For each computational data fusion method, from top to bottom the first panel shows mC and RNA modalities on the joint UMAP embedding after data fusion. The second panel shows the normalized self-radius. The third and fourth panels show the co-cluster level cell composition and clustering accuracy.


A
Within modality cluster cross validation
DNA methylomes (mCH)
Transcriptomes (RNA)
Odd to even chromosomes: 39 - 68 clusters
Odd to even chromosomes: 25 - 37 clusters
Even to odd chromosomes: 39 - 66 clusters
Even to odd chromosomes: 23 - 41 clusters
Normalized mean squared error ± standard error (SE)
Number of clusters
A to B: Cluster with feature set A; Validate with feature set B.
BIC
AIC
Train
Test
Minimum
Min. + S.E.
Errorbar is from running the same setting repeatedly: [clustering with 10 different random seeds] x [10-time, 3-fold cross validations].
B
Cluster in mCH space
Validate in RNA space
Ideal
Under-split
Over-split
Mean frac. of k-partners in same cluster
k / Cluster size
Cumulative fraction of cells
Self-radius / Cluster size
C
Over-splitting quantification
17 major clusters
52 sub clusters
Fraction of neighbors from the same cluster
Number of neighbors/Cluster size
D
Under-splitting quantification
17 major clusters
52 Sub clusters
Cumulative fraction of cells with ≤ Self-radius/Cluster_size
Self-radius/Cluster_size
E
No over-split level
Inh_CGE_VIP
Inh_MGE_CALB1
Exc_L4-5_TOX
Inh_CGE_NDNF
Exc_L4_PLCH1
Exc_L1-3_CUX2
Inh_CGE_LAMP5
Exc_L2-4_RORB
Exc_L4-6_LRRK1
Inh_MGE_B3GAT2
Exc_L5-6_PDZRN4
Inh_MGE_PVALB
Inh_CGE-MGE_CHST9
Inh_MGE_UNC5B
Exc_L4-5_FOXP2
Exc_L6_TLE4
Exc_L6_TSHZ2
Major cluster
Sub cluster split from major
Over-splitting score
F
Homogeneous level (no under-split)
Inh_CGE_VIP
Exc_L4-6_LRRK1
Exc_L2-4_RORB
Inh_CGE_NDNF
Exc_L5-6_PDZRN4
Exc_L1-3_CUX2
Inh_MGE_CALB1
Inh_MGE_PVALB
Inh_MGE_B3GAT2
Inh_CGE_LAMP5
Exc_L4_PLCH1
Inh_CGE-MGE_CHST9
Exc_L6_TLE4
Exc_L4-5_TOX
Exc_L6_TSHZ2
Exc_L4-5_FOXP2
Inh_MGE_UNC5B
Major cluster
Sub cluster split from major
Under-splitting score
G
SingleCellFusion
H
Seurat
I
Harmony
J
LIGER
K
Scanorama
DNA methylomes
Transcriptomes
222 Cells Self-radius > 0.3
260 Cells Self-radius > 0.3
415 Cells Self-radius > 0.3
539 Cells Self-radius > 0.3
548 Cells Self-radius > 0.3
Count
Self Radius
Number of cells
Fraction of consistent cells
Overall 87.3%
Overall 69.3%
Overall 48%
Overall 33.6%
Overall 31.0%
17 co-clusters

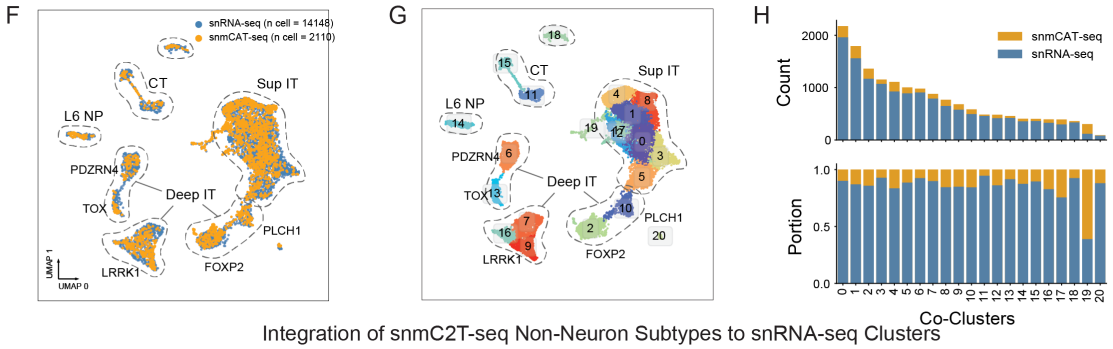
Supplementary Figure 3.4 Diverse correlations between gene expression and gene body mCH. (A) Schematic diagram of 3 types of genes with different correlations between gene expression and DNA methylation. (B) Scatter plots of gene body mCH (unnormalized) and gene expression of example genes (KCNIP4, ADARB2, GPC5) across all neuronal cells. Cells are colored by major cell types defined in Fig. 3.4. (C) Contour density plot of gene body mCH rank versus gene expression rank for 3 example genes: GPC5, KCNIP4, and ADARB2. (D) Contour density plot of delta gene body mCH rank (rank of gene body mCH - cluster mean gene body mCH) versus delta gene expression rank (rank of gene expression - cluster mean gene expression) for 3 example genes: GPC5, KCNIP4, and ADARB2. (E) Venn diagram showing a strong overlap of mCH-RNA coupled genes identified using cells from subject UMB5580, subject UMB5577, and from both subjects. (F) Distribution of Spearman correlation coefficient between gene expression and gene-level mCH quantified at gene body (red), gene body + 1 kilo-base upstream (blue), + 2 kilo-base upstream (orange) and + 4 kilo-base upstream (green). (G) Distribution of Spearman correlation coefficient between gene expression and gene-level mCH quantified at gene body with or without data smoothing before correlation analysis. (H) Scatter plot comparing the effect of data smoothing on Spearman correlation coefficients computed between gene expression and gene-level mCH. (I) Scatter plot comparing gene body mCH-RNA correlation versus gene body mCG-RNA correlation. (J-L) Boxplot of gene length (J), mean gene expression level (K) and mean open chromatin level (L) for mCH-RNA uncoupled and coupled genes. (M) Gene ontology enrichment of mCH-RNA coupled genes. (N) Scatter plots comparing the fraction of variance explained by cell type (η^2) for each gene from different datasets or data modalities: RNA (from snmCAT-seq), mCH (from snmCAT-seq) and 10X (snRNA-Seq from 10X protocols). (O) The distribution of different histone marks at TSS over 5 gene bins grouped according to gene expression ratio of early fetal (PCW 8-9) to adult (>2 yrs).


A
Uncorrelated
Correlated across and within clusters
Correlated across clusters
Gene expressoin
DNA methylation
B
GPC5
r = -0.02
KCNIP4
r = -0.55
ADARB2
r = -0.45
RNA expression log10(TPM+1)
Gene body mCH
C
GPC5
KCNIP4
ADARB2
Gene expression rank
Gene-body mCH rank
D
GPC5
KCNIP4
ADARB2
Δ gene expression rank
Δ gene-body mCH rank
E
Subject UMB5580
Subject UMB5577
363
249
2966
1300
1456
Both subjects
F
1 KB flanking
2 KB flanking
4 KB flanking
gene body
Density
Spearman correlation (r)
G
smoothed data
raw data
Density
Spearman correlation (r)
H
r = 0.8586
Smoothed correlation
Unsmoothed correlation
I
9043 genes with sufficient coverage
r=0.63
Spearman correlation (mCG-RNA)
Spearman correlation (mCH-RNA)
J
Gene length
Uncoupled
coupled
K
Mean gene expression log10(TPM+1)
Uncoupled
coupled
L
Mean open chromatin
Uncoupled
coupled
M
modulation of chemical synaptic transmission
glutamate receptor signaling pathway
axon guidance
chemical synaptic transmission
nervous system development
anterograde trans-synaptic signaling
gamma-aminobutyric acid signaling pathway
regulation of synaptic transmission, glutamatergic
axonogenesis
homophilic cell adhesion via plasma membrane adhesion molecules
-log10(Adjusted P-value)
N
η² RNA (snmCAT-Seq)
η² mCH (snmC2T-Seq)
η² RNA (10X snRNA-Seq)
η² RNA (snmCAT-Seq)
η² RNA (10X snRNA-Seq)
η² mCH (snmC2T-Seq)
O
H3K4me1
H3K4me3
H3K27ac
H3K36me3
TSS H3K4me1 signal
TSS H3K4me3 signal
TSS H3K27ac signal
TSS H3K36me3 signal
Gene bin

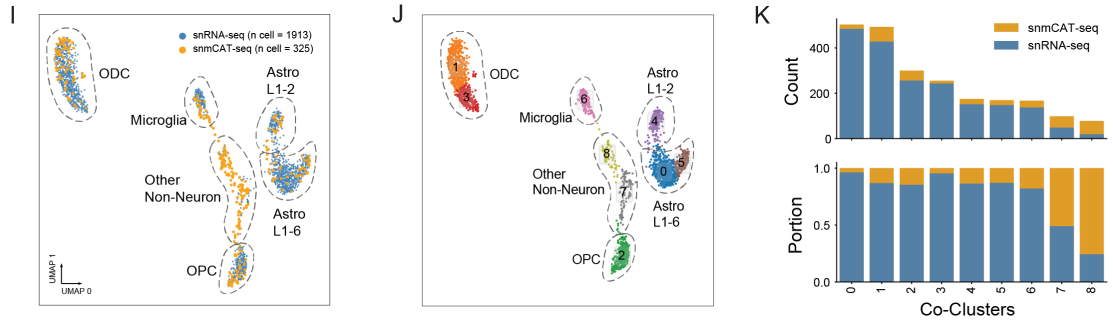
Supplementary Figure 3.5 snmCAT-seq recapitulates transcriptome and methylome signatures of neuronal subtypes.
(A) Confusion matrix showing the overlap scores between inhibitory subtypes identified by ensemble methylome analysis and snRNA-seq. Known inhibitory cell type groups are annotated by boxes. The upper bar plot indicates the snRNA-seq cell counts per cluster; the right bar plot indicates the snmCAT-seq cell counts per cluster. (B) Cluster 13 in Fig. 3.5A-C includes snmCAT-seq transcriptome profiles with lower RNA read counts. (C) snmCAT-seq methylome profiles of single nuclei in cluster 13 (colored in blue) can be readily integrated with other inhibitory cells. (D-E) Confusion matrix showing the overlap scores between methylome ensemble subtypes and the snRNA-seq clusters for excitatory neuron clusters (D) and non-neuron clusters (E). Known cell type groups are annotated by boxes. The upper bar plot annotates the snRNA-seq cell counts per cluster, the right bar plot annotates the snmCAT-seq cell counts per cluster. (F-G) UMAP embedding of all excitatory neurons profiled by snmCAT-seq and snRNA-seq after MNN-based integration, colored by technology (F) and joint clusters (G). Known cluster groups are also circled and annotated on UMAP. (H) The composition of cells profiled by snmCAT-seq and snATAC-seq in excitatory neuron joint clusters. The upper and lower bar plots show the counts and portion of cells profiled by the two technologies in each joint cluster, respectively. (I-J) UMAP embedding of all non-neuronal cells from snmCAT-seq and snRNA-seq after integration, colored by technology (I) and joint clusters (J). (K) The composition of cells profiled by snmCAT-seq and snATAC-seq in non-neuronal cell joint clusters.



Integration of snmCAT-seq Excitatory Subtypes to snRNA-seq Clusters

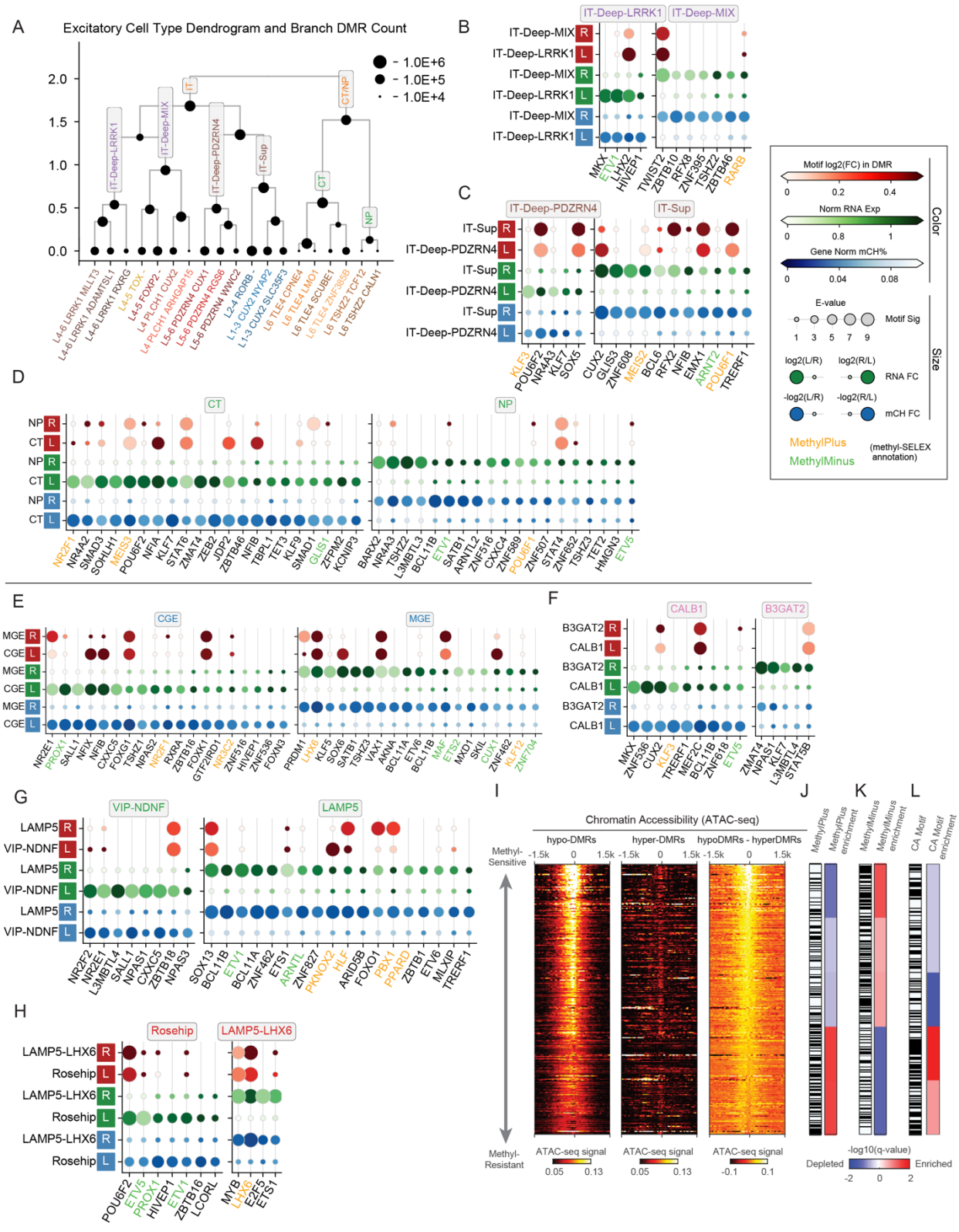


Integration of snmC2T-seq Non-Neuron Subtypes to snRNA-seq Clusters

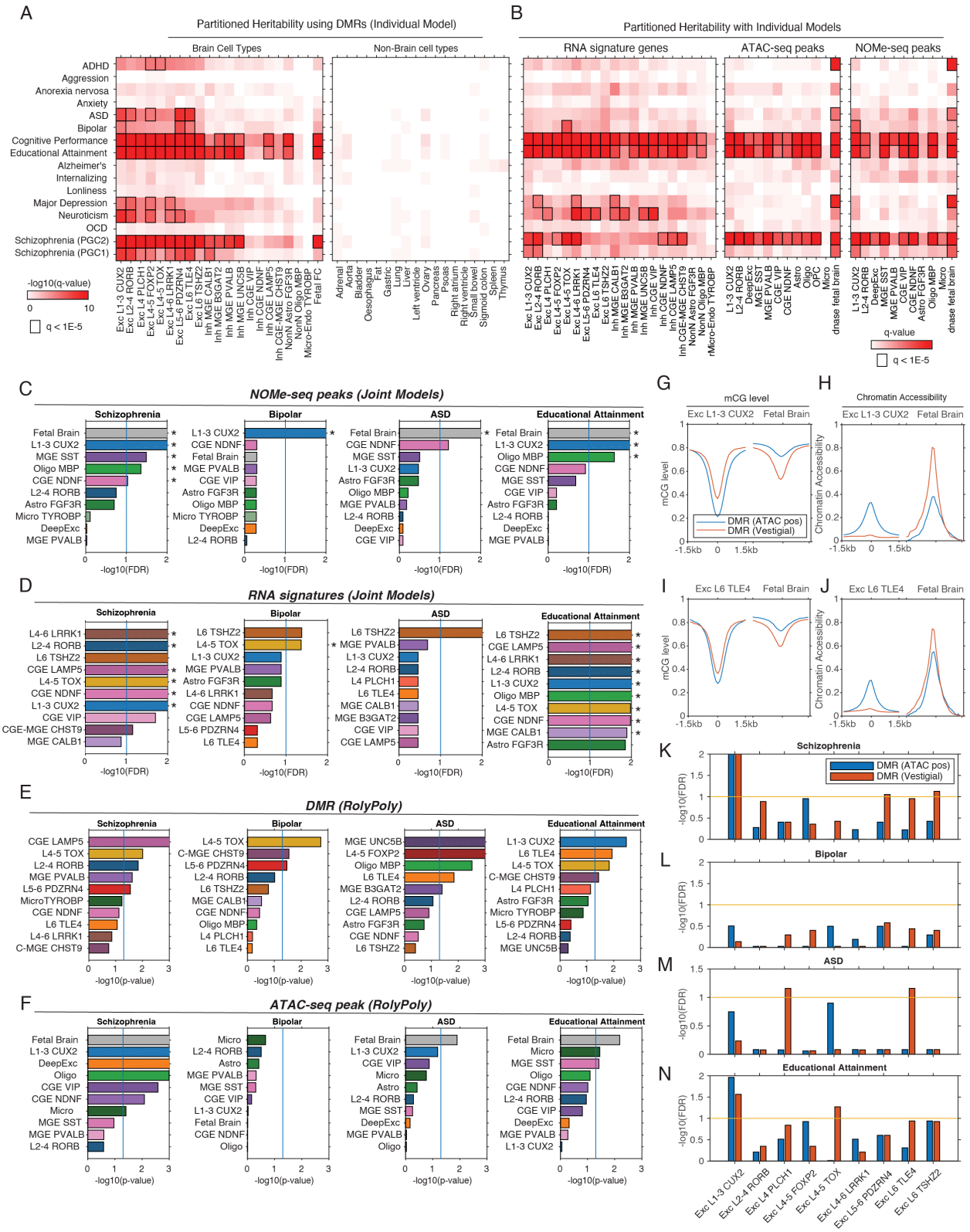


Supplementary Figure 3.6 TF binding motif enrichment across the human cortical neuronal hierarchy.

(A) Excitatory neuron subtype dendrogram. The node size represents the number of DMRs detected between the left and right branches. (B-D) Dot plots view for TFs showing lineage-specific motif enrichment, expression and gene body mCH between lineages: intratelencephalic (IT)-Deep-LRRK1 vs IT-Deep-MIX (B); IT-Deep-PDZRN4 vs IT-Sup (C); Cortico-Thalamic projection (CT) vs. layer 5/6 near-projecting neurons (NP) (D). Colors for every two rows from top to top: lineage mean gene body mCH level, lineage mean expression log(1 + CPM), TF motif enrichment log2(fold change). Sizes for every two rows from bottom to top: relative fold change of mCH level from this branch to the other, relative fold change of expression level, E-value of the motif enrichment test. Colors for the motif names: TF motif methylation preference annotated by methyl-SELEX experiment [6], orange indicate MethylPlus, green indicate MethylMinus. (E-H) Dot plot view for TFs showing lineage-specific motif enrichment, expression and gene body mCH between CGE vs. MGE (E), CALB1 vs. B3GAT2 (F), VIP/NDNF vs. LAMP5 (G) and Rosehip vs LAMP5-LHX6 (H). (**I**) The binding of TFs to hypermethylated regions validated by chromatin accessibility measurement using the snATAC-seq profile. (**J-L**) Enrichment or depletion of MethylPlus TFs (J), MethyMinus TFs (K) and TFs whose binding motif containing CA dinucleotides (L).


A
Excitatory Cell Type Dendrogram and Branch DMR Count
B
C
D
E
F
G
H
I
J
K
L
Chromatin Accessibility (ATAC-seq)
Motif log2(FC) in DMR
Norm RNA Exp
Gene Norm mCH%
E-value
Motif Sig
RNA FC
mCH FC
Color
Size
MethylPlus
MethylMinus
(methyl-SELEX annotation)
Methyl-Sensitive
Methyl-Resistant
ATAC-seq signal
-log10(q-value)
Depleted
Enriched

Supplementary Figure 3.7 Prediction of causal cell types for neuropsychiatric traits using partitioned heritability analysis.
(A) Partitioned heritability analysis individually comparing (individual models) adult brain cell-type-specific DMRs, fetal cortex lowly methylated regions, and non-brain tissues to the baseline (B) Partitioned heritability analysis (individual models) using adult brain RNA signature genes, ATAC-seq peaks, and NOMe-seq peaks. (C-D) Multiple regression partitioned heritability analysis using NOMe-seq peaks (C) and gene expression signatures (D). (E-F) Prioritization of neuropsychiatric traits-associated brain cell types using RolyPoly. (G-J) CG methylation and chromatin accessibility profiles at adult brain regulatory elements [DMR (ATAC-pos)] and vestigial enhancers. (K-N) Multiple regression partitioned heritability analysis comparing adult regulatory elements and [DMR (ATAC-pos)] and vestigial enhancers.


A
Partitioned Heritability using DMRs (Individual Model)
Brain Cell Types
Non-Brain cell types
ADHD
Aggression
Anorexia nervosa
Anxiety
ASD
Bipolar
Cognitive Performance
Educational Attainment
Alzheimer's
Internalizing
Lonliness
Major Depression
Neuroticism
OCD
Schizophrenia (PGC2)
Schizophrenia (PGC1)
-log10(q-value)
0
10
q < 1E-5
Exc L1-3 CUX2
Exc L2-4 RORB
Exc L4 PLCH1
Exc L4-5 FOXP2
Exc L4-5 TOX
Exc L4-6 LRRK1
Exc L5-6 PDZRN4
Exc L6 TLE4
Exc L6 TSHZ2
Inh MGE CALB1
Inh MGE B3GAT2
Inh MGE PVALB
Inh MGE SST
Inh CGE VIP
Inh CGE NDNF
Inh CGE LAMP5
Inh CGE-MGE CHST9
NonN Astro FGF3R
NonN Oligo MBP
Micro-Endo TYROBP
Fetal FC
Adrenal
Aorta
Bladder
Oesophagus
Fat
Gastric
Lung
Liver
Left ventricle
Ovary
Pancreas
Psoas
Right atrium
Right ventricle
Small bowel
Sigmoid colon
Spleen
Thymus
B
Partitioned Heritability with Individual Models
RNA signature genes
ATAC-seq peaks
NOMe-seq peaks
Exc L1-3 CUX2
Exc L2-4 RORB
Exc L4 PLCH1
Exc L4-5 TOX
Exc L4-6 LRRK1
Exc L6 TLE4
Exc L6 TSHZ2
Inh MGE CALB1
Inh MGE B3GAT2
Inh MGE PVALB
Inh MGE UNC5B
Inh CGE VIP
Inh CGE NDNF
Inh CGE LAMP5
Inh CGE-MGE CHST9
NonN Astro FGF3R
NonN Oligo MBP
rMicro-Endo TYROBP
Exc L1-3 CUX2
L2-4 RORB
DeepExc
MGE SST
MGE PVALB
CGE VIP
CGE NDNF
Astro
Oligo
OPC
Micro
dnase fetal brain
q-value
q < 1E-5
C
NOMe-seq peaks (Joint Models)
Schizophrenia
Bipolar
ASD
Educational Attainment
Fetal Brain
L1-3 CUX2
MGE SST
Oligo MBP
CGE NDNF
L2-4 RORB
Astro FGF3R
Micro TYROBP
DeepExc
MGE PVALB
CGE NDNF
Fetal Brain
MGE PVALB
CGE VIP
Astro FGF3R
Oligo MBP
Micro TYROBP
DeepExc
L2-4 RORB
Fetal Brain
CGE NDNF
MGE SST
L1-3 CUX2
Astro FGF3R
Oligo MBP
MGE PVALB
L2-4 RORB
DeepExc
CGE VIP
Fetal Brain
L1-3 CUX2
Oligo MBP
CGE NDNF
MGE SST
CGE VIP
Astro FGF3R
L2-4 RORB
DeepExc
MGE PVALB
-log10(FDR)
D
RNA signatures (Joint Models)
Schizophrenia
Bipolar
ASD
Educational Attainment
L4-6 LRRK1
L2-4 RORB
L6 TSHZ2
CGE LAMP5
L4-5 TOX
CGE NDNF
L1-3 CUX2
CGE VIP
CGE-MGE CHST9
MGE CALB1
L6 TSHZ2
L4-5 TOX
L1-3 CUX2
MGE PVALB
Astro FGF3R
L4-6 LRRK1
CGE NDNF
CGE LAMP5
L5-6 PDZRN4
L6 TLE4
L6 TSHZ2
MGE PVALB
L1-3 CUX2
L4 PLCH1
L6 TLE4
MGE CALB1
MGE B3GAT2
CGE VIP
CGE LAMP5
L6 TSHZ2
CGE LAMP5
L4-6 LRRK1
CGE NDNF
L2-4 RORB
L4-5 TOX
L1-3 CUX2
Oligo MBP
CGE NDNF
MGE CALB1
Astro FGF3R
-log10(FDR)
E
DMR (RolyPoly)
Schizophrenia
Bipolar
ASD
Educational Attainment
CGE LAMP5
L4-5 TOX
L2-4 RORB
MGE PVALB
L5-6 PDZRN4
MicroTYROBP
CGE NDNF
L6 TLE4
L4-6 LRRK1
C-MGE CHST9
L4-5 TOX
C-MGE CHST9
L5-6 PDZRN4
L2-4 RORB
L6 TSHZ2
MGE CALB1
CGE NDNF
Oligo MBP
L4 PLCH1
L6 TLE4
MGE UNC5B
L4-5 FOXP2
Oligo MBP
L6 TLE4
MGE B3GAT2
L2-4 RORB
CGE LAMP5
Astro FGF3R
CGE NDNF
L6 TSHZ2
L1-3 CUX2
L6 TLE4
L4-5 TOX
C-MGE CHST9
L4 PLCH1
Astro FGF3R
Micro TYROBP
L5-6 PDZRN4
L2-4 RORB
MGE UNC5B
-log10(p-value)
F
ATAC-seq peak (RolyPoly)
Schizophrenia
Bipolar
ASD
Educational Attainment
Fetal Brain
L1-3 CUX2
DeepExc
Oligo
CGE VIP
CGE NDNF
Micro
MGE SST
MGE PVALB
L2-4 RORB
Micro
L2-4 RORB
Astro
MGE PVALB
MGE SST
CGE VIP
L1-3 CUX2
Fetal Brain
CGE NDNF
Oligo
Fetal Brain
L1-3 CUX2
CGE VIP
Micro
Astro
L2-4 RORB
MGE SST
DeepExc
MGE PVALB
Oligo
Fetal Brain
Micro
MGE SST
Oligo
CGE NDNF
L2-4 RORB
CGE VIP
DeepExc
MGE PVALB
L1-3 CUX2
-log10(p-value)
G
mCG level
Exc L1-3 CUX2
Fetal Brain
mCG level
DMR (ATAC pos)
DMR (Vestigial)
-1.5kb
0
1.5kb
H
Chromatin Accessibility
Exc L1-3 CUX2
Fetal Brain
Chromatin Accessibility
I
Exc L6 TLE4
Fetal Brain
J
Exc L6 TLE4
Fetal Brain
K
Schizophrenia
DMR (ATAC pos)
DMR (Vestigial)
-log10(FDR)
L
Bipolar
M
ASD
N
Educational Attainment
Exc L1-3 CUX2
Exc L2-4 RORB
Exc L4 PLCH1
Exc L4-5 FOXP2
Exc L4-5 TOX
Exc L4-6 LRRK1
Exc L5-6 PDZRN4
Exc L6 TLE4
Exc L6 TSHZ2

CHAPTER 4

ALLCools: a comprehensive and scalable single-cell DNA methylation analysis framework

4.1 Abstract

ALLCools is a python package designed explicitly for single-cell DNA methylome and methylome-based multi-omic datasets. ALLCools includes data formats and associated functions to process, store, analyze, and visualize this unique epigenomic modality with high efficiency and scalability. Its core matrix formats, MCDS and RegionDS, support storing DNA methylome data with various features and methylation types, providing cellular and genomic analysis infrastructures. With ALLCools, we established two methods for cell clustering and tested them in DNA methylome datasets from various tissue sources. Besides, ALLCools' python-based implementation allows seamless integration with other established packages for more specific analysis purposes. After cellular analysis, we also demonstrated the integration of multiple single-cell epigenomic technologies to predict cell-type-specific enhancers and linked enhancers to potential targeting genes in mouse brain hippocampus.

4.2 ALLCools Analysis Overview

We design ALLCools as a comprehensive and scalable framework for single-cell DNA methylation analysis. The schematic overview (Fig. 4.1) shows that the ALLCools stores single-cell cytosine base counts in the ALLC format (Fig. 4.1A). The analysis starts from aggregating methylation information into N-dimensional labeled arrays corresponding to different feature sets (such as gene, genomic bins, promoter, CGI), molecular modalities (methylation, accessibility, expression), methylation contexts (CpG, CpH, GpC), and quantification types (count, fraction,

hypo- or hyper-score) via a single command ("allcools generate-dataset"). These arrays are stored in a dataset named MCDS or RegionDS (Fig. 4.1B, C). The MCDS and RegionDS share the same data structure but have different analysis functions: the MCDS serves as the basis of cellular analysis, while the RegionDS is associated with genomic analysis. Both dataset formats are based on the Xarray package[179] in memory and serialized on disk using the Zarr package[180]. Both packages natively support handling N-dimensional dense arrays and are highly scalable to large data through chunk access and on-disk computation.

For cellular analysis with the MCDS, we provide a methylation-specific version of canonical functions for quantification, cell and feature preprocessing, dimension reduction, graph-based clustering, and embedding. Importantly, we developed two methods for methylation clustering, one with methylation fractions of large genomic features, the other with methylation scores of small genomic features. These two methods cover different feature sizes, from cis-regulatory elements and promoters to large gene bodies (Supplementary Table 4.2). Besides, the python basis of MCDS provides interoperability with the popular single-cell data structure, AnnData[44], allowing to adapt of existing algorithms or packages for specific analysis purposes, such as the harmony[52] and scanorama[37] for batch correction and data integration; the Scrublet[181] for doublets detection; the MOFA+[48] and Seurat v3 WNN[182] for multi-omic factor analysis (through the MUON package[51]).

In genomic analysis, we usually create the RegionDS with Differentially Methylated Regions (DMR) analysis by the methylpy algorithm[41]. However, RegionDS can also be created from any genomic region (such as genes, promoters, CGIs, or DMRs from other packages). In addition to quantifying the methylation level of these regions, the RegionDS provides functions for annotation by BigWig and BED files, scanning DNA motifs, performing motif enrichment

analysis, and calculating region-to-region correlation. Furthermore, the rich annotated datasets provide a basis for applying machine learning models to predict epigenomic activities or features[53]. For example, we demonstrated an enhancer prediction analysis with the REPTILE model. Furthermore, by involving an snm3C-seq[24] dataset from the same type of tissue[15], we predicted the potential targeting genes of these enhancers via the chromatin contact loops[183].

4.3 ALLCools Data Structures

The ALLCools data processing starts with base-level cytosine count information saved in the ALLC file, a tab-separated table compressed and indexed by bgzip and tabix[184] to allow querying with genomic coordinates. In addition, the ALLCools package provides functions to generate ALLC files after mapping (allcools bam-to-allc) or transfer from other methylation count formats (allcools table-to-allc). Additional functions in ALLCools can also help extract, merge and convert ALLC files throughout the analysis pipeline. Notably, one ALLC file only stores methylome information for one sample (a cell or a pseudo-bulk profile).

Next, we aggregate multiple ALLC files into a cell-by-feature format called MCDS to collectively analyze hundreds of thousands of cells from a single-cell assay. Unlike existing cell-by-feature formats designed for single-cell RNA data[43,51,185], the MCDS has several features to support large-scale DNA methylation and multi-omic technology specifically. First, the DNA methylation dynamics occur at a wide range of scales (10^2-10^7 bp) in the genome (Supplementary Table. 4.2), studying multiple features in the same dataset is a common task for methylome analysis[3,15,17]. The MCDS, generated in a single step (allcools generate-dataset), natively supports storing and aligning multiple sets of features together. Second, most single-cell mC technologies are based on bisulfite conversion, which chemically converts unmethylated C to T[186] and quantifies DNA methylation fraction based on the total base count and mutated base count.

Besides, the native or artificially introduced difference between methylation context (CpG, CpH, CpA, or GpC) also requires counting cytosines in multiple categories. The MCDS supports all these requirements by using N-dimensional labeled arrays to store data. Noteworthy, this strategy also supports storing quantifications from additional modalities in MCDS. The underline implementations are adapted from the xarray package, originally designed to handle geographical and meteorological data[179], whose N-dimensional structure is analogous to single-cell mC and multi-omic datasets. Finally, besides handling the above complexities, MCDS has high scalability in both the cell dimension and feature dimensions (See below for scalability discussion).

While MCDS is cellular analysis centric, the same infrastructure is also suitable for genomic analysis, where the focus becomes genome regions, and cells or cell clusters can be treated as features to annotate these regions. Therefore, we provide another in-memory representation of the data structure called RegionDS, which provides functions performing analysis on genome regions, but shares the same underlying data structure with MCDS. The MCDS and RegionDS provide the foundations for all the cellular and genomic analysis in ALLCools.

4.4 Cellular Analysis with ALLCools MCDS

We first demonstrate the core functions of ALLCools for dimension reduction, which is the essential step for single-cell data visualization and cell-type identification. Using 16,985 snmC-seq2 cells from mouse hippocampus (HIP)[15], we demonstrated the basic clustering process using the MCDS file with 100kb genomic bins as features. After quality control to filter out low-quality cells (low read number, high non-conversion rate) and features (low read coverage, overlap ENCODE blacklist[73]), ALLCools compute posterior methylation level and normalize within each cell and select highly variable features to feed into PCA for dimension reduction. These steps were

applied to both mCH and mCG, and the principal components were combined for further UMAP visualization and consensus cluster analysis. We identify 28 clusters and annotate them by checking the gene body mCH (for neurons) or mCG (for non-neuronal cells) level of known marker genes for major cell types (Fig. 4.3A).

Neuronal cells have highly diverse methylation patterns, including the mCH signatures across the whole gene bodies and the large super-enhancer-like regions with differentially CG methylation across cell types[187]. Together, these make it possible to use methylation levels at 100kb resolution to resolve cell types at high granularity[3,15]. Contrarily in other tissues where fewer long and specific features were explicitly observed, classifying cell types using DNA methylation data only remains challenging. Existing methods often focus on the methylation level of promoters or predefined enhancers[17,18]. However, recent studies have also revealed limited methylation diversity near TSSs across cell types, while relying on annotated enhancers limits the generalizability of the analysis in the samples without histone data. To develop a universally adaptable framework, we use 5kb genomic bins as features to perform dimension reduction, with the rationale to capture methylation dynamics at shorter regulatory elements. Given the coverage of single-cell DNA methylation assays, the majority of 5kb bins across the mammalian genome are not covered, making it challenging to save the full cell-by-5k bin matrices either in memory or on disk. To alleviate this issue, ALLCools compute a hypo-methylation score in each 5kb bin for every single cell, assuming regulatory elements are usually hypo-methylated. The score accounts for the methylation level and coverage with a binomial model. Only the bins with a score > 0.9 are stored, while all other bins are saved as 0. This resulting cell-by-bin matrix is a sparse matrix with ~2% non-zero values, considerably reducing computation cost and memory and disk usage. After

generating a hypo-methylation score matrix, we binarize the matrix with a threshold and perform latent semantic indexing (LSI) for dimension reduction, similar to scATAC data analysis[43,188].

We first benchmarked this method on the mouse HPF dataset and observed comparable, if not better, capability to separate the annotated cell types (Fig. 4.3B). To further test the generalizability of this method, we used FACS to select eight specific cell populations from peripheral blood mononuclear cells (PBMC) and profiled them with snmC-seq2[28]. This dataset collects Naive helper T cells (CD3+, CD4+, CCR7+, CD45RA+), Memory helper T cells (CD3+, CD4+, CD45RA-), Naive cytotoxic T cells (CD3+, CD8+, CCR7+, CD45RA+), Memory cytotoxic T cells (CD3+, CD8+, CD45RA-), B cells (CD3-, CD19+), Monocytes (CD3-, CD19-, CD14+), NK cells (CD3-, CD19-, CD14-, CD16+, CD56+), and other cells (CD3-, CD19-, CD14-, CD16-, CD56-), with the FACS providing a standard for cell-type annotation. The 5kb LSI framework in ALLCools can separate the above cell types and resolve more rare cell types (Fig. 4.3C), including CD16+ monocytes, Dendritic cells (DC), and Plasmacytoid DC (pDC). The cell type annotation is corroborated by integrating published scATAC-seq data using 5kb bins as features. The naive PCA-based framework failed to distinguish between helper and cytotoxic cells with either 100kb bins, gene bodies, or promoters as features (Fig. 4.3D). To obtain a more intuitive interpretation of why 5kb bins could work better, we identified DMRs between Memory and Naive T cells and between helper and cytotoxic T cells. Indeed, we find the differences between Memory and Naive cells are at a whole-genome scale, with a 10% global methylation level change, resulting in widespread DMRs across large bins. On the contrary, most of the differences between helper and cytotoxic cells are localized near the marker genes and expand to >10 kb regions less frequently (Fig. 4.3E).

We also noted that specific methods or features could outperform these general frameworks on certain datasets. For example, MOFA[48] can separate embryonic cell states better than PCA when using predefined H3K27Ac peaks as features (Fig. 4.3G), potentially due to its model handling missing values in the matrix. Since ALLCools features its scalability on large datasets (discussed below), we did not use MOFA as a default embedding approach. Nevertheless, the ALLCools can seamlessly include MOFA into the analysis with the recently developed MUON package[51]. The flexibility allows users to choose any dimension reduction methods they prefer in the large Python community.

4.5 Genomic Analysis with ALLCools RegionDS

Next, we use ten major cell types (Fig. 4.3A) from the mouse HIP dataset to demonstrate the genomic analysis using ALLCools RegionDS (Fig. 4.4A). First, based on the cell type labels from clustering analysis, we merged the relatively shallow single-cell ALLC files into pseudo-bulk ALLC files with the "allcools merge" command to increase the genome coverage per cluster. We then use the methylpy algorithm [5,41] to identify Differentially Methylated Regions (DMR). The methylpy algorithm implemented in ALLCools will produce a RegionDS containing the mCG counts of DMRs directly. ALLCools also support creating RegionDS from ALLC files and a set of regions in a BED file, such as DMRs from other tools like DSS[189] or DMRseq[190]. After storing the DMR information into the RegionDS, we can annotate the regions with BigWig and BED files with one function call. This annotation process facilitates further analysis of DMR, such as motif analysis, DMR-gene correlation, and enhancer prediction.

In this example, we identified a total of 732,544 DMRs from the 10 HIP cell types (Fig. 4.4B, Supplementary Fig. 4.1). We then used the pseudo-bulk profile of snATAC-seq data[29] from matched cell types to annotate HIP DMRs with chromatin accessibility information. As expected,

the hypo-DMRs for a cell type are more accessible in the same cell type than the others (Fig. 4.4C, Supplementary Fig. 4.2), indicating the correct match of cell types from both datasets. The combination of two modalities allows us to train a REPTILE model[53] to predict candidate enhancer regions. After prediction, we get 102,888 candidate enhancers for each cell type (Fig. 4.4D, Supplementary Fig. 4.3). Interestingly, the chromatin contact (from snm3C-seq) between the enhancers and the promoters of DMGs is stronger in matched cell types than in unmatched cell types, indicating the genome 3D conformation is also involved in forming the cell-type-specific gene regulation (Fig. 4.4E, Supplementary Fig. 4.4).

As suggested by this differential interaction between enhancers and gene promoters, we further identified chromatin contact loops to predict the potential target genes of the enhancers. For example, *Celf2*, an RNA alternative splicing regulatory protein related to axon extension[191], contains hypo-methylated, enhancer-overlapping DMRs in the CA1 cells but not astrocytes. These potential cis-regulatory elements (CRE) are located inside the gene body and up to 1.2 Mb of the transcription start site (TSS) of Celf2. The chromatin conformation map indicates that both the gene and CREs are inside a topologically associating domain (TAD), with chromatin loops further linking some enhancer elements to the TSS of *Celf2* (Fig. 4.5A, Supplementary Fig. 4.5). Likewise, the gene *Slc1a3* is a glial high-affinity glutamate transporter highly expressed in astrocytes[192]. The CREs identified 200 Kbs upstream of the gene body are linked to TSS via an ASC specific chromatin loop (Fig. 4.5B, Supplementary Fig. 4.6). In another gene, neurodevelopmental transcription factor *Bcl11b*[193], the furthest loop-linked CA1-specific CRE is 1.5 Mbs downstream to the TSS (Fig. 4.5C), which also has CA1 specific hypo-mCG and high accessibility (Supplementary Fig. 4.7), provides an example of ultra-long-range enhancer-promoter interaction associated by the chromatin loops.

As a deliverable resource of this pipeline, we systematically annotated the DMRs of these 10 clusters in a RegionDS, with annotation and labeling on whether the DMR is overlapping with predicted enhancers or whether a chromatin loop in the corresponding cell type link this DMR to the TSS to the DMG of this cell type (Fig. 4.5D, E, Supplementary Fig. 4.8). Together, the linkage between DMGs and DMRs with the chromatin contact information prioritizes the candidate enhancer validation for future studies.

4.6 Discussion

The booming of single-cell methylome and multi-omic technologies brings remarkable opportunities to study DNA methylation in diverse cell types and link it with other molecular modalities. The ALLCools package is specifically designed to meet the unique computational challenges arising from these new technologies. Based on the python environment, ALLCools contains core data structures, MCDS and RegionDS, modeling multi-dimensional labeled data arrays, which is vital for the DNA methylome analysis under multiple sets of features, different methylation contexts, and various ways of quantification. In addition, ALLCools also provides all essential functions for single-cell analysis, inspired by the popular single-cell analysis packages[44,182], with essential adaptations for the methylation dataset. Internally, ALLCools utilize mature packages in the python ecosystem, such as Xarray[179], Zarr[180] and Dask[194], to efficiently handle vast datasets, providing massive scalability for future atlas-scale studies.

The ALLCools analysis workflow contains two main parts. First, the cellular analysis is based on the MCDS and its associated functions, covering data preprocessing, filtering, dimension reduction, embedding, clustering, differentially methylated gene analysis, and more. Notably, we established two complementary methods for different sizes of DNA methylation features, providing great generalizability on datasets from various tissues and species. Second, the genomic

analysis is based on RegionDS and other functions, covering differentially methylated region analysis, annotating genomic regions, motif analysis, enhancer prediction, and linking regions to target genes.

Together, ALLCools aims to provide a comprehensive and scalable platform for handling the single-cell DNA methylome dataset in the python ecosystem. Other single-cell or genomic packages focusing on various downstream analysis topics can quickly utilize and combine with the ALLCools workflow to complete more general analysis goals. In addition to DNA methylome, more assays covering additional epigenomic modalities, such as snhmC-seq for 5hmC[195], CLEVER-seq for 5fC[196], are ready to be analyzed by the ALLCools framework. Besides, future development of ALLCools will focus on providing more seamless integration with other multi-omic assays containing diverse molecular information, such as the multiome assay from 10X genomics, Paired-Tag[197] or similar assay, or more recently, the spatial transcriptome[198,199] and epigenome technologies[200,201]. Bringing all the multi-omic information together will enable a deeper investigation of the functional impact of DNA methylation, providing opportunities to understand the gene regulation mechanisms in complex biological systems.

4.7 Methods

4.7.1 Implementation of ALLCools

ALLCools has been implemented in the Python programming languages and based on a number of scientific and bioinformatic open-source libraries. In particular, NumPy[202], Scipy[203], Pandas[204], Xarray[179], Zarr[180], Numba[205] and Dask[194] is used to establish the MCDS and RegionDS data structure; htslib[206], biopython[207], Bedtools[208], Samtools[155], pyBigWig[209] is used to handle genomic file formats; scikit-learn[210], anndata[185], scanpy[44], pynndescent[211], matplotlib[212], and seaborn[213] were used to build analysis functions.

4.7.2 Implementation of ALLC, MCDS and RegionDS format

The ALLC (ALL Cytosine) format is a tab-separated table containing base level methylation and coverage counts. Each row in an ALLC file corresponds to one cytosine in the genome. An ALLC file contains 7 mandatory columns describing the position, coverage, context, and level of methylation (Supplementary Table 4.3). ALLCools always compress and index ALLC files by bgzip and tabix[184] to allow access to the compressed files via genomic regions. There are three ways to generate an ALLC file:

1) Generate from a single-cell BAM file after mapping via "allcools bam-to-allc" (https://lhqing.github.io/ALLCools/command_line/allcools_allc.html);

2) Merge multiple ALLC files into an pseudo-bulk ALLC file via "allcools merge" (https://lhqing.github.io/ALLCools/command_line/allcools_merge.html);

3) Convert from other form of methylation count table via "allcools table-to-allc" (https://lhqing.github.io/ALLCools/command_line/allcools_table_to_allc.html).

Both the MCDS and RegionDS format represent a collection of labeled multi-dimensional arrays. The in-memory representation of MCDS and RegionDS builds on the Dataset class in the Xarray package[179], which associates raw multi-dimensional data arrays with labeled dimensions, coordinates and attributes. This labeling system is a key feature to allow accessing the data consistently and associating them with the cell or gene metadata. The on-disk serialization of MCDS and RegionDS builds on the zarr package[180], which is a storage format for chunked, compressed, multi-dimensional arrays. MCDS is generated from a set of ALLC files and multiple genome region sets (gene, chromosome bins, promoter, CGI, etc.), via "allcools generate-dataset" (https://lhqing.github.io/ALLCools/command_line/allcools_dataset.html). RegionDS is created from DMR analysis

(https://lhqing.github.io/ALLCools/cluster_level/RegionDS/01a.call_dmr.html) or form a set of genomic regions identified by other tools (https://lhqing.github.io/ALLCools/cluster_level/RegionDS/01b.other_option_to_init_region_ds.html), and further annotated by BigWig or BED files (https://lhqing.github.io/ALLCools/cluster_level/RegionDS/02.annotation.html).

4.7.3 Fraction-based clustering using genome 100kb bins

We use mc_{ij} and cov_{ij} to denote the total cytosine basecalls in a specific sequence content in cell i and feature j. For each cell, we calculated the mean (m) and variance (v) of the mCH (or mCG) level across the features. A beta distribution was then fit for each cell i, where the parameters were estimated by

$$\alpha_i = m_i(\frac{m_i(1-m_i)}{v_i} - 1)$$

$$\beta_i = (1-m_i)(\frac{m_i(1-m_i)}{v_i} - 1)$$

We next calculated the posterior mC level of each bin j by

$$ratio_{ij} = \frac{\alpha_i + mc_{ij}}{\alpha_i + \beta_i + cov_{ij}}$$

We normalized this rate by the cell's global mean methylation by

$$global_i = \frac{\alpha_i}{\alpha_i + \beta_i}$$

$$M_{ij} = \frac{ratio_{ij}}{global_i}$$

The values greater than 10 in M were set to 10. After normalization, M_{ij} is close to 1 when cov_{ij} is close to 0. PCA was perform on M to generate embeddings for UMAP visualization and cluster analysis.

4.7.4 LSI-based clustering using genome 5kb bins

In a single cell i, we modeled its mCG base call M_{ij} for a 5kb bin j using a binomial distribution $M_{ij} \sim Bi(cov_{ij}, p_i)$, where p represents the global mCG level of the cell. We then computed $P(M_{ij} > mc_{ij})$ as the hypomethylation score of cell i at bin j. The less likely to observe smaller or equal methylated basecalls, the more hypomethylated the bin is. We next binarized the hypomethylation score matrix M by setting the values greater than 0.95 as 1, otherwise 0, to generate a sparse binary matrix A. Latent semantic analysis with log term frequency was applied to A to compute the embedding. Specifically, we selected the columns having 1 in more than 5 rows, then computed the column sum of the matrix ($colsum_j = \sum_{i=1}^{\#cell} A_{ij}$) and kept only the bins with Z-scored $log_2 colsum$ between -2 and 2. The filtered matrix was normalized by dividing the row sum of the matrix to generate TF, and further converted to X used for singular value decomposition $X = USV^T$, where $X_{ij} = log(TF_{ij} \times 100000 + 1) \times log(1 + \frac{\#cell}{colsum_j})$. Top dimensions of U were then used for UMAP visualization and cluster analysis.

4.7.5 Multi-omic analysis

For the mouse gastrulation scNMT-seq dataset[17], we generated mCG count matrix of E7.5 H3K27Ac peaks with ALLCools generate-mcds. The peaks having zero coverage in all cells were removed from analysis. The methylation fraction matrix was computed by $\frac{mc_{ij}}{cov_{ij}}$. The peaks not sequenced in a cell were assigned with nan in the matrix. MOFA+ implemented in muon was used for dimension reduction of the fraction matrix with 10 factors, followed by UMAP for visualization.

4.7.7 Enhancer prediction workflow

The enhancer prediction workflow has been previously described in Liu et al.[15] A brief description of the main steps are as follow:

1) Merge ALLC and Call DMR: after cell clustering and integration, we merged single cell ALLC files via "allcools merge" into ten pseudo-bulk ALLC files. Then we used the "call_dmr" function in ALLCools package to identify DMRs and generate the initial RegionDS (https://lhqing.github.io/ALLCools/cluster_level/RegionDS/01a.call_dmr.html#call-differentially-methylated-sites). This function is a reimplementation of the "methylpy callDMR" method based on the code in (70). The algorithm is formally described in Schultz et al (26).

2) Filter DMRs: We calculated DMRs for both snmC and snm3C pseudo-bulk profiles (20 samples in total). We filtered the DMRs by requiring each DMR has >= 2 differentially methylated CpG pairs (DMS), and have the same trend (hypo- or hyper-methylated) in both the snmC and snm3C samples. Other parameters are as the default of the "call_dmr" function.

3) We then annotated the RegionDS by cluster matched snATAC-seq pseudo-bulk profiles. We then used the REPTILE model in ALLCools to predict active enhancers using both the mCG fraction and chromatin accessibility information (https://lhqing.github.io/ALLCools/cluster_level/REPTILE/02.run_reptile.html). The REPTILE model is implemented as described in He et al.[53] and adapted into the RegionDS framework.

4) Call final enhancers: The enhancer is called by the REPTILE with default parameters (prediction score > 0.3). The REPTILE model predicts enhancers both on DMR and genome bin coordinates (2Kb-long and 100bp-step moving windows). However, the final enhancer call from the REPTILE model is neither DMR nor genome bins, but called from the combination of them whenever a higher prediction score is seen[53]. We followed this framework but annotated the enhancer-overlapping DMR using the enhancer identified from REPTILE as well. Both coordinates are provided.

4.7.8 Chromatin conformation analysis workflow

Processing of the 3D genome conformation modality from the snm3C-seq is based on the schicluster[91] and cooler[214] package as previously described[15]. We first save single cell contact matrices into SCOOL format defined in the cooler package[214], then use schicluster's random walk with restart (RWWR) algorithm to impute the chromatin matrix at 10Kb resolution for identifying chromatin loops. The loop calling is based on a recent package designed for single-cell HiC dataset, SnapHiC[183]. This package leverages the cell-level normalization and statistical test, reaching higher sensitivity than bulk HiC loop calling methods, therefore identifying more chromatin loops given the same coverage[183].

4.8 Data and Code Availability

The published single cell methylome and multiomic data set is downloaded from GEO, all accession numbers are listed in Supplementary Table 1. All analysis was done in ALLCools v1.0.0. The ALLCools documentation is available at https://lhqing.github.io/ALLCools/intro.html, and the source code is available at https://github.com/lhqing/ALLCools.

4.9 Acknowledgements

Chapter 4, in full, is currently being prepared for submission for publication of the material. "ALLCools: a scalable and comprehensive single-cell DNA methylation analysis framework." Hanqing Liu, Jingtian Zhou, Wei Tian, Jiayin Xu, Joseph R. Ecker. The dissertation author was the primary investigator and author of this material.

4.10 Figures

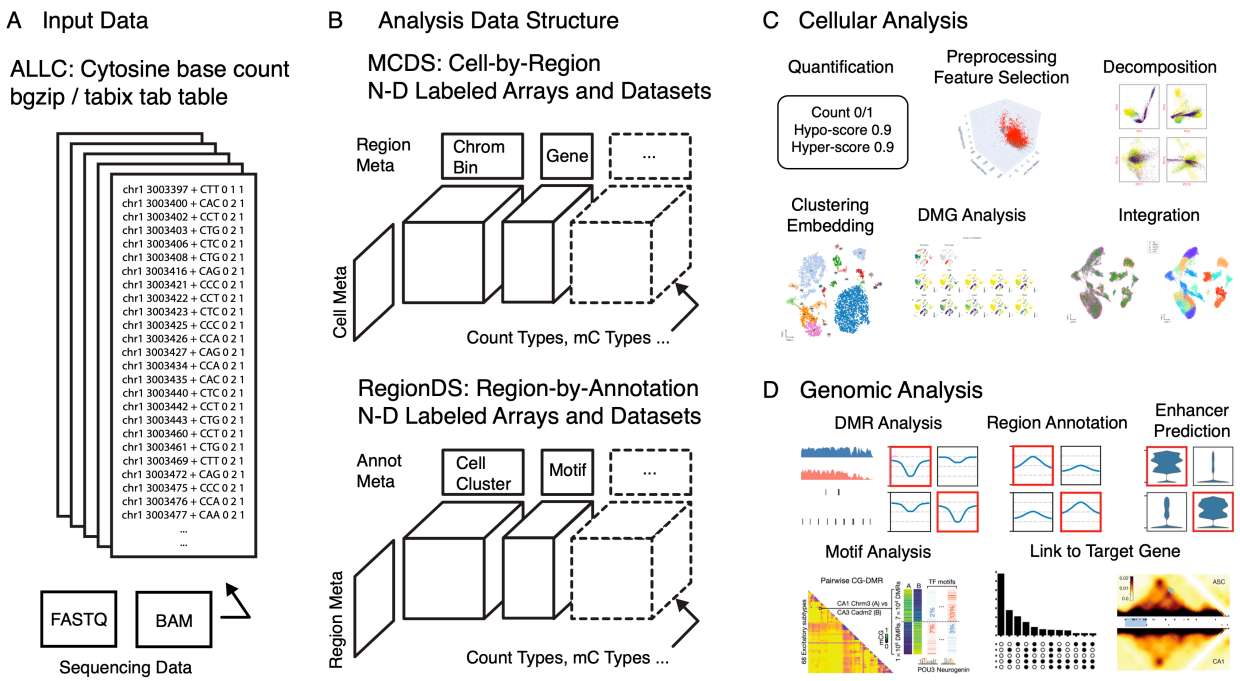


Figure 4.1 ALLCools analysis overview

(A) ALLC format is the input of ALLCools data analysis, it contains all the raw methyl-cytosine base count across genome. (B) ALLCools defines two N-dimensional labeled array data structure for cellular and genomic analysis. (C) Example of cellular analysis supported by ALLCools MCDS and associated functions. (D) Example of genomic analysis supported by ALLCools RegionDS and associated functions.

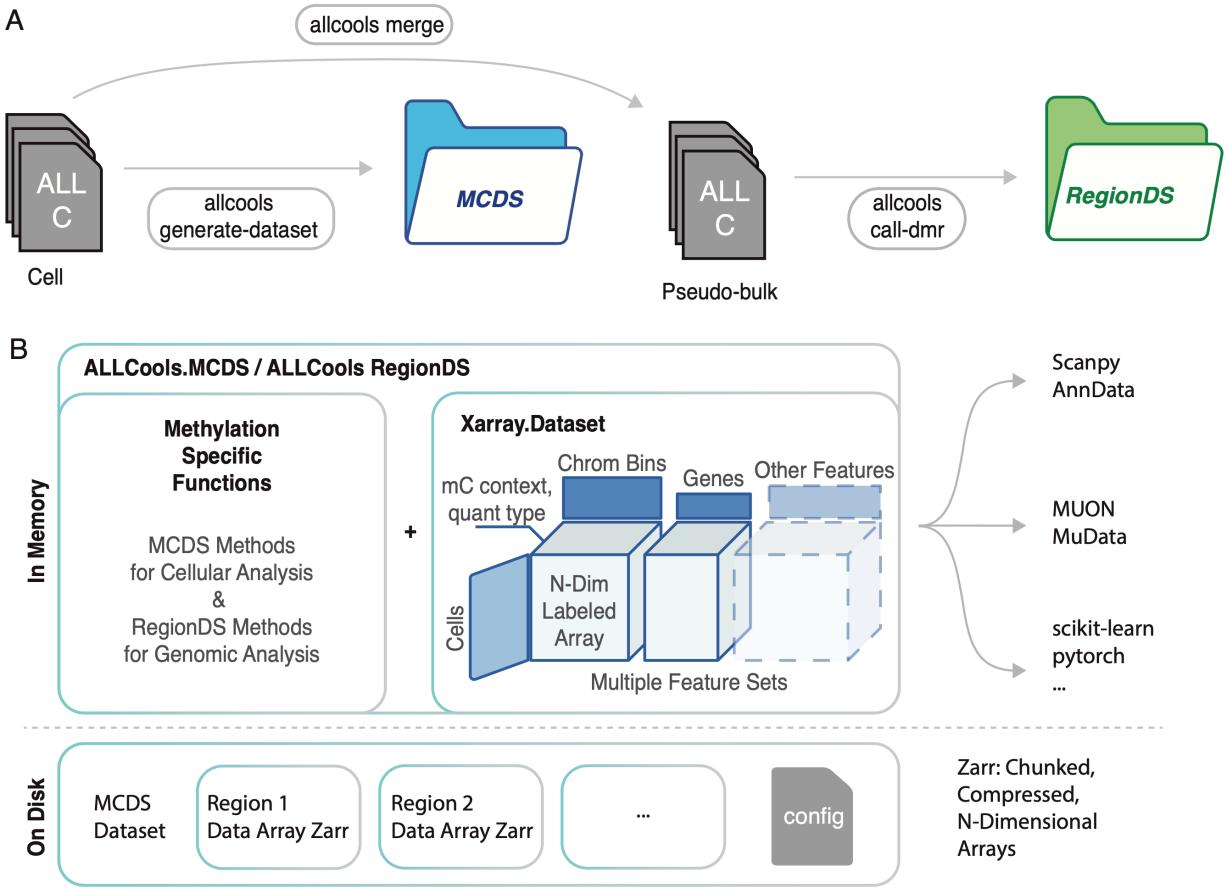


Figure 4.2 ALLCools Data Model

(A) ALLC uses ALLC files as input to generate cell-by-feature datasets and saved in MCDS. After cellular analysis, single-cell ALLC files can be merged into pseudo-bulk ALLC files, then, after calling DMR or given any regions of interest, a RegionDS can be generated to perform following genomic analysis. (B) The implementation detail of ALLCools MCDS and RegionDS. The in-memory representation is mainly based on Xarray package[179]. The on-disk serialization uses Zarr package[180].

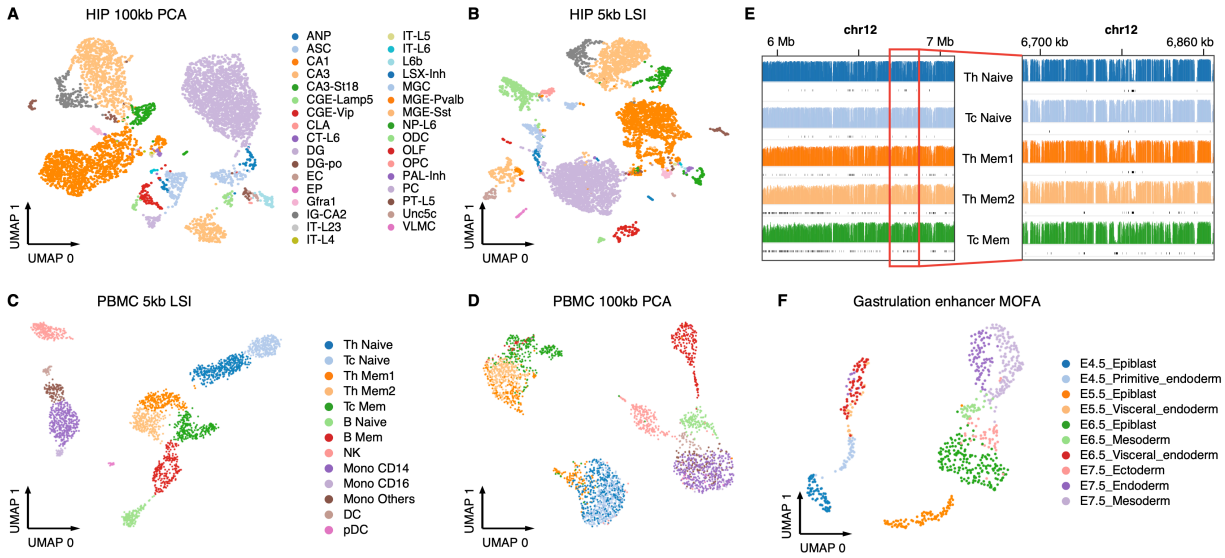


Figure 4.3 Cellular Analysis of ALLCools

(A) UMAP embedding of mouse HIP cells using 100kb mCH and mCG fractions as features and PCA as dimension reduction method. (B) UMAP embedding of mouse HIP cells using 5kb mCG hypo-methylation score as features and LSI as dimension reduction method. (C) UMAP embedding of human PBMC cells using 100kb mCH and mCG fractions as features and PCA as dimension reduction method. (D) UMAP embedding of human PBMC cells using 5kb mCG hypo-methylation score as features and LSI as dimension reduction method. (E) Genome browser view of the mCG fractions and DMRs (small tracks below each large ones) of human PBMC Th Naïve, Tc Naïve, Th Memory 1, Th memory 2, and Tc Memory cells. (F) UMAP embedding of mouse gastrulation data using enhancer regions as features and MOFA+ as dimension reduction method.

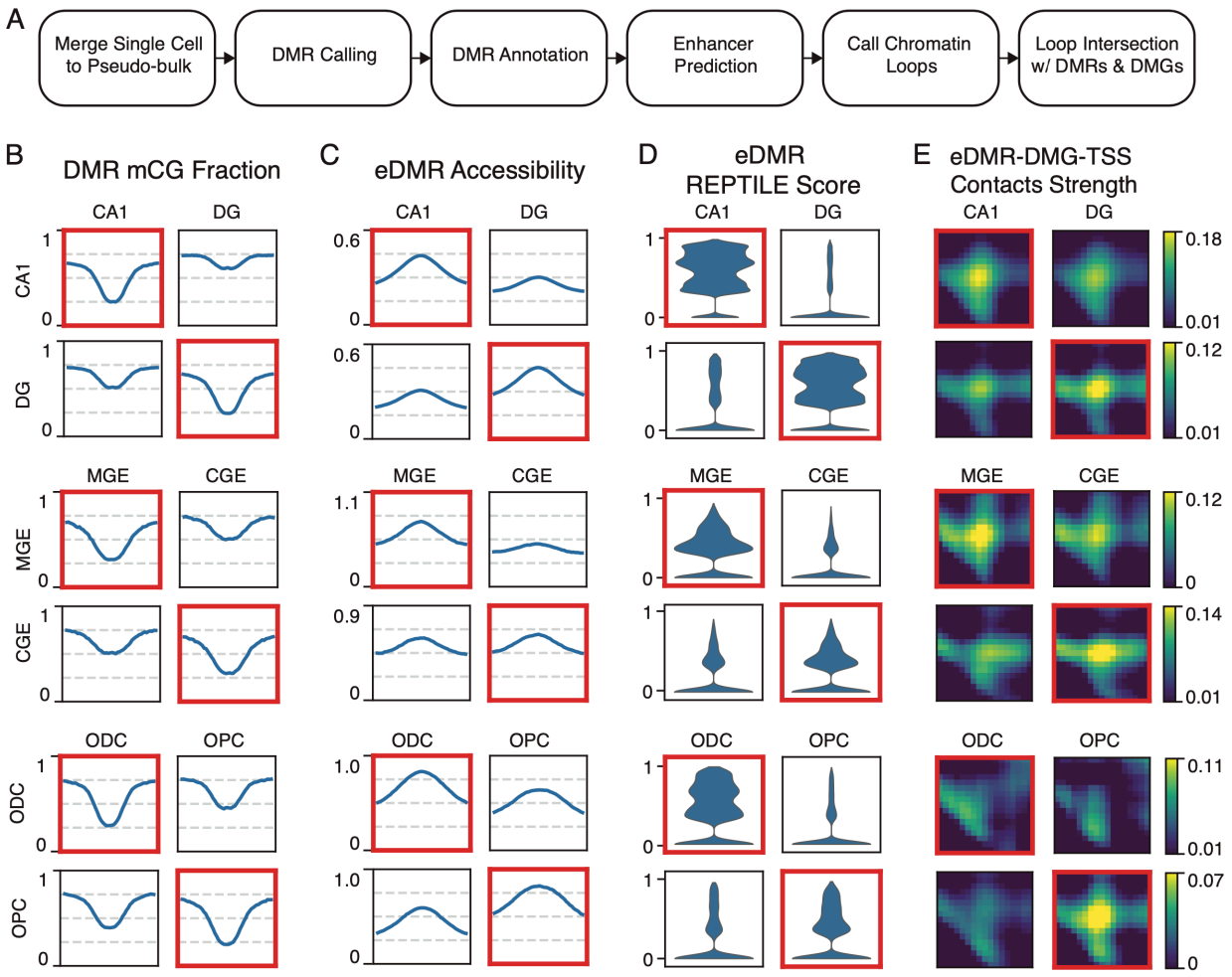


Figure 4.4 Enhancer Prediction in mouse HIP dataset

(A) The enhancer prediction workflow. (B) DMR mCG fraction profile plot, x-axis is centralized at DMR center with 10Kb flanking. The red box indicates the DMR is hypo-methylated in the corresponding cell type of that row. (C) Enhancer-overlapping DMR accessibility profile plots, similar to (B). (D) Enhancer-overlapping DMR REPTILE prediction score violin plot. The red box indicates the DMR overlap with enhancers in the corresponding cell type of that row. (E) Normalized contact strength between Enhancer-overlapping DMRs and DMG TSS regions. The red box indicates the DMR and DMG are selected from the corresponding cell type of that row.

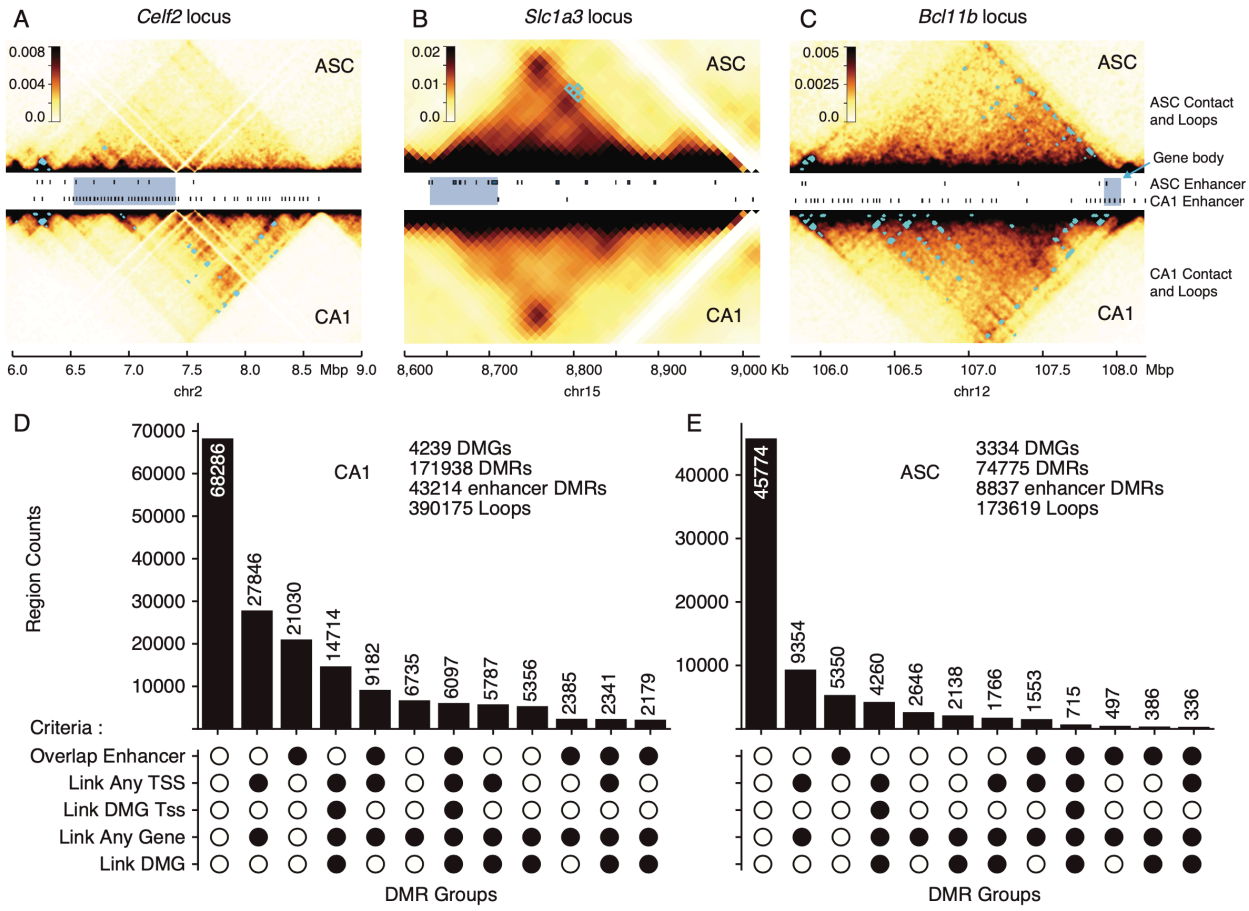
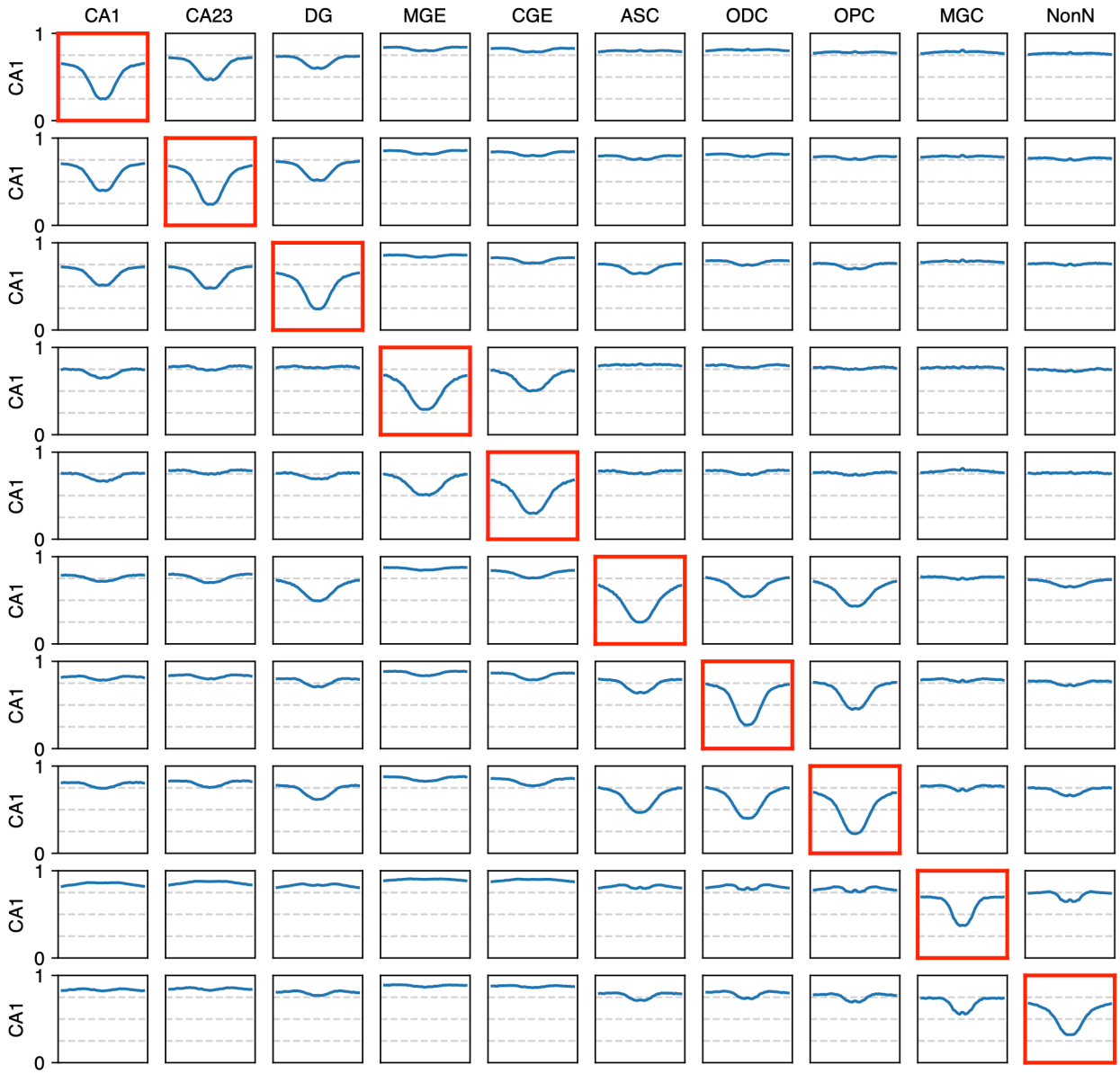


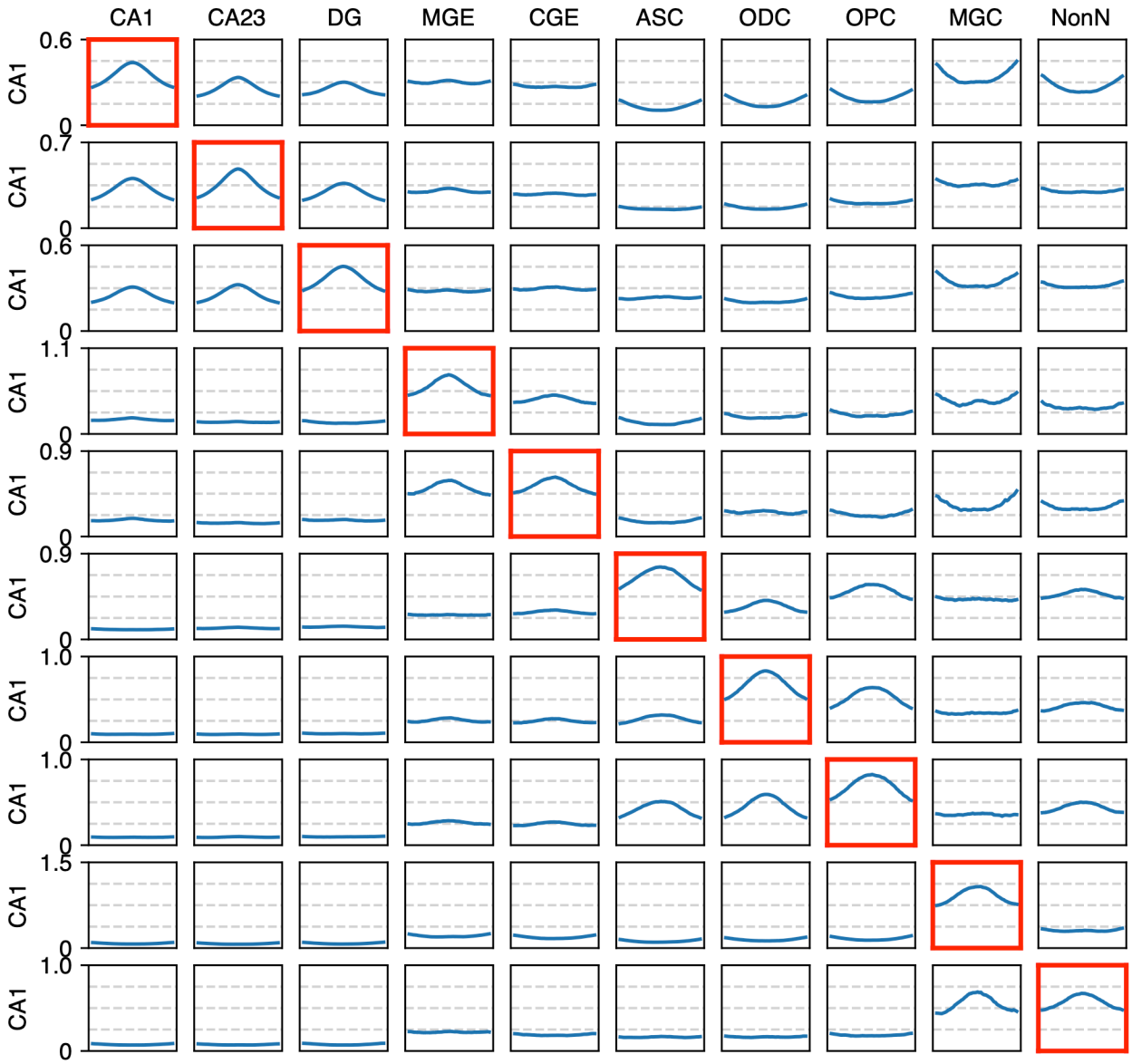
Figure 4.5 Chromatin contact loops link enhancer to gene TSS

(A-C) The normalized chromatin contact matrix at the *Celf2* (A), *Slc1a3* (B), and *Bcl11b* (C) locus. The top part shows chromatin conformation of astrocytes (ASC), the bottom part shows CA1 neurons. The cyan dots on heatmaps indicates SnapHiC-identified chromatin loops. (D) The number of CA1 DMRs dissected by different criteria. From top to bottom, the five criteria are: 1) Whether the DMR is overlapping with REPTILE enhancers from CA1; 2) Whether the DMR is linked to TSS by a chromatin loop; 3) Whether the DMR is link to TSS of a DMG by a chromatin loop; 4) Whether the DMR is link to any gene body region by a chromatin loop; 5) Whether the DMR is link to the gene body region of a DMG by a chromatin loop. (E) The number of ASC DMRs dissected by different criteria, similar to (D).

4.11 Supplementary Figures

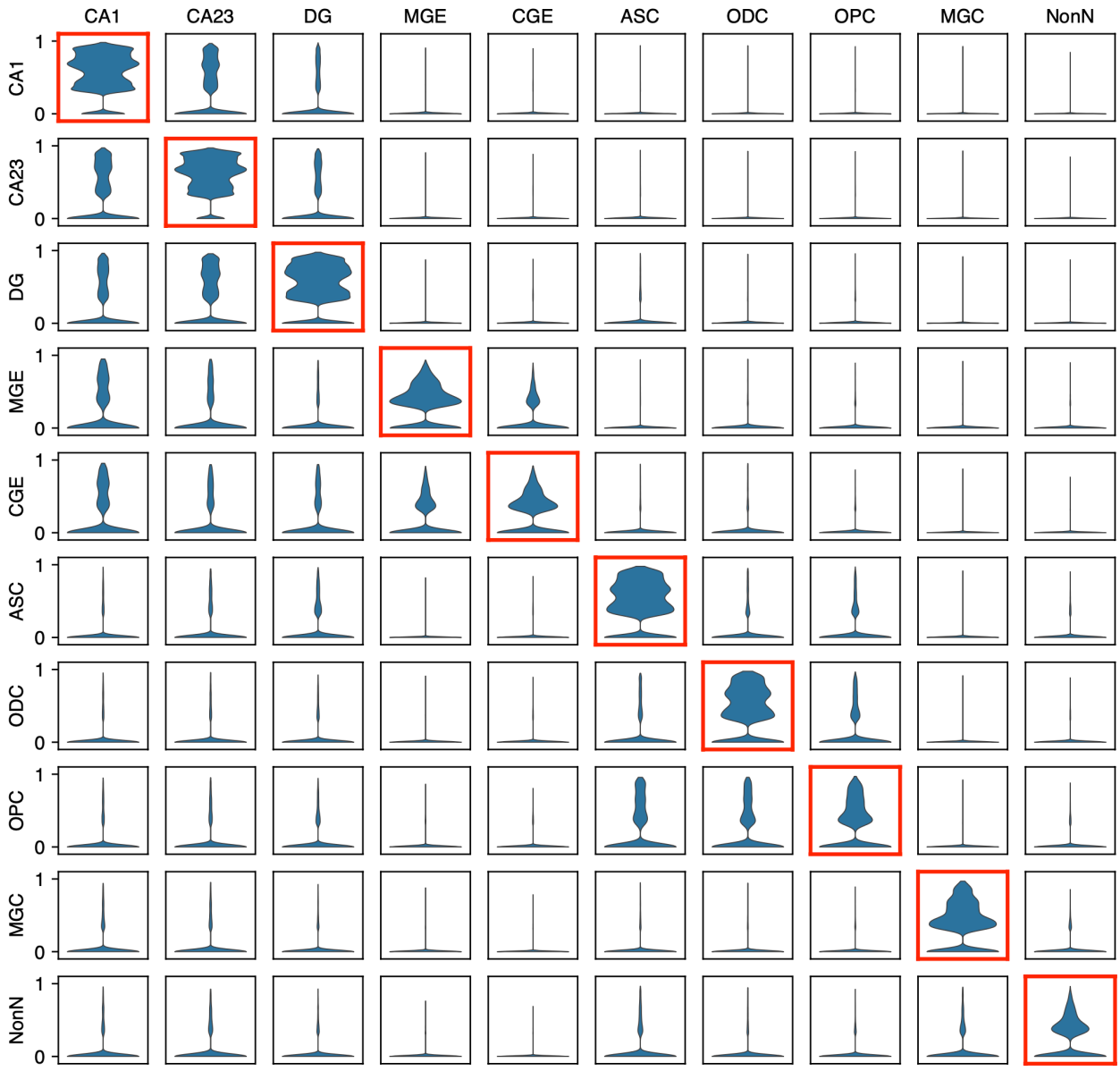


Supplementary Figure 4.1 DMR mCG fraction profiles in mouse HIP cell types.
The x-axis is centralized at DMR center with 10Kb flanking. The red box indicates the DMR is hypo-methylated in the corresponding cell type of that row.



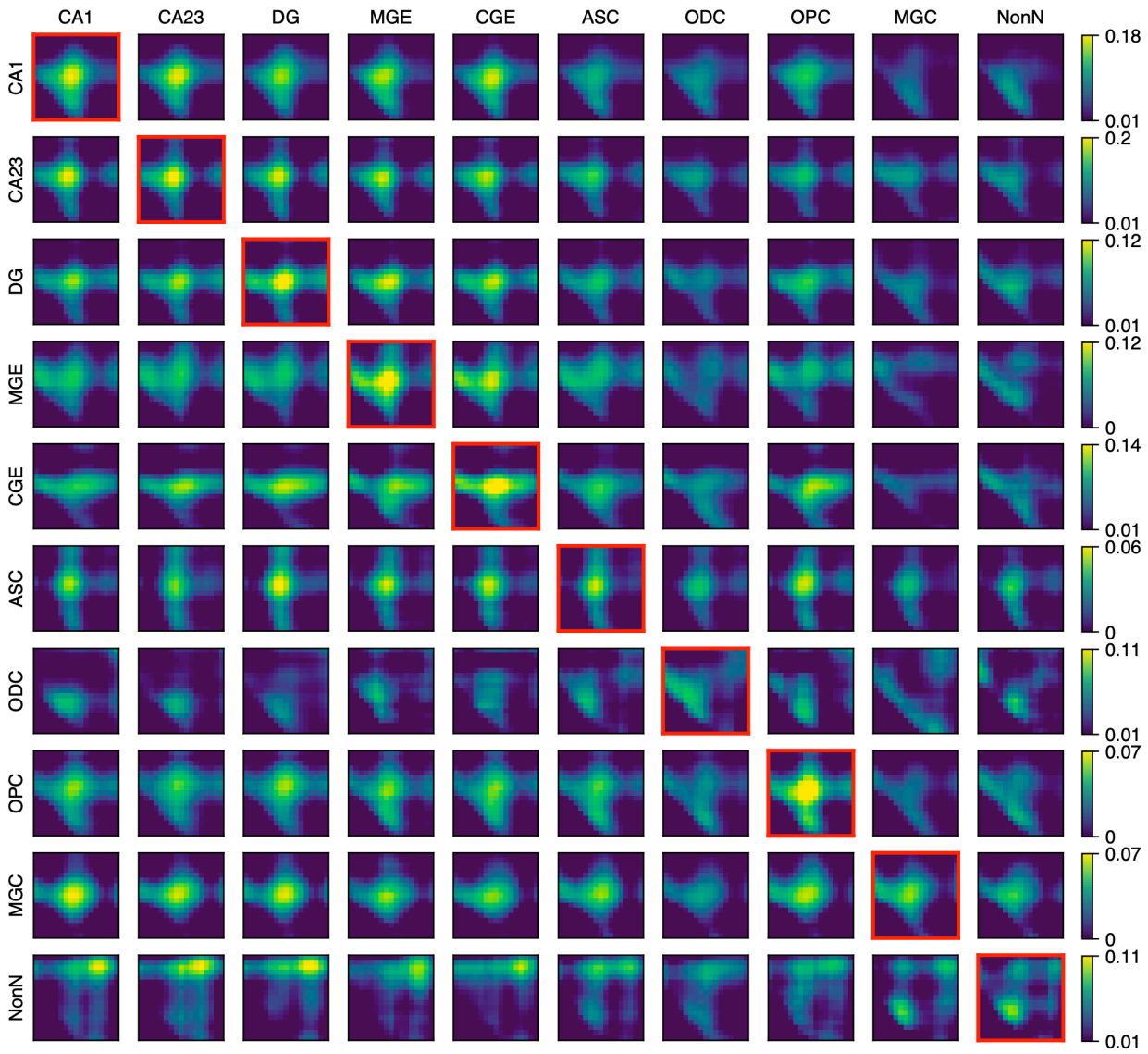
Supplementary Figure 4.2 Enhancer-overlapping DMR accessibility profiles in mouse HIP cell types.

The x-axis is centralized at DMR center with 10Kb flanking. The red box indicates the DMR is hypo-methylated in the corresponding cell type of that row.

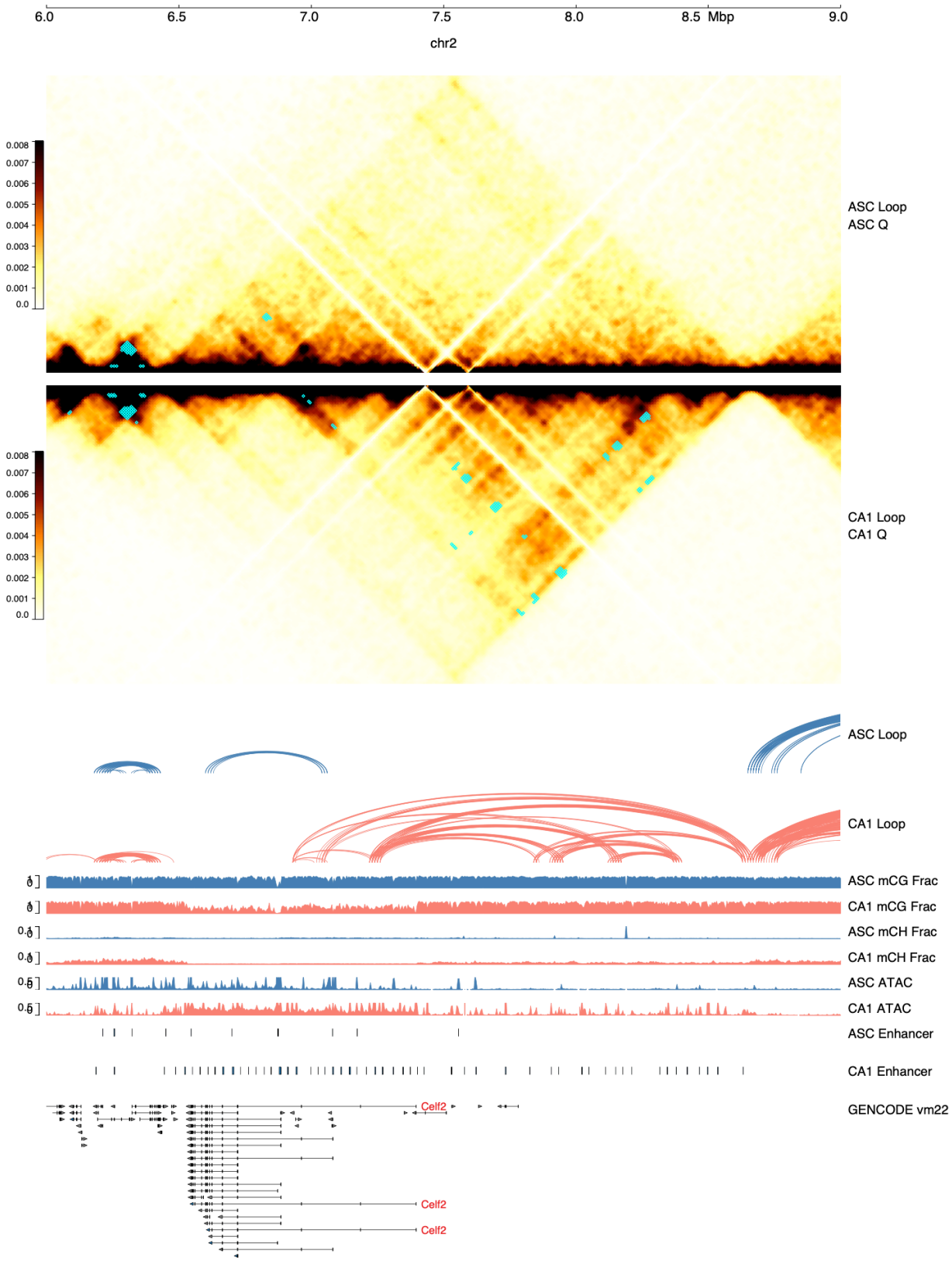


Supplementary Figure 4.3 REPTILE prediction scores of the enhancer-overlapping DMRs in mouse HIP cell types.

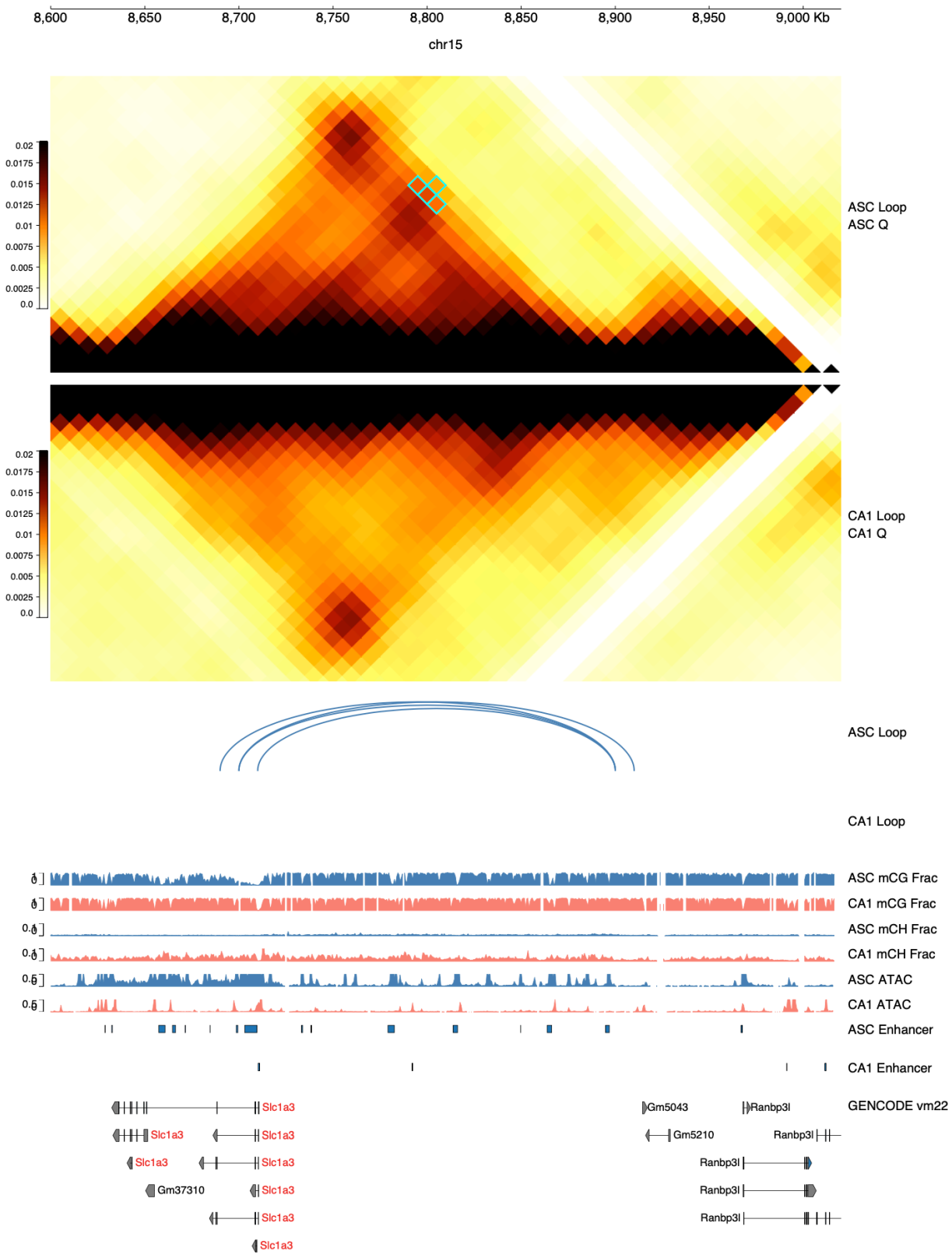
The x-axis is centralized at DMR center with 10Kb flanking. The red box indicates the DMR is hypo-methylated in the corresponding cell type of that row.



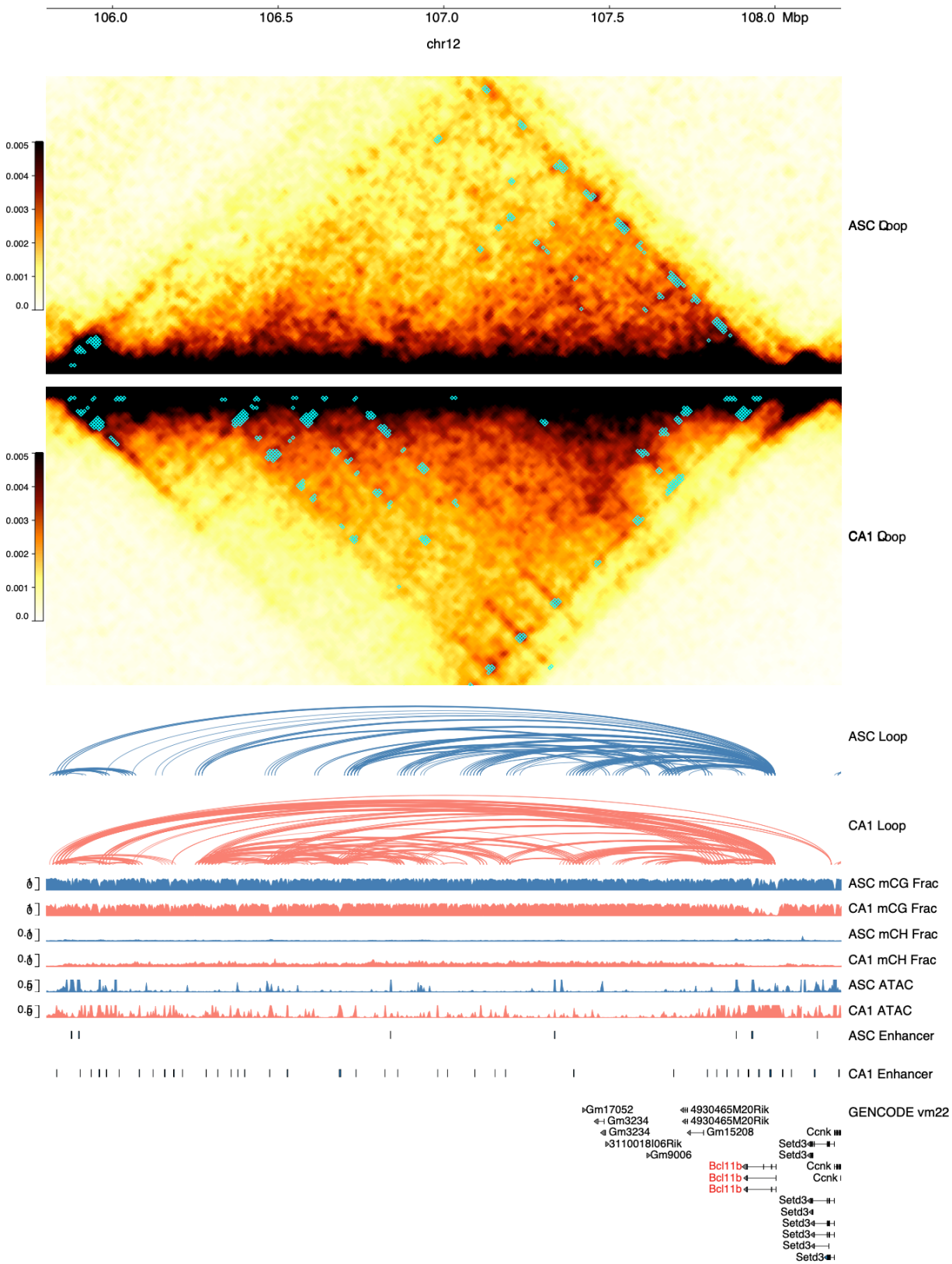
Supplementary Figure 4.4 Normalized chromatin contact strength of the enhancer-overlapping DMRs and DMGs in mouse HIP cell types.
The x-axis is centralized at DMR center with 10Kb flanking. The red box indicates the DMR and DMG is hypo-methylated in the corresponding cell type of that row.



Supplementary Figure 4.5 Genome browser tracks at the *Celf2* locus.

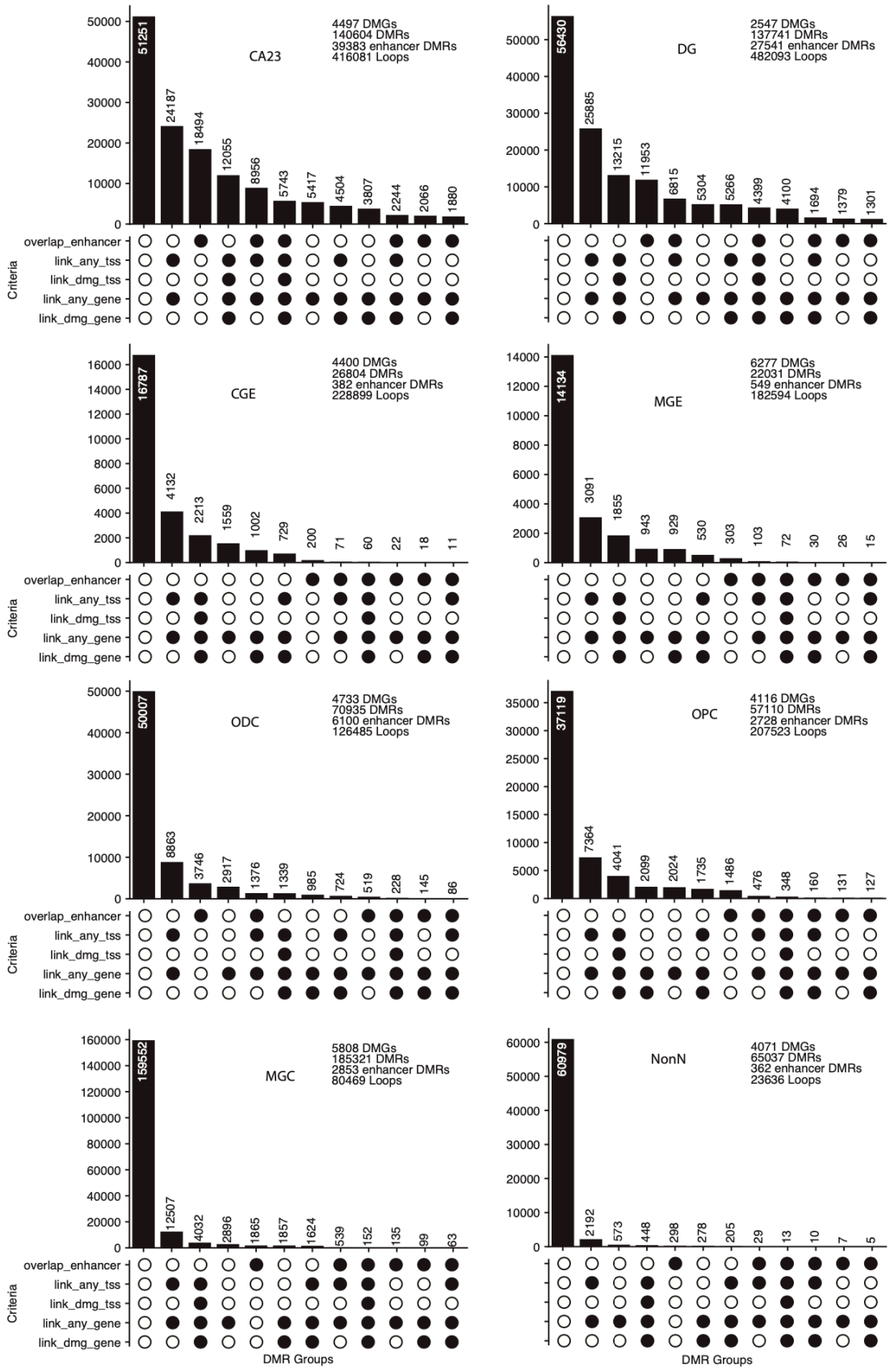


Supplementary Figure 4.6 Genome browser tracks at the *Slc1a3* locus.



Supplementary Figure 4.7 Genome browser tracks at the *Bcl11b* locus.

Supplementary Figure 4.8 Subset cell type enhancers with chromatin contact loops
Each panel shows the number of a cell type's DMRs dissected by different criteria. From top to bottom, the five criteria are: 1) Whether the DMR is overlapping with REPTILE enhancers from CA1; 2) Whether the DMR is linked to TSS by a chromatin loop; 3) Whether the DMR is link to TSS of a DMG by a chromatin loop; 4) Whether the DMR is link to any gene body region by a chromatin loop; 5) Whether the DMR is link to the gene body region of a DMG by a chromatin loop.


CA23
4497 DMGs
140604 DMRs
39383 enhancer DMRs
416081 Loops
51251
24187
18494
12055
8956
5743
5417
4504
3807
2244
2066
1880
DG
2547 DMGs
137741 DMRs
27541 enhancer DMRs
482093 Loops
56430
25885
13215
11953
6815
5304
5266
4399
4100
1694
1379
1301
CGE
4400 DMGs
26804 DMRs
382 enhancer DMRs
228899 Loops
16787
4132
2213
1559
1002
729
200
71
60
22
18
11
MGE
6277 DMGs
22031 DMRs
549 enhancer DMRs
182594 Loops
14134
3091
1855
943
929
530
303
103
72
30
26
15
ODC
4733 DMGs
70935 DMRs
6100 enhancer DMRs
126485 Loops
50007
8863
3746
2917
1376
1339
985
724
519
228
145
86
OPC
4116 DMGs
57110 DMRs
2728 enhancer DMRs
207523 Loops
37119
7364
4041
2099
2024
1735
1486
476
348
160
131
127
MGC
5808 DMGs
185321 DMRs
2853 enhancer DMRs
80469 Loops
159552
12507
4032
2896
1865
1857
1624
539
152
135
99
63
NonN
4071 DMGs
65037 DMRs
362 enhancer DMRs
23636 Loops
60979
2192
573
448
298
278
205
29
13
10
7
5
Criteria
overlap_enhancer
link_any_tss
link_dmg_tss
link_any_gene
link_dmg_gene
DMR Groups

REFERENCES

1. Luo, C., Hajkova, P. & Ecker, J. R. Dynamic DNA methylation: In the right place at the right time. *Science* **361,** 1336–1340 (2018).

2. 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).

3. 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., Margarita Behrens, M. & Ecker, J. R. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357,** 600–604 (2017).

4. 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).

5. He, Y., Hariharan, M., Gorkin, D. U., Dickel, D. E., Luo, C., Castanon, R. G., Nery, J. R., Lee, A. Y., Zhao, Y., Huang, H., Williams, B. A., Trout, D., Amrhein, H., Fang, R., Chen, H., Li, B., Visel, A., Pennacchio, L. A., Ren, B. & Ecker, J. R. Spatiotemporal DNA methylome dynamics of the developing mouse fetus. *Nature* **583,** 752–759 (2020).

6. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C. & Taipale, J. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356,** eaaj2239 (2017).

7. Hon, G. C., Rajagopal, N., Shen, Y., McCleary, D. F., Yue, F., Dang, M. D. & Ren, B. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nat. Genet.* **45,** 1198–1206 (2013).

8. Guo, J. U., Su, Y., Shin, J. H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., Zhu, H., Chang, Q., Gao, Y., Ming, G.-L. & Song, H. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci.* **17,** 215–222 (2014).

9. Lagger, S., Connelly, J. C., Schweikert, G., Webb, S., Selfridge, J., Ramsahoye, B. H., Yu, M., He, C., Sanguinetti, G., Sowers, L. C., Walkinshaw, M. D. & Bird, A. MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. *PLoS Genet.* **13,** e1006793 (2017).

10. Stroud, H., Su, S. C., Hrvatin, S., Greben, A. W., Renthal, W., Boxer, L. D., Nagy, M. A., Hochbaum, D. R., Kinde, B., Gabel, H. W. & Greenberg, M. E. Early-Life Gene Expression in Neurons Modulates Lasting Epigenetic States. *Cell* **171,** 1151-1164.e16 (2017).

11. Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U. & Zoghbi, H. Y. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23,** 185–188 (1999).

12. Guo, H., Zhu, P., Wu, X., Li, X., Wen, L. & Tang, F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23,** 2126–2135 (2013).

13. Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W. & Kelsey, G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11,** 817–820 (2014).

14. Mulqueen, R. M., Pokholok, D., Norberg, S. J., Torkenczy, K. A., Fields, A. J., Sun, D., Sinnamon, J. R., Shendure, J., Trapnell, C., O'Roak, B. J., Xia, Z., Steemers, F. J. & Adey, A. C. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36,** 428–431 (2018).

15. Liu, H., Zhou, J., Tian, W., Luo, C., Bartlett, A., Aldridge, A., Lucero, J., Osteen, J. K., Nery, J. R., Chen, H., Rivkin, A., Castanon, R. G., Clock, B., Li, Y. E., Hou, X., Poirion, O. B., Preissl, S., Pinto-Duarte, A., O'Connor, C., Boggeman, L., Fitzpatrick, C., Nunn, M., Mukamel, E. A., Zhang, Z., Callaway, E. M., Ren, B., Dixon, J. R., Behrens, M. M. & Ecker, J. R. DNA methylation atlas of the mouse brain at single-cell resolution. *Nature* **598,** 120–128 (2021).

16. Ruf-Zamojski, F., Zhang, Z., Zamojski, M., Smith, G. R., Mendelev, N., Liu, H., Nudelman, G., Moriwaki, M., Pincas, H., Castanon, R. G., Nair, V. D., Seenarine, N., Amper, M. A. S., Zhou, X., Ongaro, L., Toufaily, C., Schang, G., Nery, J. R., Bartlett, A., Aldridge, A., Jain, N., Childs, G. V., Troyanskaya, O. G., Ecker, J. R., Turgeon, J. L., Welt, C. K., Bernard, D. J. & Sealfon, S. C. Single nucleus multi-omics regulatory landscape of the murine pituitary. *Nat. Commun.* **12,** 2677 (2021).

17. Argelaguet, R., Clark, S. J., Mohammed, H., Stapel, L. C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C. W., Smallwood, S., Ibarra-Soria, X., Buettner, F., Sanguinetti, G., Xie, W., Krueger, F., Göttgens, B., Rugg-Gunn, P. J., Kelsey, G., Dean, W., Nichols, J., Stegle, O., Marioni, J. C. & Reik, W. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576,** 487–491 (2019).

18. Zhu, P., Guo, H., Ren, Y., Hou, Y., Dong, J., Li, R., Lian, Y., Fan, X., Hu, B., Gao, Y., Wang, X., Wei, Y., Liu, P., Yan, J., Ren, X., Yuan, P., Yuan, Y., Yan, Z., Wen, L., Yan, L., Qiao, J. & Tang, F. Single-cell DNA methylome sequencing of human preimplantation embryos. *Nat. Genet.* **50,** 12–19 (2018).

19. Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S. A., Ponting, C. P., Voet, T., Kelsey, G., Stegle, O. & Reik, W. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13,** 229–232 (2016).

20. Clark, S. J., Argelaguet, R., Kapourani, C.-A., Stubbs, T. M., Lee, H. J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J. C., Stegle, O. & Reik, W. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9,** 781 (2018).

21. Luo, C., Liu, H., Xie, F., Armand, E. J., Siletti, K., Bakken, T. E., Fang, R., Doyle, W. I., Stuart, T., Hodge, R. D., Hu, L., Wang, B.-A., Zhang, Z., Preissl, S., Lee, D.-S., Zhou, J., Niu, S.-Y., Castanon, R., Bartlett, A., Rivkin, A., Wang, X., Lucero, J., Nery, J. R., Davis, D. A., Mash, D. C., Satija, R., Dixon, J. R., Linnarsson, S., Lein, E., Margarita Behrens, M., Ren, B., Mukamel, E. A. & Ecker, J. R. Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. *Cell Genomics* **2,** (2022).

22. Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y. & Peng, J. Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26,** 304–319 (2016).

23. Guo, F., Li, L., Li, J., Wu, X., Hu, B., Zhu, P., Wen, L. & Tang, F. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27,** 967–988 (2017).

24. Lee, D.-S., Luo, C., Zhou, J., Chandran, S., Rivkin, A., Bartlett, A., Nery, J. R., Fitzpatrick, C., O'Connor, C., Dixon, J. R. & Ecker, J. R. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods* **16,** 999–1006 (2019).

25. Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., Kellis, M. & Ren, B. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nat. Methods* **16,** 991–993 (2019).

26. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat. Methods* **17,** 11–14 (2020).

27. Karemaker, I. D. & Vermeulen, M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends Biotechnol.* **36,** 952–965 (2018).

28. Luo, C., Rivkin, A., Zhou, J., Sandoval, J. P., Kurihara, L., Lucero, J., Castanon, R., Nery, J. R., Pinto-Duarte, A., Bui, B., Fitzpatrick, C., O'Connor, C., Ruga, S., Van Eden, M. E., Davis, D. A., Mash, D. C., Behrens, M. M. & Ecker, J. R. Robust single-cell DNA methylome profiling with snmC-seq2. *Nat. Commun.* **9,** 3824 (2018).

29. Li, Y. E., Preissl, S., Hou, X., Zhang, Z., Zhang, K., Qiu, Y., Poirion, O. B., Li, B., Chiou, J., Liu, H., Pinto-Duarte, A., Kubo, N., Yang, X., Fang, R., Wang, X., Han, J. Y., Lucero, J., Yan, Y., Miller, M., Kuan, S., Gorkin, D., Gaulton, K. J., Shen, Y., Nunn, M., Mukamel, E. A., Behrens, M. M., Ecker, J. R. & Ren, B. An atlas of gene regulatory elements in adult mouse cerebrum. *Nature* **598,** 129–136 (2021).

30. 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,** 432–439 (2018).

31. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* **358,** 69–75 (2017).

32. Ecker, J. R., Geschwind, D. H., Kriegstein, A. R., Ngai, J., Osten, P., Polioudakis, D., Regev, A., Sestan, N., Wickersham, I. R. & Zeng, H. The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. *Neuron* **96,** 542–557 (2017).

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

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. Tasic, B., Yao, Z., Graybuck, L. T., Smith, K. A., Nguyen, T. N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M. N., Viswanathan, S., Penn, O., Bakken, T., Menon, V., Miller, J., Fong, O., Hirokawa, K. E., Lathia, K., Rimorin, C., Tieu, M., Larsen, R., Casper, T., Barkan, E., Kroll, M., Parry, S., Shapovalova, N. V., Hirschstein, D., Pendergraft, J., Sullivan, H. A., Kim, T. K., Szafer, A., Dee, N., Groblewski, P., Wickersham, I., Cetin, A., Harris, J. A., Levi, B. P., Sunkin, S. M., Madisen, L., Daigle, T. L., Looger, L., Bernard, A., Phillips, J., Lein, E., Hawrylycz, M., Svoboda, K., Jones, A. R., Koch, C. & Zeng, H. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563,** 72–78 (2018).

36. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., 3rd, Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888-1902.e21 (2019).

37. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37,** 685–691 (2019).

38. Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C. & Macosko, E. Z. Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177,** 1873-1887.e17 (2019).

39. Mukamel, E. A. & Ngai, J. Perspectives on defining cell types in the brain. *Curr. Opin. Neurobiol.* **56,** 61–68 (2019).

40. Hu, Y., Huang, K., An, Q., Du, G., Hu, G., Xue, J., Zhu, X., Wang, C.-Y., Xue, Z. & Fan, G. Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17,** 88 (2016).

41. Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S.,

Ren, B., Sejnowski, T. J., Wang, W. & Ecker, J. R. Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523,** 212–216 (2015).

42. Zhang, Z., Zhou, J., Tan, P., Pang, Y., Rivkin, A. C., Kirchgessner, M. A., Williams, E., Lee, C.-T., Liu, H., Franklin, A. D., Miyazaki, P. A., Bartlett, A., Aldridge, A. I., Vu, M., Boggeman, L., Fitzpatrick, C., Nery, J. R., Castanon, R. G., Rashid, M., Jacobs, M. W., Ito-Cole, T., O'Connor, C., Pinto-Duartec, A., Dominguez, B., Smith, J. B., Niu, S.-Y., Lee, K.-F., Jin, X., Mukamel, E. A., Behrens, M. M., Ecker, J. R. & Callaway, E. M. Epigenomic diversity of cortical projection neurons in the mouse brain. *Nature* **598,** 167–173 (2021).

43. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18,** 1333–1341 (2021).

44. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19,** 15 (2018).

45. Granja, J. M., Corces, M. R., Pierce, S. E., Bagdatli, S. T., Choudhry, H., Chang, H. Y. & Greenleaf, W. J. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* (2021). doi:10.1038/s41588-021-00790-6

46. Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A. K., Zhou, X., Xie, F., Mukamel, E. A., Zhang, K., Zhang, Y., Behrens, M. M., Ecker, J. R. & Ren, B. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* **12,** 1337 (2021).

47. Danese, A., Richter, M. L., Chaichoompu, K., Fischer, D. S., Theis, F. J. & Colomé-Tatché, M. EpiScanpy: integrated single-cell epigenomic analysis. *Nat. Commun.* **12,** 5228 (2021).

48. Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C. & Stegle, O. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21,** 111 (2020).

49. Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M. & Theis, F. J. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19,** 41–50 (2022).

50. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37,** 547–554 (2019).

51. Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. *Genome Biol.* **23,** 42 (2022).

52. Korsunsky, I., Millard, N., 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. *Nat. Methods* **16,** 1289–1296 (2019).

53. He, Y., Gorkin, D. U., Dickel, D. E., Nery, J. R., Castanon, R. G., Lee, A. Y., Shen, Y., Visel, A., Pennacchio, L. A., Ren, B. & Ecker, J. R. Improved regulatory element prediction based

on tissue-specific local epigenomic signatures. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E1633–E1640 (2017).

54. Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S. & Engreitz, J. M. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51,** 1664–1669 (2019).

55. Wang, Q., Ding, S.-L., Li, Y., Royall, J., Feng, D., Lesnar, P., Graddis, N., Naeemi, M., Facer, B., Ho, A., Dolbeare, T., Blanchard, B., Dee, N., Wakeman, W., Hirokawa, K. E., Szafer, A., Sunkin, S. M., Oh, S. W., Bernard, A., Phillips, J. W., Hawrylycz, M., Koch, C., Zeng, H., Harris, J. A. & Ng, L. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181,** 936-953.e20 (2020).

56. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018). at <http://arxiv.org/abs/1802.03426>

57. Ming, G.-L. & Song, H. Adult neurogenesis in the mammalian brain: significant answers and significant questions. *Neuron* **70,** 687–702 (2011).

58. Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R. S., Aldridge, A. I., Ament, S. A., Bartlett, A., Behrens, M. M., Van den Berge, K., Bertagnolli, D., de Bézieux, H. R., Biancalani, T., Booeshaghi, A. S., Bravo, H. C., Casper, T., Colantuoni, C., Crabtree, J., Creasy, H., Crichton, K., Crow, M., Dee, N., Dougherty, E. L., Doyle, W. I., Dudoit, S., Fang, R., Felix, V., Fong, O., Giglio, M., Goldy, J., Hawrylycz, M., Herb, B. R., Hertzano, R., Hou, X., Hu, Q., Kancherla, J., Kroll, M., Lathia, K., Li, Y. E., Lucero, J. D., Luo, C., Mahurkar, A., McMillen, D., Nadaf, N. M., Nery, J. R., Nguyen, T. N., Niu, S.-Y., Ntranos, V., Orvis, J., Osteen, J. K., Pham, T., Pinto-Duarte, A., Poirion, O., Preissl, S., Purdom, E., Rimorin, C., Risso, D., Rivkin, A. C., Smith, K., Street, K., Sulc, J., Svensson, V., Tieu, M., Torkelson, A., Tung, H., Vaishnav, E. D., Vanderburg, C. R., van Velthoven, C., Wang, X., White, O. R., Huang, Z. J., Kharchenko, P. V., Pachter, L., Ngai, J., Regev, A., Tasic, B., Welch, J. D., Gillis, J., Macosko, E. Z., Ren, B., Ecker, J. R., Zeng, H. & Mukamel, E. A. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature* **598,** 103–110 (2021).

59. Huang, Z. J. & Paul, A. The diversity of GABAergic neurons and neural communication elements. *Nat. Rev. Neurosci.* **20,** 563–572 (2019).

60. Arendt, D., Musser, J. M., Baker, C. V. H., Bergman, A., Cepko, C., Erwin, D. H., Pavlicev, M., Schlosser, G., Widder, S., Laubichler, M. D. & Wagner, G. P. The origin and evolution of cell types. *Nat. Rev. Genet.* **17,** 744–757 (2016).

61. Smith, J. B., Alloway, K. D., Hof, P. R., Orman, R., Reser, D. H., Watakabe, A. & Watson, G. D. R. The relationship between the claustrum and endopiriform nucleus: A perspective towards consensus on cross-species homology. *J. Comp. Neurol.* **527,** 476–499 (2019).

62. Crick, F. C. & Koch, C. What is the function of the claustrum? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **360,** 1271–1279 (2005).

63. Hrvatin, S., Tzeng, C. P., Nagy, M. A., Stroud, H., Koutsioumpa, C., Wilcox, O. F., Assad, E. G., Green, J., Harvey, C. D., Griffith, E. C. & Greenberg, M. E. A scalable platform for the development of cell-type-specific viral drivers. *Elife* **8,** (2019).

64. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* **164,** 1110–1121 (2016).

65. Ferland, R. J., Cherry, T. J., Preware, P. O., Morrisey, E. E. & Walsh, C. A. Characterization of Foxp2 and Foxp1 mRNA and protein in the developing and mature brain. *J. Comp. Neurol.* **460,** 266–279 (2003).

66. Siddiqui, T. J., Tari, P. K., Connor, S. A., Zhang, P., Dobie, F. A., She, K., Kawabe, H., Wang, Y. T., Brose, N. & Craig, A. M. An LRRTM4-HSPG complex mediates excitatory synapse development on dentate gyrus granule cells. *Neuron* **79,** 680–695 (2013).

67. Yao, Z., van Velthoven, C. T. J., Nguyen, T. N., Goldy, J., Sedeno-Cortes, A. E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., Ding, S.-L., Fong, O., Garren, E., Glandon, A., Gouwens, N. W., Gray, J., Graybuck, L. T., Hawrylycz, M. J., Hirschstein, D., Kroll, M., Lathia, K., Lee, C., Levi, B., McMillen, D., Mok, S., Pham, T., Ren, Q., Rimorin, C., Shapovalova, N., Sulc, J., Sunkin, S. M., Tieu, M., Torkelson, A., Tung, H., Ward, K., Dee, N., Smith, K. A., Tasic, B. & Zeng, H. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* (2021). doi:10.1016/j.cell.2021.04.021

68. Nieto, M., Monuki, E. S., Tang, H., Imitola, J., Haubst, N., Khoury, S. J., Cunningham, J., Gotz, M. & Walsh, C. A. Expression of Cux-1 and Cux-2 in the subventricular zone and upper layers II-IV of the cerebral cortex. *J. Comp. Neurol.* **479,** 168–180 (2004).

69. Hochgerner, H., Zeisel, A., Lönnerberg, P. & Linnarsson, S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat. Neurosci.* **21,** 290–299 (2018).

70. O'Leary, D. D. M., Chou, S.-J. & Sahara, S. Area patterning of the mammalian cortex. *Neuron* **56,** 252–269 (2007).

71. Zhang, T.-Y., Keown, C. L., Wen, X., Li, J., Vousden, D. A., Anacker, C., Bhattacharyya, U., Ryan, R., Diorio, J., O'Toole, N., Lerch, J. P., Mukamel, E. A. & Meaney, M. J. Environmental enrichment increases transcriptional and epigenetic differentiation between mouse dorsal and ventral dentate gyrus. *Nat. Commun.* **9,** 298 (2018).

72. Szulwach, K. E., Li, X., Li, Y., Song, C.-X., Wu, H., Dai, Q., Irier, H., Upadhyay, A. K., Gearing, M., Levey, A. I., Vasanthakumar, A., Godley, L. A., Chang, Q., Cheng, X., He, C. & Jin, P. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat. Neurosci.* **14,** 1607–1616 (2011).

73. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9,** 9354 (2019).

74. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9,** 5233 (2019).

75. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **46,** 389–422 (2002).

76. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. in *2010 20th International Conference on Pattern Recognition* 3121–3124 (2010).

77. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **18,** 1–5 (2017).

78. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9,** 2579–2605 (2008).

79. Linderman, G. C., Rachh, M., Hoskins, J. G., Steinerberger, S. & Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat. Methods* **16,** 243–245 (2019).

80. Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L. E., La Manno, G., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U. & Linnarsson, S. Molecular Architecture of the Mouse Nervous System. *Cell* **174,** 999-1014.e22 (2018).

81. Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., Boguski, M. S., Brockway, K. S., Byrnes, E. J., Chen, L., Chen, L., Chen, T.-M., Chin, M. C., Chong, J., Crook, B. E., Czaplinska, A., Dang, C. N., Datta, S., Dee, N. R., Desaki, A. L., Desta, T., Diep, E., Dolbeare, T. A., Donelan, M. J., Dong, H.-W., Dougherty, J. G., Duncan, B. J., Ebbert, A. J., Eichele, G., Estin, L. K., Faber, C., Facer, B. A., Fields, R., Fischer, S. R., Fliss, T. P., Frensley, C., Gates, S. N., Glattfelder, K. J., Halverson, K. R., Hart, M. R., Hohmann, J. G., Howell, M. P., Jeung, D. P., Johnson, R. A., Karr, P. T., Kawal, R., Kidney, J. M., Knapik, R. H., Kuan, C. L., Lake, J. H., Laramee, A. R., Larsen, K. D., Lau, C., Lemon, T. A., Liang, A. J., Liu, Y., Luong, L. T., Michaels, J., Morgan, J. J., Morgan, R. J., Mortrud, M. T., Mosqueda, N. F., Ng, L. L., Ng, R., Orta, G. J., Overly, C. C., Pak, T. H., Parry, S. E., Pathak, S. D., Pearson, O. C., Puchalski, R. B., Riley, Z. L., Rockett, H. R., Rowland, S. A., Royall, J. J., Ruiz, M. J., Sarno, N. R., Schaffnit, K., Shapovalova, N. V., Sivisay, T., Slaughterbeck, C. R., Smith, S. C., Smith, K. A., Smith, B. I., Sodt, A. J., Stewart, N. N., Stumpf, K.-R., Sunkin, S. M., Sutram, M., Tam, A., Teemer, C. D., Thaller, C., Thompson, C. L., Varnam, L. R., Visel, A., Whitlock, R. M., Wohnoutka, P. E., Wolkey, C. K., Wong, V. Y., Wood, M., Yaylaoglu, M. B., Young, R. C., Youngstrom, B. L., Yuan, X. F., Zhang, B., Zwingman, T. A. & Jones, A. R. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445,** 168–176 (2007).

82. 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).

83. Suzuki, R. & Shimodaira, H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22,** 1540–1542 (2006).

84. Hodge, R. D., Bakken, T. E., Miller, J. A., Smith, K. A., Barkan, E. R., Graybuck, L. T., Close, J. L., Long, B., Johansen, N., Penn, O., Yao, Z., Eggermont, J., Höllt, T., Levi, B. P., Shehata, S. I., Aevermann, B., Beller, A., Bertagnolli, D., Brouner, K., Casper, T., Cobbs, C., Dalley, R., Dee, N., Ding, S.-L., Ellenbogen, R. G., Fong, O., Garren, E., Goldy, J., Gwinn, R. P., Hirschstein, D., Keene, C. D., Keshk, M., Ko, A. L., Lathia, K., Mahfouz, A., Maltzer, Z., McGraw, M., Nguyen, T. N., Nyhus, J., Ojemann, J. G., Oldre, A., Parry, S., Reynolds, S., Rimorin, C., Shapovalova, N. V., Somasundaram, S., Szafer, A., Thomsen, E. R., Tieu, M., Quon, G., Scheuermann, R. H., Yuste, R., Sunkin, S. M., Lelieveldt, B., Feng, D., Ng, L., Bernard, A., Hawrylycz, M., Phillips, J. W., Tasic, B., Zeng, H., Jones, A. R., Koch, C. & Lein, E. S. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573,** 61–68 (2019).

85. Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W. & Mathelier, A. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48,** D87–D92 (2020).

86. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27,** 1017–1018 (2011).

87. Wingender, E., Schoeps, T., Haubrock, M., Krull, M. & Dönitz, J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic Acids Res.* **46,** D343–D347 (2018).

88. 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).

89. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat. Protoc.* **10,** 475–483 (2015).

90. Gorkin, D. U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A. Y., Li, B., Chiou, J., Wildberg, A., Ding, B., Zhang, B., Wang, M., Strattan, J. S., Davidson, J. M., Qiu, Y., Afzal, V., Akiyama, J. A., Plajzer-Frick, I., Novak, 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., Han, J. Y., Williams, B. A., Trout, D., Amrhein, H., Yang, H., Cherry, J. M., Wang, W., Gaulton, K., Ecker, J. R., Shen, Y., Dickel, D. E., Visel, A., Pennacchio, L. A. & Ren, B. An

atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583,** 744–751 (2020).

91. Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T. J., Dixon, J. R. & Ecker, J. R. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 14011–14018 (2019).

92. Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P. & Tang, H. GOATOOLS: A Python library for Gene Ontology analyses. *Sci. Rep.* **8,** 10872 (2018).

93. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33,** 1029–1047 (2013).

94. Zhang, H., Emerson, D. J., Gilgenast, T. G., Titus, K. R., Lan, Y., Huang, P., Zhang, D., Wang, H., Keller, C. A., Giardine, B., Hardison, R. C., Phillips-Cremins, J. E. & Blobel, G. A. Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature* **576,** 158–162 (2019).

95. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. (2014). at <https://dl.acm.org/doi/abs/10.5555/2627435.2670313>

96. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv [cs.DC]* (2016). at <http://arxiv.org/abs/1603.04467>

97. Dudek, S. M., Alexander, G. M. & Farris, S. Rediscovering area CA2: unique properties and functions. *Nat. Rev. Neurosci.* **17,** 89–102 (2016).

98. San Antonio, A., Liban, K., Ikrar, T., Tsyganovskiy, E. & Xu, X. Distinct physiological and developmental properties of hippocampal CA2 subfield revealed by using anti-Purkinje cell protein 4 (PCP4) immunostaining. *J. Comp. Neurol.* **522,** 1333–1354 (2014).

99. Fukaya, M., Yamazaki, M., Sakimura, K. & Watanabe, M. Spatial diversity in gene expression for VDCCgamma subunit family in developing and adult mouse brains. *Neurosci. Res.* **53,** 376–383 (2005).

100. Phillips, H. S., Hains, J. M., Laramee, G. R., Rosenthal, A. & Winslow, J. W. Widespread expression of BDNF but not NT3 by target areas of basal forebrain cholinergic neurons. *Science* **250,** 290–294 (1990).

101. Adamek, G. D., Shipley, M. T. & Sanders, M. S. The indusium griseum in the mouse: architecture, Timm's histochemistry and some afferent connections. *Brain Res. Bull.* **12,** 657–668 (1984).

102. Heimer, L. & Wilson, R. D. The subcortical projections of the allocortex: similarities in the neural connections of the hippocampus, the piriform cortex and the neocortex. Santini M, editor. *Perspectives in neurobiology* (1975).

103. Voorn, P., Vanderschuren, L. J. M. J., Groenewegen, H. J., Robbins, T. W. & Pennartz, C. M. A. Putting a spin on the dorsal-ventral divide of the striatum. *Trends Neurosci.* **27,** 468–474 (2004).

104. Smith, J. B., Klug, J. R., Ross, D. L., Howard, C. D., Hollon, N. G., Ko, V. I., Hoffman, H., Callaway, E. M., Gerfen, C. R. & Jin, X. Genetic-Based Dissection Unveils the Inputs and Outputs of Striatal Patch and Matrix Compartments. *Neuron* **91,** 1069–1084 (2016).

105. Ross, S. E., Greenberg, M. E. & Stiles, C. D. Basic helix-loop-helix factors in cortical development. *Neuron* **39,** 13–25 (2003).

106. Snyder, J. S., Radik, R., Wojtowicz, J. M. & Cameron, H. A. Anatomical gradients of adult neurogenesis and activity: young neurons in the ventral dentate gyrus are activated by water maze training. *Hippocampus* **19,** 360–370 (2009).

107. Bekiari, C., Grivas, I., Tsingotjidou, A. & Papadopoulos, G. C. Adult neurogenesis and gliogenesis in the dorsal and ventral canine hippocampus. *J. Comp. Neurol.* **528,** 1216–1230 (2020).

108. Bayraktar, O. A., Fuentealba, L. C., Alvarez-Buylla, A. & Rowitch, D. H. Astrocyte development and heterogeneity. *Cold Spring Harb. Perspect. Biol.* **7,** a020362 (2014).

109. Floriddia, E. M., Zhang, S. & van Bruggen, D. Distinct oligodendrocyte populations have different spatial distributions and injury-specific responses. *bioRxiv* (2019). at <https://www.biorxiv.org/content/10.1101/580985v2.abstract>

110. Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P. & Sandberg, R. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30,** 777–782 (2012).

111. Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G. & Sandberg, R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013).

112. Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B. & Ecker, J. R. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322 (2009).

113. Pott, S. Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife* **6,** (2017).

114. Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P. & Jones, P. A. Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22,** 2497–2506 (2012).

115. Ellis, B. C., Molloy, P. L. & Graham, L. D. CRNDE: A Long Non-Coding RNA Involved in CanceR, Neurobiology, and DEvelopment. *Front. Genet.* **3,** 270 (2012).

116. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36,** 421–427 (2018).

117. Nordström, K. J. V., Schmidt, F., Gasparoni, N., Salhab, A., Gasparoni, G., Kattler, K., Müller, F., Ebert, P., Costa, I. G., DEEP consortium, Pfeifer, N., Lengauer, T., Schulz, M. H. & Walter, J. Unique and assay specific features of NOMe-, ATAC- and DNase I-seq data. *Nucleic Acids Res.* **47,** 10580–10596 (2019).

118. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9,** 884 (2018).

119. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36,** 411–420 (2018).

120. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* (2019). doi:10.1038/s41576-019-0093-7

121. Luo, C., Lancaster, M. A., Castanon, R., Nery, J. R., Knoblich, J. A. & Ecker, J. R. Cerebral Organoids Recapitulate Epigenomic Signatures of the Human Fetal Brain. *Cell Rep.* **17,** 3369–3384 (2016).

122. Kozlenkov, A., Li, J., Apontes, P., Hurd, Y. L., Byne, W. M., Koonin, E. V., Wegner, M., Mukamel, E. A. & Dracheva, S. A unique role for DNA (hydroxy)methylation in epigenetic regulation of human inhibitory neurons. *Sci Adv* **4,** eaau6190 (2018).

123. Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T. J., Dixon, J. R. & Ecker, J. R. Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation. *Proc. Natl. Acad. Sci. U. S. A.* **116,** 14011–14018 (2019).

124. Koopmans, F., van Nierop, P., Andres-Alonso, M., Byrnes, A., Cijsouw, T., Coba, M. P., Cornelisse, L. N., Farrell, R. J., Goldschmidt, H. L., Howrigan, D. P., Hussain, N. K., Imig, C., de Jong, A. P. H., Jung, H., Kohansalnodehi, M., Kramarz, B., Lipstein, N., Lovering, R. C., MacGillavry, H., Mariano, V., Mi, H., Ninov, M., Osumi-Sutherland, D., Pielot, R., Smalla, K.-H., Tang, H., Tashman, K., Toonen, R. F. G., Verpelli, C., Reig-Viader, R., Watanabe, K., van Weering, J., Achsel, T., Ashrafi, G., Asi, N., Brown, T. C., De Camilli, P.,

Feuermann, M., Foulger, R. E., Gaudet, P., Joglekar, A., Kanellopoulos, A., Malenka, R., Nicoll, R. A., Pulido, C., de Juan-Sanz, J., Sheng, M., Südhof, T. C., Tilgner, H. U., Bagni, C., Bayés, À., Biederer, T., Brose, N., Chua, J. J. E., Dieterich, D. C., Gundelfinger, E. D., Hoogenraad, C., Huganir, R. L., Jahn, R., Kaeser, P. S., Kim, E., Kreutz, M. R., McPherson, P. S., Neale, B. M., O'Connor, V., Posthuma, D., Ryan, T. A., Sala, C., Feng, G., Hyman, S. E., Thomas, P. D., Smit, A. B. & Verhage, M. SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103,** 217-234.e4 (2019).

125. Deneris, E. S. & Hobert, O. Maintenance of postmitotic neuronal cell identity. *Nat. Neurosci.* **17,** 899–907 (2014).

126. Kepecs, A. & Fishell, G. Interneuron cell types are fit to function. *Nature* **505,** 318–326 (2014).

127. Boldog, E., Bakken, T. E., Hodge, R. D., Novotny, M., Aevermann, B. D., Baka, J., Bordé, S., Close, J. L., Diez-Fuertes, F., Ding, S.-L., Faragó, N., Kocsis, Á. K., Kovács, B., Maltzer, Z., McCorrison, J. M., Miller, J. A., Molnár, G., Oláh, G., Ozsvár, A., Rózsa, M., Shehata, S. I., Smith, K. A., Sunkin, S. M., Tran, D. N., Venepally, P., Wall, A., Puskás, L. G., Barzó, P., Steemers, F. J., Schork, N. J., Scheuermann, R. H., Lasken, R. S., Lein, E. S. & Tamás, G. Transcriptomic and morphophysiological evidence for a specialized human cortical GABAergic cell type. *Nat. Neurosci.* **21,** 1185–1195 (2018).

128. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R. & Weirauch, M. T. The Human Transcription Factors. *Cell* **172,** 650–665 (2018).

129. Castro-Mondragon, J. A., Jaeger, S., Thieffry, D., Thomas-Chollier, M. & van Helden, J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.* **45,** e119 (2017).

130. Paul, A., Crow, M., Raudales, R., He, M., Gillis, J. & Huang, Z. J. Transcriptional Architecture of Synaptic Communication Delineates GABAergic Neuron Identity. *Cell* **171,** 522-539.e20 (2017).

131. Piper, M., Barry, G., Harvey, T. J., McLeay, R., Smith, A. G., Harris, L., Mason, S., Stringer, B. W., Day, B. W., Wray, N. R., Gronostajski, R. M., Bailey, T. L., Boyd, A. W. & Richards, L. J. NFIB-mediated repression of the epigenetic factor Ezh2 regulates cortical development. *J. Neurosci.* **34,** 2921–2930 (2014).

132. Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., Tiwari, V. K. & Schübeler, D. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480,** 490–495 (2011).

133. Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T.-Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A. & Meissner, A. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500,** 477–481 (2013).

134. Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., Shin, J., Cox, E., Rho, H. S., Woodard, C., Xia, S., Liu, S., Lyu, H., Ming, G.-L., Wade, H., Song, H., Qian, J. & Zhu, H. DNA methylation presents distinct binding sites for human transcription factors. *Elife* **2,** e00726 (2013).

135. Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, RACI Consortium, Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., Okada, Y., Raychaudhuri, S., Daly, M. J., Patterson, N., Neale, B. M. & Price, A. L. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).

136. 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).

137. Skene, N. G., Bryois, J., Bakken, T. E., Breen, G., Crowley, J. J., Gaspar, H. A., Giusti-Rodriguez, P., Hodge, R. D., Miller, J. A., Muñoz-Manchado, A. B., O'Donovan, M. C., Owen, M. J., Pardiñas, A. F., Ryge, J., Walters, J. T. R., Linnarsson, S., Lein, E. S., Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Sullivan, P. F. & Hjerling-Leffler, J. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50,** 825–833 (2018).

138. Birnbaum, R. & Weinberger, D. R. Genetic insights into the neurodevelopmental origins of schizophrenia. *Nat. Rev. Neurosci.* **18,** 727–740 (2017).

139. Calderon, D., Bhaskar, A., Knowles, D. A., Golan, D., Raj, T., Fu, A. Q. & Pritchard, J. K. Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. *Am. J. Hum. Genet.* **101,** 686–699 (2017).

140. Bhaduri, A., Andrews, M. G., Mancia Leon, W., Jung, D., Shin, D., Allen, D., Jung, D., Schmunk, G., Haeussler, M., Salma, J., Pollen, A. A., Nowakowski, T. J. & Kriegstein, A. R. Cell stress in cortical organoids impairs molecular subtype specification. *Nature* **578,** 142–148 (2020).

141. Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C. & Shendure, J. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* (2018). doi:10.1126/science.aau0730

142. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0290-0

143. Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M. M., Hu, M. & Ren, B. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26,** 1063–1070 (2019).

144. Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y.-C., Regev, A. & Buenrostro, J. D. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183,** 1103-1116.e20 (2020).

145. Cadwell, C. R., Palasantza, A., Jiang, X., Berens, P., Deng, Q., Yilmaz, M., Reimer, J., Shen, S., Bethge, M., Tolias, K. F., Sandberg, R. & Tolias, A. S. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nat. Biotechnol.* **34,** 199–203 (2016).

146. Fuzik, J., Zeisel, A., Máté, Z., Calvigioni, D., Yanagawa, Y., Szabó, G., Linnarsson, S. & Harkany, T. Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. *Nat. Biotechnol.* **34,** 175–183 (2016).

147. Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A. J. M., Faridani, O. R. & Sandberg, R. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38,** 708–714 (2020).

148. Greenberg, M. M. Abasic and oxidized abasic site reactivity in DNA: enzyme inhibition, cross-linking, and nucleosome catalyzed reactions. *Acc. Chem. Res.* **47,** 646–655 (2014).

149. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27,** 1571–1572 (2011).

150. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

151. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

152. 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 Biol.* **9,** R137 (2008).

153. Fang, R., Preissl, S., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A. K., Mukamel, E. A., Zhang, Y., Margarita Behrens, M., Ecker, J. & Ren, B. Fast and Accurate Clustering of Single Cell Epigenomes Reveals Cis-Regulatory Elements in Rare Cell Types. *bioRxiv* 615179 (2019). doi:10.1101/615179

154. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

155. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

156. Libertini, E., Lebreton, A., Lakisic, G., Dillies, M.-A., Beck, S., Coppée, J.-Y., Cossart, P. & Bierne, H. Overexpression of the Heterochromatinization Factor BAHD1 in HEK293 Cells Differentially Reshapes the DNA Methylome on Autosomes and X Chromosome. *Front. Genet.* **6,** 339 (2015).

157. Aktaş, T., Avşar Ilık, İ., Maticzka, D., Bhardwaj, V., Pessoa Rodrigues, C., Mittler, G., Manke, T., Backofen, R. & Akhtar, A. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* **544,** 115–119 (2017).

158. Endersby, J. Lumpers and splitters: Darwin, Hooker, and the search for order. *Science* **326,** 1496–1499 (2009).

159. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer Science & Business Media, 2009).

160. van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., Bierie, B., Mazutis, L., Wolf, G., Krishnaswamy, S. & Pe'er, D. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174,** 716-729.e27 (2018).

161. Gao, C., Preissl, S., Luo, C., Castanon, R., Sandoval, J., Rivkin, A., Nery, J. R., Behrens, M. M., Ecker, J. R., Ren, B. & Welch, J. D. Iterative Refinement of Cellular Identity from Single-Cell Data Using Online Learning. *bioRxiv* 2020.01.16.909861 (2020). doi:10.1101/2020.01.16.909861

162. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11,** 165 (2010).

163. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F. & Mathelier, A. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46,** D260–D266 (2018).

164. Martin, J., Walters, R. K., Demontis, D., Mattheisen, M., Lee, S. H., Robinson, E., Brikell, I., Ghirardi, L., Larsson, H., Lichtenstein, P., Eriksson, N., 23andMe Research Team,

Psychiatric Genomics Consortium: ADHD Subgroup, iPSYCH–Broad ADHD Workgroup, Werge, T., Mortensen, P. B., Pedersen, M. G., Mors, O., Nordentoft, M., Hougaard, D. M., Bybjerg-Grauholm, J., Wray, N. R., Franke, B., Faraone, S. V., O'Donovan, M. C., Thapar, A., Børglum, A. D. & Neale, B. M. A Genetic Investigation of Sex Bias in the Prevalence of Attention-Deficit/Hyperactivity Disorder. *Biol. Psychiatry* **83,** 1044–1053 (2018).

165. Pappa, I., St Pourcain, B., Benke, K., Cavadino, A., Hakulinen, C., Nivard, M. G., Nolte, I. M., Tiesler, C. M. T., Bakermans-Kranenburg, M. J., Davies, G. E., Evans, D. M., Geoffroy, M.-C., Grallert, H., Groen-Blokhuis, M. M., Hudziak, J. J., Kemp, J. P., Keltikangas-Järvinen, L., McMahon, G., Mileva-Seitz, V. R., Motazedi, E., Power, C., Raitakari, O. T., Ring, S. M., Rivadeneira, F., Rodriguez, A., Scheet, P. A., Seppälä, I., Snieder, H., Standl, M., Thiering, E., Timpson, N. J., Veenstra, R., Velders, F. P., Whitehouse, A. J. O., Smith, G. D., Heinrich, J., Hypponen, E., Lehtimäki, T., Middeldorp, C. M., Oldehinkel, A. J., Pennell, C. E., Boomsma, D. I. & Tiemeier, H. A genome-wide approach to children's aggressive behavior: The EAGLE consortium. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **171,** 562–572 (2016).

166. Watson, H. J., Yilmaz, Z., Thornton, L. M., Hübel, C., Coleman, J. R. I., Gaspar, H. A., Bryois, J., Hinney, A., Leppä, V. M., Mattheisen, M., Medland, S. E., Ripke, S., Yao, S., Giusti-Rodríguez, P., Anorexia Nervosa Genetics Initiative, Hanscombe, K. B., Purves, K. L., Eating Disorders Working Group of the Psychiatric Genomics Consortium, Adan, R. A. H., Alfredsson, L., Ando, T., Andreassen, O. A., Baker, J. H., Berrettini, W. H., Boehm, I., Boni, C., Perica, V. B., Buehren, K., Burghardt, R., Cassina, M., Cichon, S., Clementi, M., Cone, R. D., Courtet, P., Crow, S., Crowley, J. J., Danner, U. N., Davis, O. S. P., de Zwaan, M., Dedoussis, G., Degortes, D., DeSocio, J. E., Dick, D. M., Dikeos, D., Dina, C., Dmitrzak-Weglarz, M., Docampo, E., Duncan, L. E., Egberts, K., Ehrlich, S., Escaramís, G., Esko, T., Estivill, X., Farmer, A., Favaro, A., Fernández-Aranda, F., Fichter, M. M., Fischer, K., Föcker, M., Foretova, L., Forstner, A. J., Forzan, M., Franklin, C. S., Gallinger, S., Giegling, I., Giuranna, J., Gonidakis, F., Gorwood, P., Mayora, M. G., Guillaume, S., Guo, Y., Hakonarson, H., Hatzikotoulas, K., Hauser, J., Hebebrand, J., Helder, S. G., Herms, S., Herpertz-Dahlmann, B., Herzog, W., Huckins, L. M., Hudson, J. I., Imgart, H., Inoko, H., Janout, V., Jiménez-Murcia, S., Julià, A., Kalsi, G., Kaminská, D., Kaprio, J., Karhunen, L., Karwautz, A., Kas, M. J. H., Kennedy, J. L., Keski-Rahkonen, A., Kiezebrink, K., Kim, Y.-R., Klareskog, L., Klump, K. L., Knudsen, G. P. S., La Via, M. C., Le Hellard, S., Levitan, R. D., Li, D., Lilenfeld, L., Lin, B. D., Lissowska, J., Luykx, J., Magistretti, P. J., Maj, M., Mannik, K., Marsal, S., Marshall, C. R., Mattingsdal, M., McDevitt, S., McGuffin, P., Metspalu, A., Meulenbelt, I., Micali, N., Mitchell, K., Monteleone, A. M., Monteleone, P., Munn-Chernoff, M. A., Nacmias, B., Navratilova, M., Ntalla, I., O'Toole, J. K., Ophoff, R. A., Padyukov, L., Palotie, A., Pantel, J., Papezova, H., Pinto, D., Rabionet, R., Raevuori, A., Ramoz, N., Reichborn-Kjennerud, T., Ricca, V., Ripatti, S., Ritschel, F., Roberts, M., Rotondo, A., Rujescu, D., Rybakowski, F., Santonastaso, P., Scherag, A., Scherer, S. W., Schmidt, U., Schork, N. J., Schosser, A., Seitz, J., Slachtova, L., Slagboom, P. E., Slof-Op 't Landt, M. C. T., Slopien, A., Sorbi, S., Świątkowska, B., Szatkiewicz, J. P., Tachmazidou, I., Tenconi, E., Tortorella, A., Tozzi, F., Treasure, J., Tsitsika, A., Tyszkiewicz-Nwafor, M., Tziouvas, K., van Elburg, A. A., van Furth, E. F., Wagner, G., Walton, E., Widen, E., Zeggini, E., Zerwas, S., Zipfel, S., Bergen, A. W., Boden, J. M., Brandt, H., Crawford, S., Halmi, K. A., Horwood, L. J., Johnson, C., Kaplan, A. S., Kaye, W. H., Mitchell, J. E., Olsen, C. M., Pearson, J. F., Pedersen, N. L., Strober, M., Werge, T., Whiteman, D. C., Woodside, D. B.,

Stuber, G. D., Gordon, S., Grove, J., Henders, A. K., Juréus, A., Kirk, K. M., Larsen, J. T., Parker, R., Petersen, L., Jordan, J., Kennedy, M., Montgomery, G. W., Wade, T. D., Birgegård, A., Lichtenstein, P., Norring, C., Landén, M., Martin, N. G., Mortensen, P. B., Sullivan, P. F., Breen, G. & Bulik, C. M. Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat. Genet.* **51,** 1207–1214 (2019).

167. Otowa, T., Hek, K., Lee, M., Byrne, E. M., Mirza, S. S., Nivard, M. G., Bigdeli, T., Aggen, S. H., Adkins, D., Wolen, A., Fanous, A., Keller, M. C., Castelao, E., Kutalik, Z., Van der Auwera, S., Homuth, G., Nauck, M., Teumer, A., Milaneschi, Y., Hottenga, J.-J., Direk, N., Hofman, A., Uitterlinden, A., Mulder, C. L., Henders, A. K., Medland, S. E., Gordon, S., Heath, A. C., Madden, P. A. F., Pergadia, M. L., van der Most, P. J., Nolte, I. M., van Oort, F. V. A., Hartman, C. A., Oldehinkel, A. J., Preisig, M., Grabe, H. J., Middeldorp, C. M., Penninx, B. W. J. H., Boomsma, D., Martin, N. G., Montgomery, G., Maher, B. S., van den Oord, E. J., Wray, N. R., Tiemeier, H. & Hettema, J. M. Meta-analysis of genome-wide association studies of anxiety disorders. *Mol. Psychiatry* **21,** 1391–1399 (2016).

168. Grove, J., Ripke, S., Als, T. D., Mattheisen, M., Walters, R. K., Won, H., Pallesen, J., Agerbo, E., Andreassen, O. A., Anney, R., Awashti, S., Belliveau, R., Bettella, F., Buxbaum, J. D., Bybjerg-Grauholm, J., Bækvad-Hansen, M., Cerrato, F., Chambert, K., Christensen, J. H., Churchhouse, C., Dellenvall, K., Demontis, D., De Rubeis, S., Devlin, B., Djurovic, S., Dumont, A. L., Goldstein, J. I., Hansen, C. S., Hauberg, M. E., Hollegaard, M. V., Hope, S., Howrigan, D. P., Huang, H., Hultman, C. M., Klei, L., Maller, J., Martin, J., Martin, A. R., Moran, J. L., Nyegaard, M., Nærland, T., Palmer, D. S., Palotie, A., Pedersen, C. B., Pedersen, M. G., dPoterba, T., Poulsen, J. B., Pourcain, B. S., Qvist, P., Rehnström, K., Reichenberg, A., Reichert, J., Robinson, E. B., Roeder, K., Roussos, P., Saemundsen, E., Sandin, S., Satterstrom, F. K., Davey Smith, G., Stefansson, H., Steinberg, S., Stevens, C. R., Sullivan, P. F., Turley, P., Walters, G. B., Xu, X., Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium, BUPGEN, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, 23andMe Research Team, Stefansson, K., Geschwind, D. H., Nordentoft, M., Hougaard, D. M., Werge, T., Mors, O., Mortensen, P. B., Neale, B. M., Daly, M. J. & Børglum, A. D. Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51,** 431–444 (2019).

169. Stahl, E. A., Breen, G., Forstner, A. J., McQuillin, A., Ripke, S., Trubetskoy, V., Mattheisen, M., Wang, Y., Coleman, J. R. I., Gaspar, H. A., de Leeuw, C. A., Steinberg, S., Pavlides, J. M. W., Trzaskowski, M., Byrne, E. M., Pers, T. H., Holmans, P. A., Richards, A. L., Abbott, L., Agerbo, E., Akil, H., Albani, D., Alliey-Rodriguez, N., Als, T. D., Anjorin, A., Antilla, V., Awasthi, S., Badner, J. A., Bækvad-Hansen, M., Barchas, J. D., Bass, N., Bauer, M., Belliveau, R., Bergen, S. E., Pedersen, C. B., Bøen, E., Boks, M. P., Boocock, J., Budde, M., Bunney, W., Burmeister, M., Bybjerg-Grauholm, J., Byerley, W., Casas, M., Cerrato, F., Cervantes, P., Chambert, K., Charney, A. W., Chen, D., Churchhouse, C., Clarke, T.-K., Coryell, W., Craig, D. W., Cruceanu, C., Curtis, D., Czerski, P. M., Dale, A. M., de Jong, S., Degenhardt, F., Del-Favero, J., DePaulo, J. R., Djurovic, S., Dobbyn, A. L., Dumont, A., Elvsåshagen, T., Escott-Price, V., Fan, C. C., Fischer, S. B., Flickinger, M., Foroud, T. M., Forty, L., Frank, J., Fraser, C., Freimer, N. B., Frisén, L., Gade, K., Gage, D., Garnham, J., Giambartolomei, C., Pedersen, M. G., Goldstein, J., Gordon, S. D., Gordon-Smith, K., Green,

E. K., Green, M. J., Greenwood, T. A., Grove, J., Guan, W., Guzman-Parra, J., Hamshere, M. L., Hautzinger, M., Heilbronner, U., Herms, S., Hipolito, M., Hoffmann, P., Holland, D., Huckins, L., Jamain, S., Johnson, J. S., Juréus, A., Kandaswamy, R., Karlsson, R., Kennedy, J. L., Kittel-Schneider, S., Knowles, J. A., Kogevinas, M., Koller, A. C., Kupka, R., Lavebratt, C., Lawrence, J., Lawson, W. B., Leber, M., Lee, P. H., Levy, S. E., Li, J. Z., Liu, C., Lucae, S., Maaser, A., MacIntyre, D. J., Mahon, P. B., Maier, W., Martinsson, L., McCarroll, S., McGuffin, P., McInnis, M. G., McKay, J. D., Medeiros, H., Medland, S. E., Meng, F., Milani, L., Montgomery, G. W., Morris, D. W., Mühleisen, T. W., Mullins, N., Nguyen, H., Nievergelt, C. M., Adolfsson, A. N., Nwulia, E. A., O'Donovan, C., Loohuis, L. M. O., Ori, A. P. S., Oruc, L., Ösby, U., Perlis, R. H., Perry, A., Pfennig, A., Potash, J. B., Purcell, S. M., Regeer, E. J., Reif, A., Reinbold, C. S., Rice, J. P., Rivas, F., Rivera, M., Roussos, P., Ruderfer, D. M., Ryu, E., Sánchez-Mora, C., Schatzberg, A. F., Scheftner, W. A., Schork, N. J., Shannon Weickert, C., Shehktman, T., Shilling, P. D., Sigurdsson, E., Slaney, C., Smeland, O. B., Sobell, J. L., Søholm Hansen, C., Spijker, A. T., St Clair, D., Steffens, M., Strauss, J. S., Streit, F., Strohmaier, J., Szelinger, S., Thompson, R. C., Thorgeirsson, T. E., Treutlein, J., Vedder, H., Wang, W., Watson, S. J., Weickert, T. W., Witt, S. H., Xi, S., Xu, W., Young, A. H., Zandi, P., Zhang, P., Zöllner, S., eQTLGen Consortium, BIOS Consortium, Adolfsson, R., Agartz, I., Alda, M., Backlund, L., Baune, B. T., Bellivier, F., Berrettini, W. H., Biernacka, J. M., Blackwood, D. H. R., Boehnke, M., Børglum, A. D., Corvin, A., Craddock, N., Daly, M. J., Dannlowski, U., Esko, T., Etain, B., Frye, M., Fullerton, J. M., Gershon, E. S., Gill, M., Goes, F., Grigoroiu-Serbanescu, M., Hauser, J., Hougaard, D. M., Hultman, C. M., Jones, I., Jones, L. A., Kahn, R. S., Kirov, G., Landén, M., Leboyer, M., Lewis, C. M., Li, Q. S., Lissowska, J., Martin, N. G., Mayoral, F., McElroy, S. L., McIntosh, A. M., McMahon, F. J., Melle, I., Metspalu, A., Mitchell, P. B., Morken, G., Mors, O., Mortensen, P. B., Müller-Myhsok, B., Myers, R. M., Neale, B. M., Nimgaonkar, V., Nordentoft, M., Nöthen, M. M., O'Donovan, M. C., Oedegaard, K. J., Owen, M. J., Paciga, S. A., Pato, C., Pato, M. T., Posthuma, D., Ramos-Quiroga, J. A., Ribasés, M., Rietschel, M., Rouleau, G. A., Schalling, M., Schofield, P. R., Schulze, T. G., Serretti, A., Smoller, J. W., Stefansson, H., Stefansson, K., Stordal, E., Sullivan, P. F., Turecki, G., Vaaler, A. E., Vieta, E., Vincent, J. B., Werge, T., Nurnberger, J. I., Wray, N. R., Di Florio, A., Edenberg, H. J., Cichon, S., Ophoff, R. A., Scott, L. J., Andreassen, O. A., Kelsoe, J., Sklar, P. & Bipolar Disorder Working Group of the Psychiatric Genomics Consortium. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nat. Genet.* **51,** 793–803 (2019).

170. Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H.-J., Shakhbazov, K., Abdellaoui, A., Agrawal, A., Albrecht, E., Alizadeh, B. Z., Amin, N., Barnard, J., Baumeister, S. E., Benke, K. S., Bielak, L. F., Boatman, J. A., Boyle, P. A., Davies, G., de Leeuw, C., Eklund, N., Evans, D. S., Ferhmann, R., Fischer, K., Gieger, C., Gjessing, H. K., Hägg, S., Harris, J. R., Hayward, C., Holzapfel, C., Ibrahim-Verbaas, C. A., Ingelsson, E., Jacobsson, B., Joshi, P. K., Jugessur, A., Kaakinen, M., Kanoni, S., Karjalainen, J., Kolcic, I., Kristiansson, K., Kutalik, Z., Lahti, J., Lee, S. H., Lin, P., Lind, P. A., Liu, Y., Lohman, K., Loitfelder, M., McMahon, G., Vidal, P. M., Meirelles, O., Milani, L., Myhre, R., Nuotio, M.-L., Oldmeadow, C. J., Petrovic, K. E., Peyrot, W. J., Polasek, O., Quaye, L., Reinmaa, E., Rice, J. P., Rizzi, T. S., Schmidt, H., Schmidt, R., Smith, A. V., Smith, J. A., Tanaka, T., Terracciano, A., van der Loos, M. J. H. M., Vitart, V., Völzke, H., Wellmann, J., Yu, L., Zhao, W., Allik, J., Attia, J. R., Bandinelli, S., Bastardot, F., Beauchamp, J., Bennett, D. A., Berger, K., Bierut, L. J., Boomsma, D. I., Bültmann, U., Campbell, H., Chabris, C. F.,

Cherkas, L., Chung, M. K., Cucca, F., de Andrade, M., De Jager, P. L., De Neve, J.-E., Deary, I. J., Dedoussis, G. V., Deloukas, P., Dimitriou, M., Eiríksdóttir, G., Elderson, M. F., Eriksson, J. G., Evans, D. M., Faul, J. D., Ferrucci, L., Garcia, M. E., Grönberg, H., Guðnason, V., Hall, P., Harris, J. M., Harris, T. B., Hastie, N. D., Heath, A. C., Hernandez, D. G., Hoffmann, W., Hofman, A., Holle, R., Holliday, E. G., Hottenga, J.-J., Iacono, W. G., Illig, T., Järvelin, M.-R., Kähönen, M., Kaprio, J., Kirkpatrick, R. M., Kowgier, M., Latvala, A., Launer, L. J., Lawlor, D. A., Lehtimäki, T., Li, J., Lichtenstein, P., Lichtner, P., Liewald, D. C., Madden, P. A., Magnusson, P. K. E., Mäkinen, T. E., Masala, M., McGue, M., Metspalu, A., Mielck, A., Miller, M. B., Montgomery, G. W., Mukherjee, S., Nyholt, D. R., Oostra, B. A., Palmer, L. J., Palotie, A., Penninx, B. W. J. H., Perola, M., Peyser, P. A., Preisig, M., Räikkönen, K., Raitakari, O. T., Realo, A., Ring, S. M., Ripatti, S., Rivadeneira, F., Rudan, I., Rustichini, A., Salomaa, V., Sarin, A.-P., Schlessinger, D., Scott, R. J., Snieder, H., St Pourcain, B., Starr, J. M., Sul, J. H., Surakka, I., Svento, R., Teumer, A., LifeLines Cohort Study, Tiemeier, H., van Rooij, F. J. A., Van Wagoner, D. R., Vartiainen, E., Viikari, J., Vollenweider, P., Vonk, J. M., Waeber, G., Weir, D. R., Wichmann, H.-E., Widen, E., Willemsen, G., Wilson, J. F., Wright, A. F., Conley, D., Davey-Smith, G., Franke, L., Groenen, P. J. F., Hofman, A., Johannesson, M., Kardia, S. L. R., Krueger, R. F., Laibson, D., Martin, N. G., Meyer, M. N., Posthuma, D., Thurik, A. R., Timpson, N. J., Uitterlinden, A. G., van Duijn, C. M., Visscher, P. M., Benjamin, D. J., Cesarini, D. & Koellinger, P. D. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340,** 1467–1471 (2013).

171. Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., DeStafano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thorton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., Vardarajan, B. N., Kamatani, Y., Lin, C. F., Gerrish, A., Schmidt, H., Kunkle, B., Dunstan, M. L., Ruiz, A., Bihoreau, M. T., Choi, S. H., Reitz, C., Pasquier, F., Cruchaga, C., Craig, D., Amin, N., Berr, C., Lopez, O. L., De Jager, P. L., Deramecourt, V., Johnston, J. A., Evans, D., Lovestone, S., Letenneur, L., Morón, F. J., Rubinsztein, D. C., Eiriksdottir, G., Sleegers, K., Goate, A. M., Fiévet, N., Huentelman, M. W., Gill, M., Brown, K., Kamboh, M. I., Keller, L., Barberger-Gateau, P., McGuiness, B., Larson, E. B., Green, R., Myers, A. J., Dufouil, C., Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J., St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D. W., Yu, L., Tsolaki, M., Bossù, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-Garcia, F., Fox, N. C., Hardy, J., Deniz Naranjo, M. C., Bosco, P., Clarke, R., Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's Disease Initiative (EADI), Genetic and Environmental Risk in Alzheimer's Disease, Alzheimer's Disease Genetic Consortium, Cohorts for Heart and Aging Research in Genomic Epidemiology, Moebus, S., Mecocci, P., Del Zompo, M., Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias, B., Gilbert, J. R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Carrasquillo, M. M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O., Younkin, S. G., Coto, E., Hamilton-Nelson, K. L., Gu, W., Razquin, C., Pastor, P., Mateo, I., Owen, M. J., Faber, K. M., Jonsson, P. V., Combarros, O., O'Donovan, M. C., Cantwell, L. B., Soininen, H., Blacker, D., Mead, S., Mosley, T. H., Jr, Bennett, D. A., Harris, T. B., Fratiglioni, L., Holmes, C., de Bruijn, R. F., Passmore, P., Montine, T. J., Bettens, K., Rotter, J. I., Brice, A., Morgan, K., Foroud, T. M., Kukull, W. A., Hannequin, D., Powell, J. F., Nalls, M. A., Ritchie, K., Lunetta, K. L., Kauwe, J. S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltuenen, M., Martin, E. R., Schmidt, R.,

Rujescu, D., Wang, L. S., Dartigues, J. F., Mayeux, R., Tzourio, C., Hofman, A., Nöthen, M. M., Graff, C., Psaty, B. M., Jones, L., Haines, J. L., Holmans, P. A., Lathrop, M., Pericak-Vance, M. A., Launer, L. J., Farrer, L. A., van Duijn, C. M., Van Broeckhoven, C., Moskvina, V., Seshadri, S., Williams, J., Schellenberg, G. D. & Amouyel, P. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45,** 1452–1458 (2013).

172. Gao, J., Davis, L. K., Hart, A. B., Sanchez-Roige, S., Han, L., Cacioppo, J. T. & Palmer, A. A. Genome-Wide Association Study of Loneliness Demonstrates a Role for Common Variation. *Neuropsychopharmacology* **42,** 811–821 (2017).

173. Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J., Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S.-A., Bækvad-Hansen, M., Beekman, A. F. T., Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschøn, H. N., Bybjerg-Grauholm, J., Cai, N., Castelao, E., Christensen, J. H., Clarke, T.-K., Coleman, J. I. R., Colodro-Conde, L., Couvy-Duchesne, B., Craddock, N., Crawford, G. E., Crowley, C. A., Dashti, H. S., Davies, G., Deary, I. J., Degenhardt, F., Derks, E. M., Direk, N., Dolan, C. V., Dunn, E. C., Eley, T. C., Eriksson, N., Escott-Price, V., Kiadeh, F. H. F., Finucane, H. K., Forstner, A. J., Frank, J., Gaspar, H. A., Gill, M., Giusti-Rodríguez, P., Goes, F. S., Gordon, S. D., Grove, J., Hall, L. S., Hannon, E., Hansen, C. S., Hansen, T. F., Herms, S., Hickie, I. B., Hoffmann, P., Homuth, G., Horn, C., Hottenga, J.-J., Hougaard, D. M., Hu, M., Hyde, C. L., Ising, M., Jansen, R., Jin, F., Jorgenson, E., Knowles, J. A., Kohane, I. S., Kraft, J., Kretzschmar, W. W., Krogh, J., Kutalik, Z., Lane, J. M., Li, Y., Li, Y., Lind, P. A., Liu, X., Lu, L., MacIntyre, D. J., MacKinnon, D. F., Maier, R. M., Maier, W., Marchini, J., Mbarek, H., McGrath, P., McGuffin, P., Medland, S. E., Mehta, D., Middeldorp, C. M., Mihailov, E., Milaneschi, Y., Milani, L., Mill, J., Mondimore, F. M., Montgomery, G. W., Mostafavi, S., Mullins, N., Nauck, M., Ng, B., Nivard, M. G., Nyholt, D. R., O'Reilly, P. F., Oskarsson, H., Owen, M. J., Painter, J. N., Pedersen, C. B., Pedersen, M. G., Peterson, R. E., Pettersson, E., Peyrot, W. J., Pistis, G., Posthuma, D., Purcell, S. M., Quiroz, J. A., Qvist, P., Rice, J. P., Riley, B. P., Rivera, M., Saeed Mirza, S., Saxena, R., Schoevers, R., Schulte, E. C., Shen, L., Shi, J., Shyn, S. I., Sigurdsson, E., Sinnamon, G. B. C., Smit, J. H., Smith, D. J., Stefansson, H., Steinberg, S., Stockmeier, C. A., Streit, F., Strohmaier, J., Tansey, K. E., Teismann, H., Teumer, A., Thompson, W., Thomson, P. A., Thorgeirsson, T. E., Tian, C., Traylor, M., Treutlein, J., Trubetskoy, V., Uitterlinden, A. G., Umbricht, D., Van der Auwera, S., van Hemert, A. M., Viktorin, A., Visscher, P. M., Wang, Y., Webb, B. T., Weinsheimer, S. M., Wellmann, J., Willemsen, G., Witt, S. H., Wu, Y., Xi, H. S., Yang, J., Zhang, F., eQTLGen, 23andMe, Arolt, V., Baune, B. T., Berger, K., Boomsma, D. I., Cichon, S., Dannlowski, U., de Geus, E. C. J., DePaulo, J. R., Domenici, E., Domschke, K., Esko, T., Grabe, H. J., Hamilton, S. P., Hayward, C., Heath, A. C., Hinds, D. A., Kendler, K. S., Kloiber, S., Lewis, G., Li, Q. S., Lucae, S., Madden, P. F. A., Magnusson, P. K., Martin, N. G., McIntosh, A. M., Metspalu, A., Mors, O., Mortensen, P. B., Müller-Myhsok, B., Nordentoft, M., Nöthen, M. M., O'Donovan, M. C., Paciga, S. A., Pedersen, N. L., Penninx, B. W. J. H., Perlis, R. H., Porteous, D. J., Potash, J. B., Preisig, M., Rietschel, M., Schaefer, C., Schulze, T. G., Smoller, J. W., Stefansson, K., Tiemeier, H., Uher, R., Völzke, H., Weissman, M. M., Werge, T., Winslow, A. R., Lewis, C. M., Levinson, D. F., Breen, G., Børglum, A. D., Sullivan, P. F. & Major Depressive Disorder Working Group of the Psychiatric Genomics

Consortium. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50,** 668–681 (2018).

174. Smith, D. J., Escott-Price, V., Davies, G., Bailey, M. E. S., Colodro-Conde, L., Ward, J., Vedernikov, A., Marioni, R., Cullen, B., Lyall, D., Hagenaars, S. P., Liewald, D. C. M., Luciano, M., Gale, C. R., Ritchie, S. J., Hayward, C., Nicholl, B., Bulik-Sullivan, B., Adams, M., Couvy-Duchesne, B., Graham, N., Mackay, D., Evans, J., Smith, B. H., Porteous, D. J., Medland, S. E., Martin, N. G., Holmans, P., McIntosh, A. M., Pell, J. P., Deary, I. J. & O'Donovan, M. C. Genome-wide analysis of over 106 000 individuals identifies 9 neuroticism-associated loci. *Mol. Psychiatry* **21,** 749–757 (2016).

175. Mattheisen, M., Samuels, J. F., Wang, Y., Greenberg, B. D., Fyer, A. J., McCracken, J. T., Geller, D. A., Murphy, D. L., Knowles, J. A., Grados, M. A., Riddle, M. A., Rasmussen, S. A., McLaughlin, N. C., Nurmi, E. L., Askland, K. D., Qin, H.-D., Cullen, B. A., Piacentini, J., Pauls, D. L., Bienvenu, O. J., Stewart, S. E., Liang, K.-Y., Goes, F. S., Maher, B., Pulver, A. E., Shugart, Y. Y., Valle, D., Lange, C. & Nestadt, G. Genome-wide association study in obsessive-compulsive disorder: results from the OCGAS. *Mol. Psychiatry* **20,** 337–344 (2015).

176. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

177. Ripke, S., O'Dushlaine, C., Chambert, K., Moran, J. L., Kähler, A. K., Akterin, S., Bergen, S. E., Collins, A. L., Crowley, J. J., Fromer, M., Kim, Y., Lee, S. H., Magnusson, P. K. E., Sanchez, N., Stahl, E. A., Williams, S., Wray, N. R., Xia, K., Bettella, F., Borglum, A. D., Bulik-Sullivan, B. K., Cormican, P., Craddock, N., de Leeuw, C., Durmishi, N., Gill, M., Golimbet, V., Hamshere, M. L., Holmans, P., Hougaard, D. M., Kendler, K. S., Lin, K., Morris, D. W., Mors, O., Mortensen, P. B., Neale, B. M., O'Neill, F. A., Owen, M. J., Milovancevic, M. P., Posthuma, D., Powell, J., Richards, A. L., Riley, B. P., Ruderfer, D., Rujescu, D., Sigurdsson, E., Silagadze, T., Smit, A. B., Stefansson, H., Steinberg, S., Suvisaari, J., Tosato, S., Verhage, M., Walters, J. T., Multicenter Genetic Studies of Schizophrenia Consortium, Levinson, D. F., Gejman, P. V., Kendler, K. S., Laurent, C., Mowry, B. J., O'Donovan, M. C., Owen, M. J., Pulver, A. E., Riley, B. P., Schwab, S. G., Wildenauer, D. B., Dudbridge, F., Holmans, P., Shi, J., Albus, M., Alexander, M., Campion, D., Cohen, D., Dikeos, D., Duan, J., Eichhammer, P., Godard, S., Hansen, M., Lerer, F. B., Liang, K.-Y., Maier, W., Mallet, J., Nertney, D. A., Nestadt, G., Norton, N., O'Neill, F. A., Papadimitriou, G. N., Ribble, R., Sanders, A. R., Silverman, J. M., Walsh, D., Williams, N. M., Wormley, B., Psychosis Endophenotypes International Consortium, Arranz, M. J., Bakker, S., Bender, S., Bramon, E., Collier, D., Crespo-Facorro, B., Hall, J., Iyegbe, C., Jablensky, A., Kahn, R. S., Kalaydjieva, L., Lawrie, S., Lewis, C. M., Lin, K., Linszen, D. H., Mata, I., McIntosh, A., Murray, R. M., Ophoff, R. A., Powell, J., Rujescu, D., Van Os, J., Walshe, M., Weisbrod, M., Wiersma, D., Wellcome Trust Case Control Consortium 2, Donnelly, P., Barroso, I., Blackwell, J. M., Bramon, E., Brown, M. A., Casas, J. P., Corvin, A. P., Deloukas, P., Duncanson, A., Jankowski, J., Markus, H. S., Mathew, C. G., Palmer, C. N. A., Plomin, R., Rautanen, A., Sawcer, S. J., Trembath, R. C., Viswanathan, A. C., Wood, N. W., Spencer, C. C. A., Band, G., Bellenguez, C., Freeman, C., Hellenthal, G., Giannoulatou, E., Pirinen, M., Pearson, R. D., Strange, A., Su, Z., Vukcevic, D., Donnelly,

P., Langford, C., Hunt, S. E., Edkins, S., Gwilliam, R., Blackburn, H., Bumpstead, S. J., Dronov, S., Gillman, M., Gray, E., Hammond, N., Jayakumar, A., McCann, O. T., Liddle, J., Potter, S. C., Ravindrarajah, R., Ricketts, M., Tashakkori-Ghanbaria, A., Waller, M. J., Weston, P., Widaa, S., Whittaker, P., Barroso, I., Deloukas, P., Mathew, C. G., Blackwell, J. M., Brown, M. A., Corvin, A. P., McCarthy, M. I., Spencer, C. C. A., Bramon, E., Corvin, A. P., O'Donovan, M. C., Stefansson, K., Scolnick, E., Purcell, S., McCarroll, S. A., Sklar, P., Hultman, C. M. & Sullivan, P. F. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45,** 1150–1159 (2013).

178. 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).

179. Hoyer, S. & Hamman, J. J. xarray: N-D labeled Arrays and Datasets in Python. *J. Open Res. Softw.* **5,** (2017).

180. Miles, A., jakirkham, Durant, M., Bussonnier, M., Bourbeau, J., Onalan, T., Hamman, J., Patel, Z., Rocklin, M., shikharsg, Abernathey, R., Moore, J., Schut, V., Dussin, R., de Andrade, E. S., Noyes, C., Jelenak, A., Banihirwe, A., Barnes, C., Sakkis, G., Funke, J., Kelleher, J., Jevnik, J., Swaney, J., Rahul, P. S., Saalfeld, S., john, Tran, T., Bot, P. io & sbalmer. *zarr-developers/zarr-python: v2.5.0*. (2020). doi:10.5281/zenodo.4069231

181. 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).

182. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P. & Satija, R. Integrated analysis of multimodal single-cell data. *Cell* **184,** 3573-3587.e29 (2021).

183. Yu, M., Abnousi, A., Zhang, Y., Li, G., Lee, L., Chen, Z., Fang, R., Lagler, T. M., Yang, Y., Wen, J., Sun, Q., Li, Y., Ren, B. & Hu, M. SnapHiC: a computational pipeline to identify chromatin loops from single-cell Hi-C data. *Nat. Methods* **18,** 1056–1059 (2021).

184. Li, H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27,** 718–719 (2011).

185. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Alexander Wolf, F. anndata: Annotated data. *bioRxiv* 2021.12.16.473007 (2021). doi:10.1101/2021.12.16.473007

186. Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L. & Paul, C. L. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 1827–1831 (1992).

187. Clemens, A. W. & Gabel, H. W. Emerging Insights into the Distinctive Neuronal Methylome. *Trends Genet.* (2020). doi:10.1016/j.tig.2020.07.009

188. 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).

189. Park, Y. & Wu, H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics* **32,** 1446–1453 (2016).

190. Korthauer, K., Chakraborty, S., Benjamini, Y. & Irizarry, R. A. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics* **20,** 367–383 (2019).

191. Chen, L., Liu, Z., Zhou, B., Wei, C., Zhou, Y., Rosenfeld, M. G., Fu, X.-D., Chisholm, A. D. & Jin, Y. CELF RNA binding proteins promote axon regeneration in C. elegans and mammals through alternative splicing of Syntaxins. *Elife* **5,** (2016).

192. Clarke, L. E., Liddelow, S. A., Chakraborty, C., Münch, A. E., Heiman, M. & Barres, B. A. Normal aging induces A1-like astrocyte reactivity. *Proc. Natl. Acad. Sci. U. S. A.* **115,** E1896–E1905 (2018).

193. Lennon, M. J., Jones, S. P., Lovelace, M. D., Guillemin, G. J. & Brew, B. J. Bcl11b-A Critical Neurodevelopmental Transcription Factor-Roles in Health and Disease. *Front. Cell. Neurosci.* **11,** 89 (2017).

194. Rocklin, M. Dask: Parallel computation with blocked algorithms and task scheduling. in *Proceedings of the 14th python in science conference* **130,** 136 (Citeseer, 2015).

195. Fabyanic, E. B., Hu, P., Qiu, Q., Wang, T., Berrios, K. N., Flournoy, J., Connolly, D. R., Zhou, Z., Kohli, R. M. & Wu, H. Quantitative single cell 5hmC sequencing reveals non-canonical gene regulation by non-CG hydroxymethylation. *Cold Spring Harbor Laboratory* 2021.03.23.434325 (2021). doi:10.1101/2021.03.23.434325

196. Gao, Y., Li, L., Yuan, P., Zhai, F., Ren, Y., Yan, L., Li, R., Lian, Y., Zhu, X., Wu, X., Kee, K., Wen, L., Qiao, J. & Tang, F. 5-Formylcytosine landscapes of human preimplantation embryos at single-cell resolution. *PLoS Biol.* **18,** e3000799 (2020).

197. Zhu, C., Zhang, Y., Li, Y. E., Lucero, J., Behrens, M. M. & Ren, B. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods* **18,** 283–292 (2021).

198. Zhang, M., Eichhorn, S. W., Zingg, B., Yao, Z., Cotter, K., Zeng, H., Dong, H. & Zhuang, X. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598,** 137–143 (2021).

199. Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F. & Macosko, E. Z. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363,** 1463–1467 (2019).

200. Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C. C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., Norris, E., Pan, A., Li, J., Xiao, Y., Halene, S. & Fan, R. High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell* **183,** 1665-1681.e18 (2020).

201. Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **0,** (2020).

202. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. Array programming with NumPy. *Nature* **585,** 357–362 (2020).

203. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17,** 261–272 (2020).

204. Reback, J., jbrockmendel, McKinney, W., Van den Bossche, J., Augspurger, T., Roeschke, M., Hawkins, S., Cloud, P., gfyoung, Sinhrks, Hoefler, P., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., Darbyshire, J. H. M., Garcia, M., Shadrach, R., Schendel, J., Hayden, A., Saxton, D., Gorelli, M. E., Li, F., Zeitlin, M., Jancauskas, V., McMaster, A., Wörtwein, T. & Battiston, P. *pandas-dev/pandas: Pandas 1.4.2*. (2022). doi:10.5281/zenodo.6408044

205. Lam, S. K., Pitrou, A. & Seibert, S. Numba: a LLVM-based Python JIT compiler. in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC* 1–6 (Association for Computing Machinery, 2015).

206. Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T. & Davies, R. M. HTSlib: C library for reading/writing high-throughput sequencing data. *Gigascience* **10,** (2021).

207. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).

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

209. Ryan, D., Grüning, B. & Ramirez, F. *pyBigWig 0.2.4*. (2016). doi:10.5281/zenodo.45238

210. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Others. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12,** 2825–2830 (2011).

211. McInnes, L. *pynndescent: A Python nearest neighbor descent for approximate nearest neighbors*. (Github). at <https://github.com/lmcinnes/pynndescent>

212. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9,** 90–95 (2007).

213. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6,** 3021 (2021).

214. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36,** 311–316 (2020).