UNIVERSITY OF CALIFORNIA SAN DIEGO

Integrated computational analysis of brain cell transcriptomes and epigenomes

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

in

Biophysics

by

Fangming Xie

Committee in charge:

Professor Eran A. Mukamel, Chair
Professor David Kleinfeld, Co-Chair
Professor Joseph R. Ecker
Professor Elena Koslover
Professor Bing Ren

2021

The dissertation of Fangming Xie is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To Qihui, Janna, Mom and Dad

EPIGRAPH

Cells are matter that dances.

Uri Alon

Physics was a point of view that the world around us is, with effort, ingenuity, and adequate resources, understandable in a predictive and reasonably quantitative fashion.

John J. Hopfield

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTAL FILES

Supplemental tables

ACKNOWLEDGEMENTS

Five years of graduate school transformed me in many ways: from physics to biology, from reading and learning to teaching and tinkering, from minding my own business to collaborating in big teams, and from one side of the Pacific Ocean to the other. These gaps were at times hopeless to traverse. It is the support from my mentors, colleagues, friends and family that upheld and nurtured my zeal for science. In this sense, I am extremely lucky and deeply grateful about my luck.

My advisor, Eran Mukamel, is the best PhD advisor I could ever ask for. He led me into the field, teaching me everything hands-on, in his words “Bob Rossing”. Meanwhile, he trusted me to work on the most challenging projects, and granted me extraordinary flexibility.

I had a great time sharing an office, and more recently a Zoom room, with Mukamel lab labmates. Junhao Li and Chris Keown were extremely helpful and patient during my disorienting start. Chris set an example for how to get a PhD. Junhao has been my labmate for the longest time, and my go-to person for all issues. I was inspired by Ethan Armand’s creativity, openness, and his zeal to organize and help everyone. Wayne Doyle always knew the biology and the best Illustrator diagram to make, when I was ignorant to both. Jo-Fan Chien and Alon Gelber joined the lab after the pandemic; they brought much needed energy to the physically distanced lab. I also learned a lot from mentoring Mingxuan Zhang and Adityna Chandrasekar, who were willing to bear with my inexperience.

In a large part, my work is defined by the experimental biologists I collaborated with. Chongyuan Luo always offered me the most novel data and the best analysis opportunities. He also offered me a desk later in his own lab, where I finished this thesis. Professor Joseph Ecker and Professor Bing Ren are gracious in sharing data, offering constructive and critical comments, and serving on my dissertation committee. In addition, Professor Hongkui Zeng, Marga Behren, Ed Callway, Paula Desplats and Xiangmin Xu had crucial impacts on my work at

different stages and from different perspectives.

I was fortunate to be embedded in a vibrant community of peer-bioinformaticians: Jingtian Zhou, Hanqing Liu, Yang Li, Rongxin Fang, Zizhen Yao, and Stephan Fischer. They make my scientific journey never lonely.

The introduction section, in part, used the material as it appears in *Neuron* 2021. E. J. Armand, J. Li, **F. Xie**, C. Luo, E. A. Mukamel, Single-Cell Sequencing of Brain Cell Transcriptomes and Epigenomes. *Neuron*. **109**, 11–26 (2021). The dissertation author was a co-first author of this paper.

Chapter 1, in full, is a reprint of the material as it appears in *Nature* 2021. Z. Yao, H. Liu, **F. Xie**, S. Fischer, R. S. Adkins, A. I. Aldridge, S. A. Ament, A. Bartlett, M. M. Behrens, K. Van den Berge, D. Bertagnolli, H. R. de Bézieux, T. Biancalani, A. S. Booeshaghi, H. C. Bravo, T. Casper, C. Colantuoni, J. Crabtree, H. Creasy, K. Crichton, M. Crow, N. Dee, E. L. Dougherty, W. I. Doyle, S. Dudoit, R. Fang, V. Felix, O. Fong, M. Giglio, J. Goldy, M. Hawrylycz, B. R. Herb, R. Hertzano, X. Hou, Q. Hu, J. Kancherla, M. Kroll, K. Lathia, Y. E. Li, J. D. Lucero, C. Luo, A. Mahurkar, D. McMillen, N. M. Nadaf, J. R. Nery, T. N. Nguyen, S.-Y. Niu, V. Ntranos, J. Orvis, J. K. Osteen, T. Pham, A. Pinto-Duarte, O. Poirion, S. Preissl, E. Purdom, C. Rimorin, D. Risso, A. C. Rivkin, K. Smith, K. Street, J. Sulc, V. Svensson, M. Tieu, A. Torkelson, H. Tung, E. D. Vaishnav, C. R. Vanderburg, C. van Velthoven, X. Wang, O. R. White, Z. J. Huang, P. V. Kharchenko, L. Pachter, J. Ngai, A. Regev, B. Tasic, J. D. Welch, J. Gillis, E. Z. Macosko, B. Ren, J. R. Ecker, H. Zeng, E. A. Mukamel, A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*. **598**, 103–110 (2021). The dissertation author was a co-first author of this paper.

Chapter 2, in part, has been submitted for publication of the material. The preprint of this manuscript is posted on *bioRxiv*. C. Luo, H. Liu, **F. Xie**, E. J. Armand, K. Siletti, T. Bakken, R. Fang, W. I. Doyle, R. D. Hodge, L. Hu, B.-A. Wang, Z. Zhang, S. Preissl, D.-S. Lee, J. Zhou, S.-Y. Niu, R. Castanon, A. Bartlett, A. Rivkin, X. Wang, J. Lucero, J. R. Nery, D. A. Davis, D. C.

Mash, J. R. Dixon, S. Linnarsson, E. Lein, M. Margarita Behrens, B. Ren, E. A. Mukamel, J. R. Ecker, Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv* (2019), p. 2019.12.11.873398. The dissertation author was a co-first author of this paper.

Chapter 3, in full, has been submitted for publication of the material. The preprint of this manuscript is posted on *bioRxiv*. **F. Xie**, E. J. Armand, Z. Yao, H. Liu, A. Bartlett, M. Margarita Behrens, Y. E. Li, J. D. Lucero, C. Luo, J. R. Nery, A. Pinto-Duarte, O. Poirion, S. Preissl, A. C. Rivkin, B. Tasic, H. Zeng, B. Ren, J. R. Ecker, E. A. Mukamel, Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes. *bioRxiv* (2021), p. 2021.10.25.465795. The dissertation author was a co-first author of this paper.

David Kleinfeld and Elena Koslover serve on my committee as physics professors. Their lectures, neuro-physics and cell-physics, helped me migrate from physics to biology. More recently, they reminded me to look back on the rigor and power of quantitative modeling. The physics department, despite me straying away, has been my nest. Sharmila Poddar is the best graduate coordinator--always helpful, energetic, and prompt. My classmates and roommates, in particular Xiang Ji, Dawei Li, Weiting Kuo, Mingdong Wang, and Wangmuge Qin, were my first and lasting friends in graduate school, with whom I shared an apartment, ideas, anguish and excitement.

My parents Yue Fang and Ling Xie, who granted me nature and nurture, have been supportive of me wandering far away from home to study neurons. In fact, they are proud and supportive of whatever I set out to do.

Finally, the best luck I had in graduate school and in life, is my wife Qihui Lyu. During graduate school, we met, married, and gave birth to our daughter Janna. Qihui supports my endeavor to scrutinize the notes of life inscribed in DNA, but most of all, she reminds me to enjoy the music of life produced around us in and of itself.

VITA

2016	Bachelor of Science, University of Science and Technology of China
2016-2019	Teaching Assistant, University of California San Diego
2017-2021	Research Assistant, University of California San Diego
2021	Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Integrated computational analysis of brain cell transcriptomes and epigenomes

by

Fangming Xie

Doctor of Philosophy in Biophysics

University of California San Diego, 2021

Professor Eran M. Mukamel, Chair
Professor David Kleinfeld, Co-Chair

The mammalian brain consists of a vast network of neurons and non-neuronal cells with diverse morphology, anatomy, physiology and behavioral roles[1,2]. These cellular phenotypes are enacted and maintained by a complex molecular program, including the abundance of gene

transcripts, i.e. the transcriptome, and epigenetic modifications of DNA, i.e., the epigenome[3]. Single cell sequencing assays, capable of measuring the entire transcriptome or epigenome for hundreds of thousands of single cells, have enabled the systematic characterization of brain cell types at unprecedented scale and with fine granularity[4–9]. However, it is challenging to integrate diverse datasets, which differ in sample and library preparation, sequencing platforms, and assay modalities, for a consistent biological understanding of cell type organization.

This thesis presents novel computational algorithms to integrate brain cell transcriptomes and epigenomes. We developed SingleCellFusion[10], which integrates disparate datasets into a common feature space based on a constrained k-nearest-neighbor graph algorithm. Using SingleCellFusion, we integrated 8 datasets with >400,000 cells from the mouse primary motor cortex (MOp). This analysis identified 56 neuronal cell types with consistent cell type specific patterns of gene expression, chromatin accessibility and DNA methylation.

To validate the accuracy of SingleCellFusion, we helped to develop a novel multimodal sequencing assay, snmCAT-seq[11], that simultaneously measures methylCytosine (mC), chromatin Accessibility (A), and Transcriptome (T) from the same cells. Applying snmCAT-seq to 3,898 human frontal cortex cells, we identified fine grained neuronal cell types. SingleCellFusion integrated single-cell transcriptomes and DNA methylomes from the same cell types with 62.6~87.3% accuracy, recapitulating snmCAT-seq results at the cell type level.

Cell type specific gene expression is in part regulated by epigenetic modifications of DNA at *cis*-regulatory elements (CREs), which are typically located thousands of base pairs away from the gene they regulate. We took advantage of co-variations in gene expression and epigenetic activity at candidate CREs across cell types to identify brain cell-type-specific gene-CRE associations[12]. We developed a method that identified more than 10,000 robust gene-CRE associations from mouse MOp, using an empirical data shuffling procedure to control for false positives due to gene co-expression.

Our results highlight the power of integrating transcriptomes and epigenomes to uncover the complex molecular regulation of brain cell types, and will directly enable design of reagents to target specific cell types for functional analysis. It also demonstrates that robust and efficient computational analysis methods are imperative to distill biological understandings from disparate large-scale single cell sequencing data.

INTRODUCTION

This thesis presents a series of studies that attempted to map brain cell type diversity through the integrated analysis of brain cell transcriptomes and epigenomes. Some of these studies were carried out in collaboration with the BRAIN Initiative Cell Census Network (BICCN)[13], a diverse consortium of molecular biologists, neurobiologists, bioinformaticians and statisticians united by the goal of comprehensively mapping the brain's cellular components. The three chapters present three consecutive initial steps toward the construction of a transcriptomic and epigenomic cell type atlas of the mammalian brain. Cell type diversity has long been recognized as a defining feature of the mammalian brain, and a central organizing principle of brain function. Over a century ago, Santiago Ramón y Cajal found brain cells have diverse morphologies, and recorded them with his artful yet rigorous drawings[1]. Generations of neuroscientists since then have further refined our understanding of how brain cells differ in their anatomy, electrophysiological properties, molecular signatures, and behavioral roles[14]. These threads of discoveries form a delicate tapestry of knowledge about the brain's parts-list, where molecules, structures, functions and behaviors intertwine. Yet, they exposed a profound challenge that even the basic units of the brain, neurons and glia, were seemingly too diverse to be comprehensively charted.

Single-cell sequencing technologies, developed and progressively improved over the last decade, are capable of measuring comprehensive molecular profiles (-omes) for hundreds of thousands of single cells in a single experiment[3]. These methods represent a unique opportunity to comprehensively survey brain cell types with unprecedented scale and resolution. Hundreds of new fine-grained brain cell types were identified by single-cell RNA-seq (scRNA-seq) and unsupervised clustering of brain cells' transcriptomic signatures[4–6].

However, the explosion of newly identified transcriptomic cell types raises the challenge of validating these discoveries. scRNA-seq data is noisy, as it captures and selectively amplifies

a small, random sample (5-20%) of a cell's messenger RNAs, which themselves are intrinsically dynamic and sensitive to sample preparation. This gives rise to batch effects--scRNA-seq datasets generated by different labs, technologies, and sample preparation protocols can be quantitatively very different. It is therefore important to reconcile batch effects and validate transcriptome-based results by complementary measurements. Moreover, the transcriptome represents only one component of a cell's molecular identity. A comprehensive cell type taxonomy should include multi-modal cell features, including epigenomes[9,15] and multi-modal contexts[16,17] (Figure 1).

To get a coherent biological understanding of cell types from disparate single-cell sequencing datasets, several computational data integration methods were developed[18,19]. Parametric methods, such as approximate canonical correlation analysis (CCA) implemented by Seurat[20] or non-negative matrix factorization (NMF)[21] project cells from multiple datasets into a common, low-dimensional space where they can be directly compared, clustered, and analyzed. Non-parametric methods such as mutual nearest neighbors (MNN) can also link cells across datasets, without learning an embedding in a common space[22,23]. These techniques link cells in one dataset with closely matching cells in another dataset (e.g., by selecting the cells with the most correlated gene-oriented signatures). However, these methods work primarily in the context of batch-effect correction of scRNA-seq datasets, rather than the integration of transcriptomes and epigenomes. It was unclear whether these tools are capable of integrating gene expression, DNA methylation, and chromatin accessibility datasets altogether. These studies also lack systematic validation of cell types at the scale and complexity of brain circuits.

This thesis presents novel computational methods to address the challenge of multi-modal data integration and cell type cross validation at the scale and complexity of the mammalian brain. It also presents how a multi-modal cell type atlas uncovers new details on the mechanisms of gene regulation by epigenetic marks.

Chapter 1 presents a multi-modal cell type atlas of the mouse primary motor cortex[10]. We developed a robust and efficient computational method, SingleCellFusion, that integrated more than 400,000 cells from 8 disparate single-cell sequencing datasets with diverse data modalities, including gene expression (scRNA-seq/snRNA-seq), DNA methylation (snmC-seq), and chromatin accessibility (snATAC-seq). We identified 56 neuronal cell types with distinct marker genes and regulatory elements, including a type of non-canonical layer 4 excitatory neurons. We also developed a cluster cross-validation method to objectively identify the maximum cell type resolution from data, and it suggests mouse MOp contains ~100 neuronal cell types reproducible across datasets.

Chapter 2 presents the analysis of snmCAT-seq, a novel multimodal sequencing assay that can simultaneously measure transcriptomes and epigenomes (DNA methylation and chromatin accessibility) in the same cells[11]. We applied this new assay to human prefrontal cortex, and used it to validate computational integration methods. We found SingleCellFusion, as well as a suite of other algorithms, can correctly cluster together transcriptomes and epigenomes from the same cell types, though they cannot match the exact cells.

Chapter 3 presents an analysis of the gene regulatory network of brain cell types. Cell type-specific gene expression are in part regulated by *cis*-regulatory elements (CREs), which are non-coding regions of the genome located thousands to millions base pairs away from the genes they controls[12]. It is therefore unclear which gene(s) a CRE targets. We took advantage of cross-cell-type variations in gene expression and CRE epigenetic activities to identify robust enhancer-gene associations. In particular, we developed a non-parametric data shuffling scheme to control for potential false positive gene-CRE associations due to widespread gene co-expression. Our methods identified more than 10,000 robust gene-CRE associations in mouse MOp.

Nonparametric methods underlie every aspect of the computational analyses. For example, computational integration of disparate datasets relies on building a nearest neighbor

graph between cells from different datasets. Spearman correlation coefficient is commonly used to measure similarity between cells or association strength between gene-CRE pairs. They are robust techniques that work well for noisy single-cell sequencing data, without relying on specific assumptions about the statistical distribution of signal and noise in these datasets. They also reflect a lack of understanding--we are far from knowing enough about brain cells, transcriptomes, or epigenomes to model data from first principles.

To make a case for a mouse connectome project, two-dozen prominent neuro-biologists wrote recently that “large scientific projects in genomics and astronomy are influential not because they answer any single question but because they enable investigation of continuously arising new questions from the same data-rich sources”[24]. In the same spirit, I hope the brain cell type atlas presented in this thesis will enable many targeted investigations in the future. If we were to compare with astronomy though, neuroscience is probably still in its pre-Kepler era: We have a rapidly expanding empirical dataset, yet we have just begun to discover underlying principles. Computational methods developed in this thesis are the first steps towards a systematic description of the brain’s parts-list. Hopefully, they will provide a foundation for mechanistic understandings of brain cell types, including how they are developed from a single fertilized egg, and how they orchestrate to command complex behaviors.

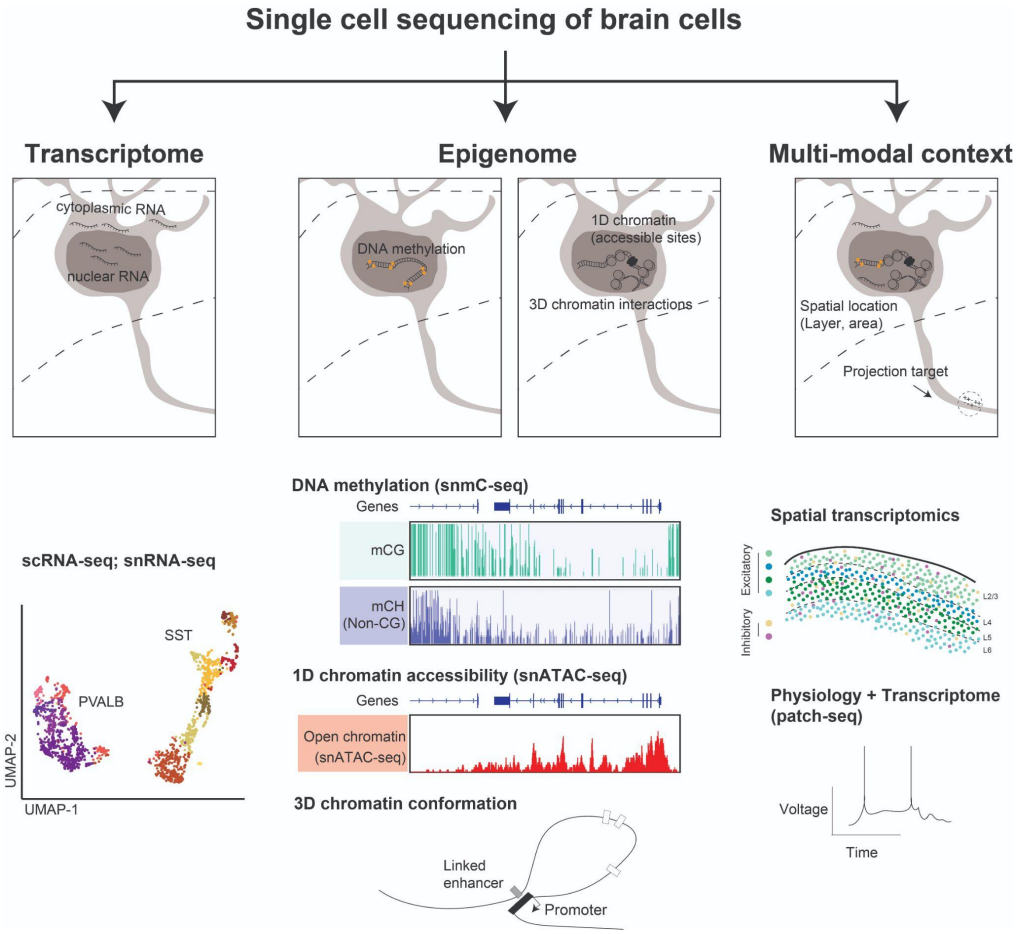


Figure 1: An overview of a brain cell's multi-modal features and measurements. A brain cell can be characterized by transcriptome, epigenome, as well as anatomical and physiological contexts.

Chapter 1. Transcriptomic and epigenomic cell atlas of mouse primary motor cortex

1.1. Abstract

Single-cell transcriptomics provides quantitative molecular signatures for large, unbiased samples of the brain's diverse cell types[4–6]. With the proliferation of multi-omics datasets, a major challenge is to validate and integrate results into a biological understanding of cell type organization. We generated transcriptomes and epigenomes from more than 500,000 individual cells in the mouse primary motor cortex (MOp), a structure with an evolutionarily conserved role in locomotion. We developed computational and statistical methods to integrate multimodal data and quantitatively validate cell type reproducibility. The resulting reference atlas, containing over 56 neuronal cell types that are highly replicable across analysis methods, sequencing technologies, and modalities, is a comprehensive molecular and genomic account of the diverse neuronal and non-neuronal cell types in MOp. It includes a population of excitatory neurons resembling layer 4 pyramidal cells in other cortical regions[25]. We further discovered thousands of concordant marker genes and gene regulatory elements for these cell types. Our results highlight the complex molecular regulation of brain cell types and will directly enable design of reagents to target specific MOp cell types for functional analysis.

1.2. Introduction

The cellular components of brain circuits are extraordinarily diverse[1,2]. Single-cell molecular assays, especially transcriptomic measurements by RNA-Seq, have accelerated cell type discovery across brain regions and in diverse species[3]. Recent advances include single-cell transcriptome datasets with $>10^5$ individual cells, identifying hundreds of neuronal and non-neuronal cell types across the mouse nervous system[4–6]. As the number of profiled cells grows into the millions, a key question is whether these data will converge toward a comprehensive, coherent taxonomy. Although a comprehensive cell atlas should incorporate

anatomical and physiological information, the high throughput of single-cell sequencing assays presents an opportunity for establishing a broad-based transcriptomic and epigenomic cell atlas. Molecular and genomic cell signatures will drive progress across modalities and help to obtain functional information.

Within the BRAIN Initiative Cell Census Network (BICCN), we aim to create an atlas of cell types across the brain of several mammalian species by integrating multiple single-cell omics approaches. We selected primary motor cortex (MOp; Figure S1a-d) as the starting point for our joint efforts due to its relatively conserved structure and function across mammalian species. MOp lacks species-specific cellular structures, such as the whisker barrels in the rodent primary somatosensory cortex and the elaborate layer 4 with multiple sublayers in the primate primary visual cortex. Traditionally, MOp is considered to lack a cytoarchitechtonically-defined granular layer (layer 4), although MOp neurons with Layer 4-like connectivity have been identified[25]. Our mouse MOp atlas is a case study of the expansive potential and the technical limitations of single-cell molecular methods for comprehensive brain-wide analysis of cell types.

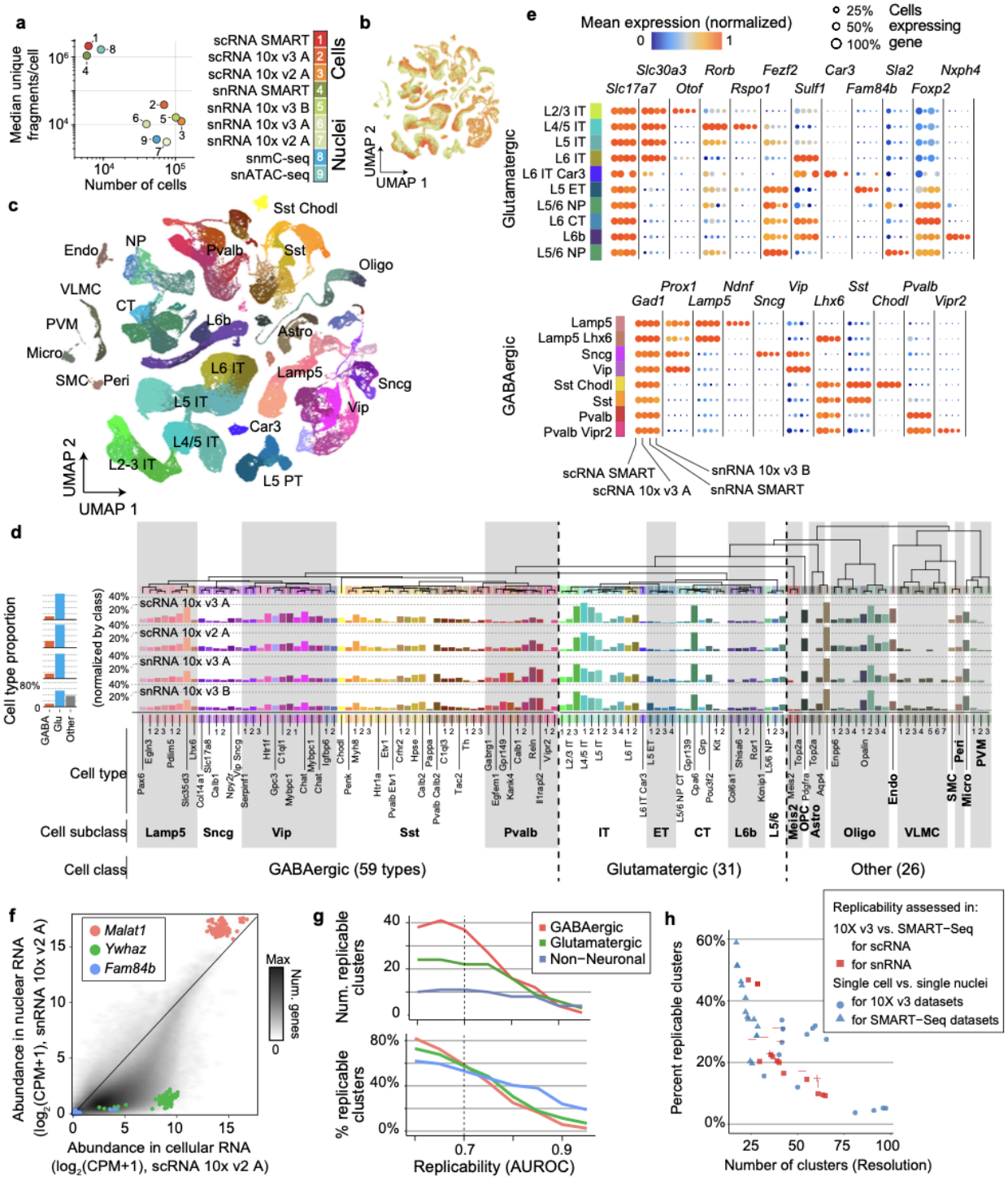
Single-cell transcriptomics identifies cell type marker genes and gene modules that shape functions such as the mode of synaptic communication[26]. Epigenomic measurements of DNA methylation and open chromatin provide signatures of gene regulation, including non-coding regulatory regions such as enhancers. Neurons acquire unique patterns of CG and non-CG DNA methylation during postnatal development[27,28] and have cell type-specific open chromatin[29]. Together, transcription and epigenetic modifications establish attractors in cell state space corresponding to cell types[30,31]. Here, we integrate large-scale single-cell transcriptome and epigenome datasets to achieve a reference taxonomy for the adult mouse MOp.

1.3. Results

1.3.1. Multimodal molecular census of mouse MOp

We produced 9 datasets, including 7 single-cell (sc) or single-nucleus (sn) transcriptome (scRNA-seq and snRNA-seq using 10x v2, v3 and SMART-Seq v4; n=526,373 high-quality cells), one single-nucleus DNA methylation (snmC-Seq2, n=9,872) and one single-nucleus open chromatin dataset (snATAC-Seq, n=81,196) (Figure S1e-f; Table S1). These span a range of technologies, assaying different numbers of cells, with different depth of sequence coverage per cell, and assessing different biological features (Figure 2a). The datasets reflect the tradeoff between the number of sequenced molecules per cell, which depends on cell size and the efficiency of RNA or DNA capture, vs. the total number of cells that can be assayed for a fixed total cost. Our datasets include single-nucleus transcriptomes from over 175,000 cells (using the 10x Chromium 3' version 3 platform), which captures 3,100-12,700 unique molecules per cell (median UMI/cell). By contrast, full-length transcript sequencing using SMART-Seq v4 captured a greater number of unique molecular fragments per cell (1-2.1 million), but covered fewer cells (~6,300 per dataset). Single-nucleus DNA methylation data provided deep coverage of the epigenome per cell (median 1.66 million unique sequenced DNA fragments, covering 6.2% of the genome) for a modest number of cells[27,32] (~9,800). Finally, snATAC-Seq data scaled to over 81,000 cells but sampled fewer DNA fragments for individual cells (median 3,778 unique fragments/cell, Table S1)[29].

Figure 2: Multi-platform transcriptomic taxonomy of MOp cell types. a, Key attributes of 9 single-cell transcriptome and epigenome datasets from mouse MOp. **b,c,** Two-dimensional projection (UMAP[33]) of cells and nuclei based on integrated analysis of seven transcriptome (sc/snRNA-seq) datasets. Cells and nuclei are colored by dataset (b) or by cell type (c). Non-neuronal cell types are depleted due to the sampling strategy which enriched neurons in all datasets except snRNA 10x v3 B. **d,** Dendrogram showing hierarchical relationship among the consensus transcriptomic cell types and proportion of cells of each type per dataset, normalized within major classes. **e,** Expression of selected marker genes for excitatory and inhibitory cell classes, across four platforms. **f,** Differential enrichment of transcripts in single cells vs. single nuclei. The long non-coding RNA *Malat1* is enriched in nuclei. **g,** Number of replicable clusters across at least two of the seven sc/snRNA-seq datasets as a function of minimal MetaNeighbor score (AUROC: area under receiver operating characteristic). **h,** Trade-off between number of clusters and replicability (percent of clusters with minimal MetaNeighbor replicability score). Lamp5/Sncg/Vip/Sst/Pvalb - Major inhibitory neuron subclasses; L2-6 - layers; IT - Intratelencephalic; ET - Extratelencephalic; CT - Corticothalamic; NP - Near-projecting; Astro - Astrocytes; OPC - Oligodendrocyte precursor; Oligo - Oligodendrocytes; Micro - Microglial cells; SMC - Smooth muscle cells; VLMC - Vascular leptomeningeal cells; Peri - Pericyte; PVM - Perivascular macrophage; Endo - Endothelial


a
Median unique fragments/cell
Number of cells
scRNA SMART
scRNA 10x v3 A
scRNA 10x v2 A
snRNA SMART
snRNA 10x v3 B
snRNA 10x v3 A
snRNA 10x v2 A
snmC-seq
snATAC-seq
Cells
Nuclei
b
UMAP 2
UMAP 1
c
Sst Chodl
Endo
NP
Pvalb
Sst
Oligo
VLMC
CT
PVM
L6b
Astro
Micro
L6 IT
Lamp5
Sncg
SMC
Peri
L5 IT
Vip
Car3
L4/5 IT
L2-3 IT
L5 PT
UMAP 2
UMAP 1
e
Mean expression (normalized)
0
1
25% Cells expressing
50% gene
100%
Slc17a7
Slc30a3
Otof
Rorb
Rspo1
Fezf2
Sulf1
Car3
Fam84b
Sla2
Foxp2
Nxph4
Glutamatergic
L2/3 IT
L4/5 IT
L5 IT
L6 IT
L6 IT Car3
L5 ET
L5/6 NP
L6 CT
L6b
L5/6 NP
Gad1
Prox1
Lamp5
Ndnf
Sncg
Vip
Lhx6
Sst
Chodl
Pvalb
Vipr2
GABAergic
Lamp5
Lamp5 Lhx6
Sncg
Vip
Sst Chodl
Sst
Pvalb
Pvalb Vipr2
scRNA SMART
scRNA 10x v3 A
snRNA 10x v3 B
snRNA SMART
d
Cell type proportion
(normalized by class)
scRNA 10x v3 A
scRNA 10x v2 A
snRNA 10x v3 A
snRNA 10x v3 B
GABA
Glu
Other
Cell type
Cell subclass
Cell class
Lamp5
Sncg
Vip
Sst
Pvalb
IT
ET
CT
L6b
L5/6
Meis2
OPC
Astro
Oligo
VLMC
Endo
SMC
Peri
Micro
PVM
GABAergic (59 types)
Glutamatergic (31)
Other (26)
f
Malat1
Ywhaz
Fam84b
Max
Num. genes
0
Abundance in nuclear RNA
(log2(CPM+1), snRNA 10x v2 A)
Abundance in cellular RNA
(log2(CPM+1), scRNA 10x v2 A)
g
GABAergic
Glutamatergic
Non−Neuronal
Num. replicable clusters
% replicable clusters
Replicability (AUROC)
h
Replicability assessed in:
10X v3 vs. SMART−Seq
for scRNA
for snRNA
Single cell vs. single nuclei
for 10X v3 datasets
for SMART−Seq datasets
Percent replicable clusters
Number of clusters (Resolution)

Subsampling RNA-Seq datasets (Figure S2b, Table S1) showed that scRNA-Seq generally detects more genes per cell (up to ~7,100 median genes/cell for 10x, 10,000 for SMART) than snRNA-Seq (up to ~4,000 for 10x, 5,800 for SMART). The 10x v3 platform detected 60-100% more genes than 10x v2. The number of genes detected per cell in the snRNA-Seq 10x v3 B dataset (median ~4,000), using an improved nucleus isolation protocol[34] (see Methods), was significantly higher than the other snRNA-Seq datasets (1,700-3,500) and was similar to the scRNA-Seq 10x v3 dataset when compared at the same sequencing depth.

We created web resources to interactively access, explore, visualize, and analyze the raw and processed datasets (Figure S1g,h).

1.3.2. consensus transcriptomic atlas of MOp

To establish a transcriptomic reference atlas of mouse MOp we jointly analyzed 7 sc/snRNA-Seq datasets. The datasets were mutually consistent, with strongly correlated expression of cell-type marker genes (Figure S2a,d,e) despite different sensitivity to genes with low expression (Figure S2c). We used computational data integration (see Methods) to jointly cluster and identify 116 cell types using all the datasets (Figure 2b,c, Figure S2d, Table S2; Table S3). Importantly, cells and nuclei, assayed by each of the technologies and in each batch, grouped primarily by cell type and not by dataset (Figure 2b). Residual systematic differences between nuclear and cellular RNA-Seq assays were observed in some clusters as a gradient of transcriptomes from different datasets. We performed hierarchical clustering to uncover the relationships among types within each major cell class: GABAergic inhibitory neurons (n=59 types), glutamatergic excitatory neurons (n=31) and non-neurons (n=26) (Figure 2d). Six of the transcriptomic datasets used cell sorting strategies to enrich neurons relative to non-neuronal cells, while the largest dataset (snRNA 10x v3 B) represents an unbiased sample of both neuronal and non-neuronal cells. Despite these differences, the relative frequency of cell types was highly consistent across datasets after normalizing for the total sample of each major class (Table S3). 86 out of 116 cell types were present across all the datasets, while the rest were

non-neuronal types that were under-sampled in many datasets, or extremely rare types (< 0.01% of all cells).

To facilitate the use of these cell types by investigators, we adopted a nomenclature that incorporates multiple anatomic and molecular identifiers. For example, we identified four clusters of excitatory neurons (expressing *Slc17a7* encoding vesicular glutamate transporter Vglut1) that express a deep layer marker, *Fezf2*, as well as *Fam84b*, a unique marker of pyramidal tract (PT)[4] or extratelencephalic (ET) projecting neurons[35] (Figure 2e). Therefore, we label these neurons “L5 ET 1-4”. We divided GABAergic neurons into 5 major subclasses based on marker genes: *Lamp5*, *Sncg*, and *Vip* labeling caudal ganglionic eminence (CGE)-derived cells, and *Sst* and *Pvalb* which label medial ganglionic eminence (MGE)-derived cells. Finer distinctions among GABAergic types are identified by secondary markers (e.g. *Sst Myh8*). Tables of cluster accession IDs and differentially expressed genes between every pair of cell types help track the cell types and their underlying molecular evidence (Table S3; Table S6)[36].

We compared our MOp atlas with a large dataset of mouse anterolateral motor cortex (ALM) and primary visual cortex (VISp) neurons assayed by scRNA-seq (SMART-Seq) (Figure S3a)[4]. We found one-to-one matches between most of the 116 MOp cell types and the 102 previously defined in ALM. Four types of Layer 5 ET neurons correspond with 3 previously described deep layer excitatory neurons with distinct subcortical projection patterns to thalamus and medulla[37] (Figure S3b,c). These types, which were associated with distinct roles in movement planning and initiation, had consistent patterns of differential gene expression across the transcriptomic datasets (Figure S4).

The motor cortex is traditionally considered to lack a discernible layer 4 based on the absence of a clear cytoarchitectonic signature[38]. However, recent anatomical studies have identified a population of pyramidal cells located between layers 3 and 5, with hallmarks of L4 neurons including thalamic input and outputs to L4 and L2/3[25]. We identified two intratelencephalic (IT) clusters, containing over 99,000 cells, which express a combination of

markers usually associated with L4[39], including *Cux2*, *Rspo1* and *Rorb* (both clusters), and those associated with L5, *e.g.*, *Fezf2* (one cluster) (Figure 2e, Figure S5a). We confirmed the specificity of the expression of these genes in MOp by *in situ* hybridization (ISH, Figure S5b). These cells represent a substantial fraction (18% or more) of all excitatory neurons in each dataset. Therefore, we labeled these clusters L4/5. Moreover, the localization of cells with these gene markers in middle layers is further supported by spatial transcriptomics[17].

Using our integrated dataset, we directly compared the nuclear and cytoplasmic transcriptomes of MOp cells. Both modalities can achieve comparable clustering resolution (Figure S2d), as previously reported[40], but they provide distinct information about some cell types and transcripts. We found that the long non-coding RNA *Malat1* was enriched in snRNA-Seq, consistent with its nuclear localization[41] (Figure 2f, Figure S2f). By contrast, mRNA of the protein-coding gene *Ywhaz* was strongly depleted from the nucleus.

We used MetaNeighbor to assess the cross-dataset replicability of clusters defined separately using each of the seven transcriptomic datasets (Table S4)[42]. We found 70 clusters with a high replicability (AUROC > 0.7 across at least two datasets, Figure 2g). Most clusters had reciprocal best matches across all datasets (Figure S8a). By comparing results of three different widely used single-cell analysis packages[20,43,44], we found lower replicability for fine-grained partitions of cells into 30 or more clusters (Figure 2h). These results highlight the importance of careful biologically informed cluster analyses.

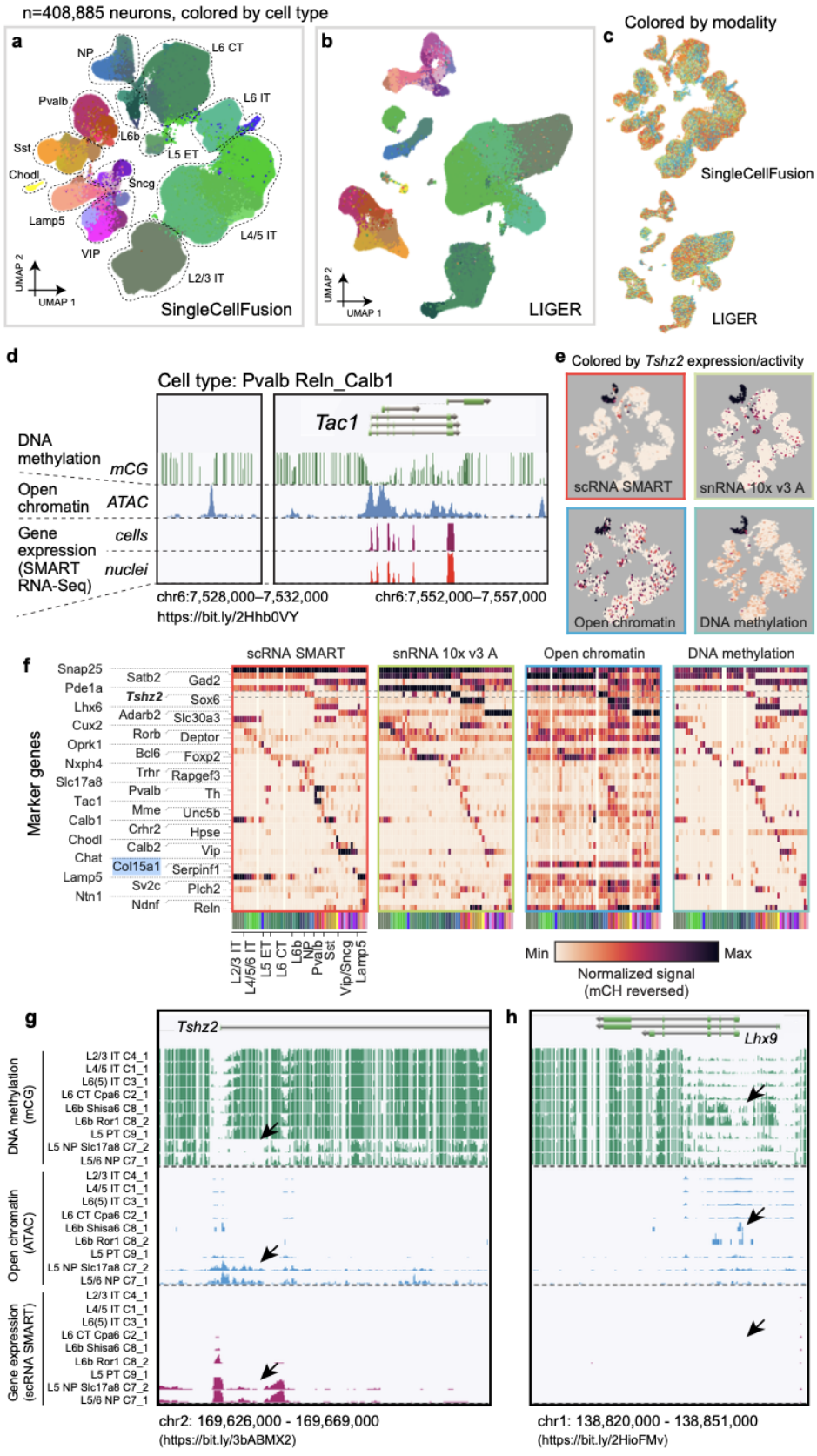
1.3.3. Integrating single-cell RNA and epigenomes

Regions of open chromatin and patterns of DNA methylation, including CG and non-CG methylation, are cell type-specific signatures of neuronal identity and can be assayed in single nuclei[27,29]. We applied snmC-seq2[32] (9,876 cells) and snATAC-Seq[45] (81,196 cells) assays to nuclei isolated from the same MOp samples. Independent analyses of each epigenomic dataset identified n=42 cell types using DNA methylation, and n=33 types using open chromatin

(Figure S6a-d; Table S4). Marker genes for major cell classes had corresponding patterns of cell type-specific depletion of non-CG methylation (low mCH, Figure S6b) and open chromatin in the gene body (Figure S6d).

We integrated 8 transcriptome and epigenome datasets using two computational methods (LIGER[21] and SingleCellFusion[11]) to produce a unified, multimodal cell census (Figure 3a-c, Figure S6e-j, 7a-b, Table S5). We reasoned that cells of the same type measured in each modality can be identified based on correlated gene-centric features. Gene expression is negatively correlated with gene body non-CG methylation[27] and positively related to the gene body ATAC-Seq read density[46]. Although distal regulatory elements (e.g. enhancers) were not used for dataset integration, they were subsequently analyzed at the level of integrated cell types.

Figure 3: Integration of epigenomes and transcriptomes. a,b, Two-dimensional projection (UMAP) of >400,000 individual cells and nuclei from 8 transcriptomic and epigenomic datasets (excluding snRNA 10x v2 A), integrated using SingleCellFusion (a) or LIGER (b). Cells are colored by joint clustering assignments from the respective integration method. **c,** UMAP projection with cells colored by dataset (color scheme as in Figure 2a). **d,** Browser view of the *Tac1* locus comparing four datasets with base-resolution transcriptomic and epigenomic information for one cell type: *Pvalb Reln_Calb1*. **e,** UMAP embeddings colored by mRNA expression level, accessibility, or DNA methylation at *Tshz2*. **f,** Selected marker genes across data modalities. **g,h** Browser showing excitatory cell type tracks. *Tshz2* consistently marks L5 NP cell types across data modalities (g), while *Lhx9* has a unique epigenetic signature in L6b cell types in DNA methylation only (h).


n=408,885 neurons, colored by cell type
a
NP
L6 CT
Pvalb
L6 IT
L6b
Sst
L5 ET
Chodl
Sncg
Lamp5
L4/5 IT
VIP
L2/3 IT
UMAP 2
UMAP 1
SingleCellFusion
b
LIGER
c
Colored by modality
SingleCellFusion
LIGER
d
Cell type: Pvalb Reln_Calb1
Tac1
DNA methylation
mCG
Open chromatin
ATAC
Gene expression (SMART RNA-Seq)
cells
nuclei
chr6:7,528,000–7,532,000
chr6:7,552,000–7,557,000
https://bit.ly/2Hhb0VY
e
Colored by Tshz2 expression/activity
scRNA SMART
snRNA 10x v3 A
Open chromatin
DNA methylation
f
scRNA SMART
snRNA 10x v3 A
Open chromatin
DNA methylation
Marker genes
Snap25
Pde1a
Lhx6
Cux2
Oprk1
Nxph4
Slc17a8
Tac1
Calb1
Chodl
Chat
Lamp5
Ntn1
Satb2
Tshz2
Adarb2
Rorb
Bcl6
Trhr
Pvalb
Mme
Crhr2
Calb2
Col15a1
Sv2c
Ndnf
Gad2
Sox6
Slc30a3
Deptor
Foxp2
Rapgef3
Th
Unc5b
Hpse
Vip
Serpinf1
Plch2
Reln
L2/3 IT
L4/5/6 IT
L5 ET
L6 CT
L6b
Pvalb
Sst
Vip/Sncg
Lamp5
Min
Max
Normalized signal
(mCH reversed)
g
Tshz2
DNA methylation (mCG)
Open chromatin (ATAC)
Gene expression (scRNA SMART)
L2/3 IT C4_1
L4/5 IT C1_1
L6(5) IT C3_1
L6 CT Cpa6 C2_1
L6b Shisa6 C8_1
L6b Ror1 C8_2
L5 PT C9_1
L5 NP Slc17a8 C7_2
L5/6 NP C7_1
chr2: 169,626,000 - 169,669,000
(https://bit.ly/3bABMX2)
h
Lhx9
chr1: 138,820,000 - 138,851,000
(https://bit.ly/2HioFMv)

By combining cells from integrated clusters into pseudo-bulk tracks, we obtained base-resolution epigenomic and transcriptomic information (Figure 3d,g,h; https://brainome.ucsd.edu/BICCN_MOp). To illustrate, we highlight the locus of *Tac1*, which encodes a precursor of the neuropeptide Substance P and marks a subset of medial ganglionic eminence (MGE)-derived interneurons[47]. We confirm *Tac1* mRNA expression in parvalbumin-expressing neurons marked by *Reln* and *Calb1*. We further observed accessible chromatin and low DNA methylation at CG sites within the body of the *Tac1* gene, and at a location ~24 kb upstream of the transcription start site (Fig 2d).

Both computational integration methods (LIGER and SingleCellFusion) identified 56 cell types, which showed a high degree of concordance between the methods and with the transcriptome-based consensus clusters (Figure S7a-d). Indeed, integrated analysis identified more cell types than the single-modality analysis of each epigenomic dataset, while largely concurring with the independent clusters (Figure S7b). Integration revealed striking examples of cross-modal cell type-specific signatures. For example, *Tshz2* is a specific marker of layer 5 near-projecting (NP) excitatory neurons, with low DNA methylation (mCG and mCH), open chromatin, and strong cell type-specific expression (Figure 3e,f,g). The close correspondence between transcriptomic and epigenomic signatures at *Tshz2,* and at 35 markers of other cell types, was evident across each of the datasets (Figure 3f). Importantly, these pseudo-bulk tracks include data, such as CG methylation and intergenic snATAC-Seq signals, that were not used for the multimodal computational integration.

In addition to concordant cross-modal signals, we also found loci where transcriptomic and epigenomic data diverged. For instance, at *Lhx9* we found high DNA methylation in L6b excitatory neurons, with little or no methylation in any other cell type (Figure 3h; Figure S7f). Despite this cell type-specific DNA methylation, we found no expression of *Lhx9* RNA in any cell type and no significant enrichment of ATAC-Seq reads. *Lhx9* has been implicated in early

developmental patterning of the caudal forebrain and may be transcriptionally silenced in the adult, potentially via Polycomb-mediated repression[48]. Other regulators of neural development, such as *Pax6* and *Dlx1/2*, have a similar epigenetic profile with cell type-specific hyper-methylation. This pattern may represent a vestigial epigenetic signature of embryonic development[49].

1.3.4. Epigenomic signatures of cell type-specific gene regulation

Epigenomic data identify potential regulatory regions, such as distal enhancers, marked by open chromatin and low DNA methylation (mCG). These modalities have complementary technical characteristics, such as the number of cells assayed (higher for open chromatin) and the genomic coverage per cell (higher for DNA methylation; Figure 2a). We first defined differentially methylated regions (DMRs) and chromatin accessibility peaks independently, identifying over 1.3 million DMRs covering 225 Mbp (8.3% of the genome) and 300,000 accessible regions (170 Mbp) (Figure 4a-b). In each cell type, a large fraction of accessible regions (28-89%) overlapped hypo-methylated DMRs (Figure 4a). By contrast, many DMRs did not overlap accessibility peaks (Figure 4b). In some cases, these DMRs coincided with broad open chromatin regions, such as whole gene bodies, which had no narrow ATAC peaks.

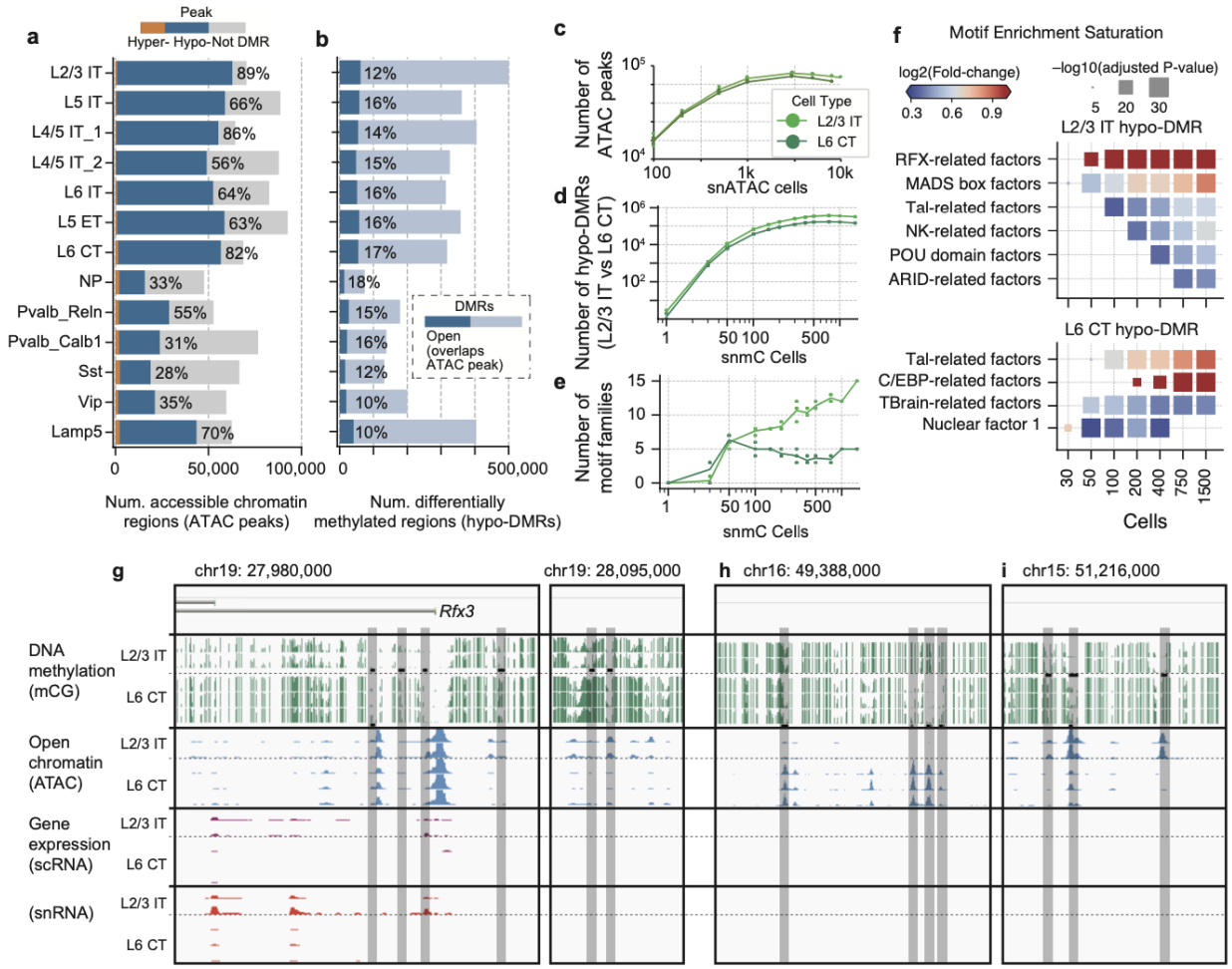


Figure 4: Epigenomic signatures of MOp regulatory elements. a,b, Regulatory regions were identified in each cell type using differentially methylated regions (DMRs, n=1,302,403) (a) and open chromatin regions (ATAC peaks, n=316,788) (b) in multimodal integrated clusters. **c,d** Saturation analysis for two excitatory subclasses shows the number of regulatory regions detected as a function of sampled cells**. e,** Saturation analysis of the number of transcription factor DNA binding sequence (TFBS) motifs enriched in each cell type's DMRs**. f,** Enrichment of motifs for selected TF families as a function of number of cells sampled. **g,** Browser views of loci containing cell type-specific regulatory elements (gray highlighted regions). The *Rfx3* gene is differentially expressed in L2/3 neurons, and has an enhancer specific to L2/3 located ~15 kb upstream of the promoter region. **h,i** Examples of intergenic regions with accessibility and demethylation specific to L6 CT (h) or L2/3 neurons (i).

By downsampling data from two abundant cell types (L2/3 IT and L6 CT), we found the number of detectable accessibility peaks saturated after sampling around 1,000 cells (Figure 4c). By contrast, the number of DMRs reached a plateau after sampling 200-300 cells (Figure 4d). Furthermore, the number of significantly enriched transcription factor (TF) motifs

increased with cell number (Figure 4e), although for L6 CT neurons it reached a plateau of ~5 key motif families after sampling ~100 cells.

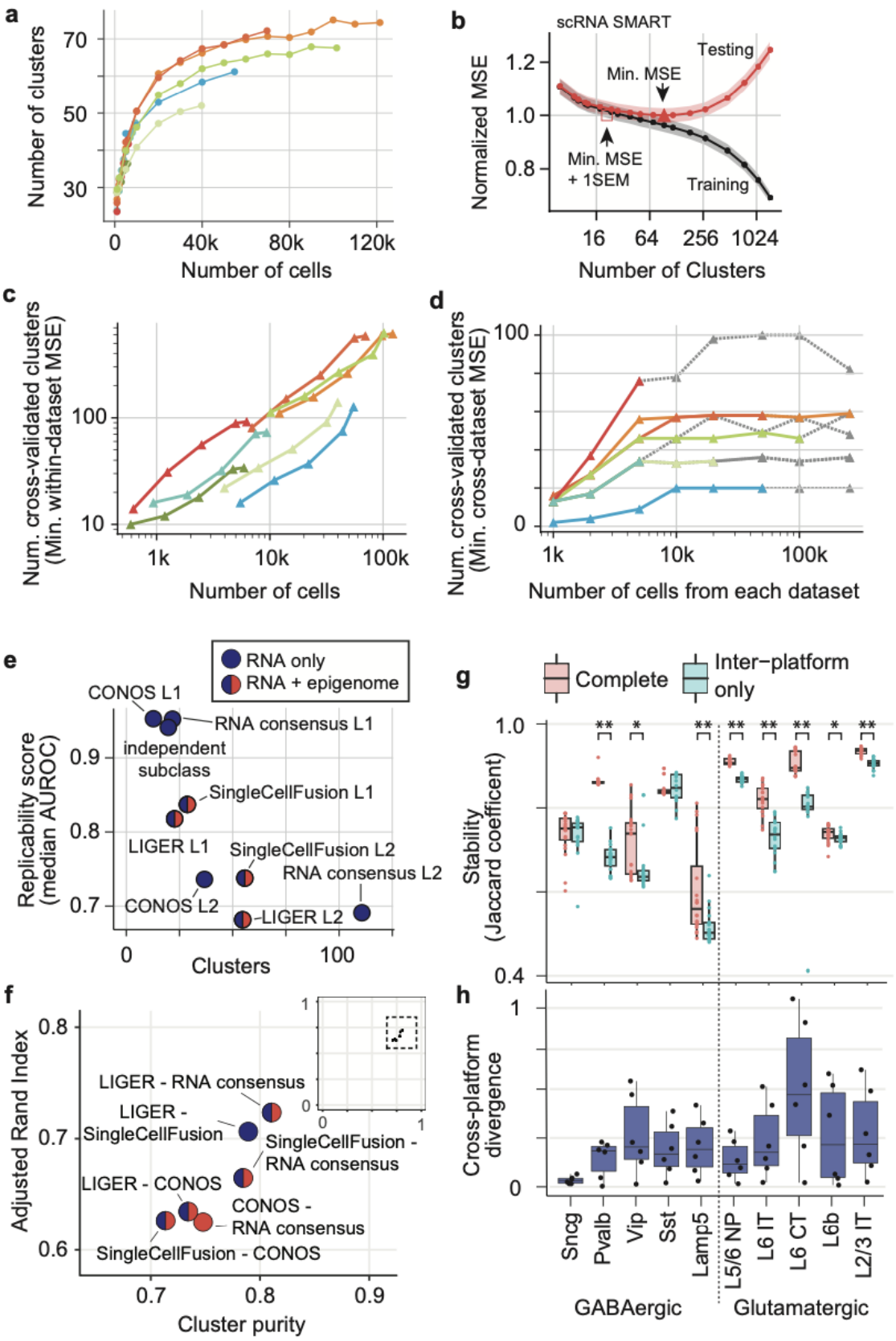
Combining both epigenomic datasets, we identified 250,000 putative enhancers with fine resolution (Table S7)[50]. Putative enhancers were often found in distal regions, at least 2 kb from the nearest transcription start site (Figure 4h,i). Sequence motifs of several TF families were enriched in each cell type (Figure 4f), such as *Rfx* motifs in L2/3 neurons. Using the transcriptomic data, we found *Rfx3*, but not other *Rfx* family members, was specifically enriched in L2/3 neurons and had low methylation and accessible chromatin in the gene body as well as ~15 kb upstream of the Rfx3 promoter (Figure 4g). These data suggest a key role for *Rfx3* in L2/3 neurons.

1.3.5. Reproducible cell types across datasets

Different molecular modalities, sampling strategies, sequencing technologies, and computational analysis procedures can lead to divergent estimates of the total number of cell types. We used systematic cross-datasets analyses to assess the statistical and biological reproducibility of cell types and constrain the range of plausible numbers of cell types based on current single-cell sequencing data.

We first addressed the impact of the number of sampled cells on the resolution of the cell atlas, by downsampling each dataset followed by clustering analysis with a fixed resolution parameter (Figure 5a). The number of detected neuronal cell types (clusters) increased logarithmically with cell number, with relatively few additional clusters detected after sampling ~80,000 cells or nuclei. Notably, the dependence of the number of clusters on the number of sampled cells was similar for all modalities and datasets, showing that the number of sampled cells is a key determinant of cluster resolution.

Figure 5: Robustness and reproducibility of cell types within and across datasets. **a,** Number of clusters estimated for each dataset after sampling a fraction of the total cells (Leiden clustering, resolution r=6; color scheme as in Figure 2a). **b,** Mean squared error (MSE) as a function of the number of clusters for scRNA SMART-seq. The minimum MSE and the min. MSE+1SEM defines a range of optimal cluster resolutions. Shaded region shows s.e.m. derived from cross-validation with n=5 random data partitions. **c,d,** Number of clusters estimated by within- (c) or across-dataset cross-validation (d) (n=5 data partitions). For cross-dataset comparison, the number of clusters is based on the minimum test MSE for one dataset after joint multimodal clustering. **e**, Trade-off between number of clusters and replicability (median MetaNeighbor AUROC) of consensus clustering methods applied at various resolutions. **f**, Agreement between consensus clustering results using different computational procedures. Inset: zoomed-out view showing that all methods have high cluster purity and adjusted Rand index. **g,** Transcriptomic platform consistency is assessed by cross-dataset cluster stability analysis (Conos) using complete networks, and using inter-platform edges only. Glutamatergic and Pvalb subclasses have reduced stability in inter-platform comparison. Data points show n=20 independent random samples, each containing 95% of the total cells. **h**, Cross-platform expression divergence (Jensen-Shannon) for major cell subclasses. Box-and-whisker plots (g,h) show the median, interquartile range (IQR: 25-75th percentile), and the smaller of the data range (min-max), or 1.5 times the IQR. *FDR≤0.05, **: FDR≤0.0001, Wilcoxon rank sum test, Benjamini-Hochberg correction.


a
Number of clusters
70
50
30
0
40k
80k
120k
Number of cells
b
scRNA SMART
Testing
Min. MSE
Min. MSE
+ 1SEM
Training
Normalized MSE
1.2
1.0
0.8
16
64
256
1024
Number of Clusters
c
Num. cross-validated clusters
(Min. within-dataset MSE)
100
10
1k
10k
100k
Number of cells
d
Num. cross-validated clusters
(Min. cross-dataset MSE)
100
0
1k
10k
100k
Number of cells from each dataset
e
RNA only
RNA + epigenome
CONOS L1
RNA consensus L1
independent
subclass
SingleCellFusion L1
LIGER L1
SingleCellFusion L2
RNA consensus L2
CONOS L2
LIGER L2
Replicability score
(median AUROC)
0.9
0.8
0.7
0
100
Clusters
f
Adjusted Rand Index
0.8
0.7
0.6
LIGER - RNA consensus
LIGER -
SingleCellFusion
SingleCellFusion -
RNA consensus
LIGER - CONOS
CONOS -
RNA consensus
SingleCellFusion - CONOS
0.7
0.8
0.9
Cluster purity
g
Complete
Inter−platform
only
Stability
(Jaccard coefficent)
1.0
0.4
h
Cross-platform
divergence
1
0
Sncg
Pvalb
Vip
Sst
Lamp5
L5/6 NP
L6 IT
L6 CT
L6b
L2/3 IT
GABAergic
Glutamatergic

Any dataset can be divided into increasingly fine-grained clusters, yet they may not reflect biologically meaningful or reproducible cell type distinctions. We used cross-validation to objectively measure the generalizability of cluster-based descriptions of the data (Figure S8b). We first used within-dataset cross-validation, dividing the features (genes or genomic bins) into clustering and validation sets. After clustering all cells using the clustering feature set, we split the cells into training and test sets. We use the training cells to learn the validation set features for each cluster. Finally, we compare the validation set features with the held out data for test cells to measure the mean squared error (MSE). We applied this procedure to each dataset with a range of clustering resolutions, resulting in a U-shaped cross-validation curve for test set error as a function of the number of clusters (Figure 5b, Figure S8c,d). The location of the minimum MSE is an estimate of the number of reliable clusters. Finally, we repeated this cross-validation procedure for each dataset in combination with systematic downsampling (Figure 5c).

All of the datasets (except snRNA SMART-seq) supported ~100 or more cell types when a sufficient number of cells were sampled. The number of cells required to achieve this resolution was larger for snATAC-Seq (with few reads per cell) compared with RNA-Seq or snmC-Seq. This observation is consistent with the relative sparseness of the snATAC-seq data. We further found that sc/snRNA-Seq datasets with the largest numbers of cells could support very high cluster resolution with up to ~600 clusters. Our cross-validation analysis shows that these fine-grained clusters capture genuine transcriptomic structure which is correlated and replicable across cells and across genomic features. However, at least some of this structure likely corresponds to continuous variation within discrete cell types, rather than discrete cell type categories[51]. Moreover, the cross-validation analysis shows no sharp error minimum at a particular value of the number of clusters. Instead, the U-shaped cross-validation curve has a broad basin covering a range of plausible values (Figure 5b, Figure S8c,d).

To more stringently test the reproducibility of cell types, we performed cross-dataset cross-validation (Figure S8b). This procedure uses a randomly chosen half of genomic features to perform data integration and joint analysis of eight datasets using SingleCellFusion. Next, we use the joint cluster labels to perform cross-validation in each dataset, as in the within-dataset procedure above. This analysis supported a maximum resolution of ~100 clusters when testing using the scRNA SMART-seq data (Figure 5d).

As an alternative to joint analysis of multiple datasets, which could potentially discern spurious correlations due to computational data integration, we also took a more stringent approach to cross-validation. Using the independent cluster analysis of each dataset, we performed MetaNeighbor analysis to assess the replicability of clusters[42]. We found that the median replicability score for all clusters was high (AUROC > 0.8) for integrated analyses with coarse resolution (<50 clusters, level 1 (L1) analyses; Figure 5e). The more fine-grained joint analyses (L2, 50-120 clusters) were also largely supported by MetaNeighbor, but with a lower median replicability score around 0.7. Notably, we found a high degree of consistency in the results of joint cluster analysis when using different computational methods (Figure 5f).

Finally, we explored whether MOp cell-type signatures were stable across different sc/snRNA-Seq platforms. Using four RNA-Seq datasets (scRNA SMART, snRNA SMART, scRNA 10x v3 A, and snRNA 10x v3 A), we performed clustering on network of samples (Conos[52]) to link cells across datasets and determine joint clusters. We compared the clustering results based on inter-platform network connections only vs. results that also included connections across datasets of the same platform (Figure S8e). Most neuron types, except Pvalb and L6 CT, had only a modest difference in cluster stability using both approaches (Figure 5g) and a low level of inter-platform divergence in their cell type transcriptomic signatures (Figure 5h).

1.4. Discussion

Our mouse primary motor cortex (MOp) cell atlas represents the most comprehensive, integrated collection of single-cell transcriptome and epigenome datasets for a single brain region to this date. We generated a high-resolution consensus transcriptomic cell type taxonomy that integrates seven sc/snRNA-Seq datasets collected from MOp with six experimental methods. Our transcriptomic taxonomy is highly consistent with a previously published transcriptomic cell census from the primary visual (VISp) and anterolateral motor (ALM) cortices based on SMART-Seq alone[4]. We found that gene expression profiles are largely consistent across methodologies, while providing complementary information about particular genes such as nucleus-enriched transcripts. The MOp atlas demonstrates the power of a two-pronged strategy combining broad sampling of diverse cell types (e.g. 10x with large number of cells and shallow sequencing) with deep sequencing (e.g. SMART-Seq) to precisely characterize gene expression profiles for each cell type. This strategy should guide future cell census efforts, by the BICCN and others, at the scale of whole brains and in other species.

We further demonstrate multimodal integration of transcriptome (sc/snRNA-Seq), DNA methylation (snmC-Seq2), and chromatin accessibility (snATAC-Seq) datasets using two computational methods (SingleCellFusion and LIGER). It is possible to directly establish links between molecular modalities through simultaneous measurement of multiple signatures in the same cell[53]. However, multimodal single-cell assays remain challenging and often sacrifice the depth or resolution of data in each modality compared with single modality assays. Moreover, it is important to show that data collected from different animals, across different laboratories and using different experimental platforms and assays, nevertheless can be integrated within a unified cell type atlas. By correlating mRNA transcripts, gene-body methylation and accessibility peaks, we showed that different types of data can be integrated without sacrificing the resolution of >50 fine-grained neuron types. Integrative analysis of transcriptional and epigenetic

signatures of cell identity will enable development of tools based on cell type-specific enhancers for cell targeting and manipulation.

Our data provide new insights into the molecular architecture of MOp cell types. The neuropeptide Substance P precursor, *Tac1,* marks a subset of Pvalb cells and is strongly upregulated in rodent MOp following motor learning[47,54]. We found that *Tac1* is expressed in two subtypes of MOp interneurons (Pvalb_Calb1 and Pvalb_Reln), and our epigenomic data identified a cell type-specific enhancer ~24 kb upstream of the gene promoter. We provide new evidence that MOp harbors an excitatory neuron population expressing markers of layer 4 thalamic-recipient neurons, including *Cux2*, *Rspo1* and *Rorb*[25]. The laminar distribution of these cells has been confirmed by ISH of these marker genes and in a parallel study by MERFISH[17]. This discovery revises the traditional understanding of MOp as an agranular cortex lacking L4. We also found multiple types of L5 ET neurons that align with recently described populations with distinct subcortical projection targets[37]. We further identified networks of gene expression regulatory elements, marked by overlapping regions of open chromatin and cell type-specific demethylation, harboring sequence motifs that identify the key transcriptional regulators. For example, by combining epigenetic and gene expression data we identified *Rfx3* as a candidate factor for L2/3 IT cells. We also identified genes with non-canonical regulatory signatures, such as enrichment of mCG in *Lhx9* specifically in L6b excitatory cells.

We took advantage of the unprecedented diversity of large-scale datasets, generated in a coordinated fashion from mouse MOp, to critically evaluate the robustness and reliability of the cell type taxonomies obtained by clustering molecular datasets. Our cross-validation analysis of individual datasets and multimodal integration objectively constrains the range of cluster resolutions supported by the data without overfitting. Rather than supporting a single, definitive number of cell types in mouse MOp, our studies instead point to a range of cluster resolutions spanning from ~30 to 116 cell types that are supported by the data. Indeed, discrete

cell type categories may be an inappropriate description at a fine-grained level of analysis, where the cells' molecular profiles vary along a continuum.

By integrating nine large-scale single-cell transcriptome and epigenome datasets, we have comprehensively classified and annotated the diversity of cell types in the adult mouse primary motor cortex (MOp). Our study demonstrates general procedures for objective cross-dataset comparison and statistical reproducibility analysis, as well as standards and best practices that can be adopted for future large-scale studies. Together with complementary BICCN datasets from spatial transcriptomics, connectivity and physiology, as well as cross-species comparative studies, our results help to establish a multi-faceted understanding of brain cell diversity. Targeted studies of individual cell types, taking advantage of the transcriptional and epigenetic signatures described here, will define their functional roles and significance in the context of neural circuits and behavior. Integrative analyses will be essential to make progress toward understanding the organizing principles of brain cell types through their molecular genetic signatures.

1.5. Methods

1.5.1. Tissue collection and isolation of cells or nuclei (RNA-Seq at Allen Institute)

The following methods apply to the following transcriptomic datasets generated at the Allen Institute: scRNA SMART, scRNA 10x v3 A, scRNA 10x v2 A, snRNA SMART, snRNA 10x v3 A, and snRNA 10x v2 A.

Mouse breeding and husbandry: All procedures were carried out in accordance with Institutional Animal Care and Use Committee protocols at the Allen Institute for Brain Science. Mice were provided food and water *ad libitum* and were maintained on a regular 12-h day/night cycle at no more than five adult animals per cage. Ambient temperature was set to 72°F and relative humidity was set to 40%. All rooms are on 12/12 hour light/dark cycle. For this study, we enriched for neurons by using *Snap25-IRES2-Cre* mice[55] (MGI:J:220523) crossed to *Ai14*[56] (MGI: J:220523), which were maintained on the C57BL/6J background (RRID:IMSR_JAX:000664). Animals were euthanized at 53−59 days of postnatal age. Tissue was collected from both males and females (scRNA SMART, snRNA SMART, scRNA 10x v3 A, snRNA 10x v2 A), only males (scRNA 10x v2 A) or only females (snRNA 10x v3 A).

Single-cell isolation: We isolated single cells by adapting previously described procedures[4,57]. The brain was dissected, submerged in ACSF[4], embedded in 2% agarose, and sliced into 250-µm (SMART-Seq) or 350-µm (10x Genomics) coronal sections on a compresstome (Precisionary Instruments). The Allen Mouse Brain Common Coordinate Framework version 3 (CCFv3, RRID:SCR_002978) [58] ontology was used to define MOp for dissections (Figure S1b).

For SMART-Seq, MOp was microdissected from the slices and dissociated into single cells with 1 mg/ml pronase (Sigma P6911-1G) and processed as previously described[4]. For 10x Genomics, tissue pieces were digested with 30 U/ml papain (Worthington PAP2) in ACSF for 30 mins at 30 °C. Enzymatic digestion was quenched by exchanging the papain solution three times with quenching buffer (ACSF with 1% FBS and 0.2% BSA). The tissue pieces in the

quenching buffer were triturated through a fire-polished pipette with 600-μm diameter opening approximately 20 times. The solution was allowed to settle and supernatant containing single cells was transferred to a new tube. Fresh quenching buffer was added to the settled tissue pieces, and trituration and supernatant transfer were repeated using 300-μm and 150-μm fire polished pipettes. The single cell suspension was passed through a 70-μm filter into a 15-ml conical tube with 500 ul of high BSA buffer (ACSF with 1% FBS and 1% BSA) at the bottom to help cushion the cells during centrifugation at 100xg in a swinging bucket centrifuge for 10 minutes. The supernatant was discarded, and the cell pellet was resuspended in a quenching buffer.

All cells were collected by fluorescence-activated cell sorting (FACS, BD Aria II, RRID: SCR_018091) using a 130-μm nozzle. Cells were prepared for sorting by passing the suspension through a 70-μm filter and adding DAPI (to the final concentration of 2 ng/ml). Sorting strategy was as previously described[4], with most cells collected using the tdTomato-positive label. For SMART-Seq, single cells were sorted into individual wells of 8-well PCR strips containing lysis buffer from the SMART-Seq v4 Ultra Low Input RNA Kit for Sequencing (Takara 634894) with RNase inhibitor (0.17 U/μl), immediately frozen on dry ice, and stored at −80 °C. For 10x Genomics, 30,000 cells were sorted within 10 minutes into a tube containing 500 μl of quenching buffer. Each aliquot of 30,000 sorted cells was gently layered on top of 200 μl of high BSA buffer and immediately centrifuged at 230xg for 10 minutes in a swinging bucket centrifuge. Supernatant was removed and 35 μl of buffer was left behind, in which the cell pellet was resuspended. The cell concentration was quantified, and immediately loaded onto the 10x Genomics Chromium controller.

1.5.2. Tissue collection & nuclei isolation (RNA-Seq at Broad Institute)

These methods apply to the dataset snRNA 10x v3 B, generated at the Broad Institute. *Animal housing:* Animals were group housed with a 12-hour light-dark schedule and allowed to acclimate to their housing environment for two weeks post arrival. Ambient temperature was set

to 70°F ± 2°F and relative humidity was set to 40% ± 10%. All rooms are on 12/12 hour light/dark cycle. All procedures involving animals at MIT were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals under protocol number 1115-111-18 and approved by the Massachusetts Institute of Technology Committee on Animal Care. All procedures involving animals at the Broad Institute were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals under protocol number 0120-09-16. Samples were collected from both male and female mice.

Brain preparation prior to 10x nuclei sequencing: At 60 days of age, C57BL/6J mice were anesthetized by administration of isoflurane in a gas chamber flowing 3% isoflurane for 1 minute. Anesthesia was confirmed by checking for a negative tail pinch response. Animals were moved to a dissection tray and anesthesia was prolonged via a nose cone flowing 3% isoflurane for the duration of the procedure. Transcardial perfusions were performed with ice cold pH 7.4 HEPES buffer containing 110 mM NaCl, 10 mM HEPES, 25 mM glucose, 75 mM sucrose, 7.5 mM $MgCl_2$, and 2.5 mM KCl to remove blood from brain and other organs sampled. The brain was removed immediately and frozen for 3 minutes in liquid nitrogen vapor and moved to -80°C for long term storage. A detailed protocol is available at protocols.io[34].

Generation of MOp nuclei profiles: Frozen mouse brains were securely mounted by the cerebellum onto cryostat chucks with OCT embedding compound such that the entire anterior half including the primary motor cortex (MOp) was left exposed and thermally unperturbed. Dissection of 500 µm anterior-posterior (A-P) spans of the MOp (Figure S1c) was performed by hand in the cryostat using an ophthalmic microscalpel (Feather safety Razor #P-715) precooled to -20°C and donning 4x surgical loupes. Each excised tissue dissectate was placed into a pre-cooled 0.25 ml PCR tube using pre-cooled forceps and stored at -80°C. In order to assess dissection accuracy, 10 µm coronal sections were taken at each 500 µm A-P dissection junction and imaged following Nissl staining. Nuclei were extracted from these frozen tissue dissectates

using gentle, detergent-based dissociation, according to a protocol (available at protocols.io) adapted from one generously provided by the McCarroll lab, and loaded into the 10x Chromium v3 system. Reverse transcription and library generation were performed according to the manufacturer's protocol.

This 10x v3 snRNA-seq protocol resulted in a higher number of genes recovered compared to other snRNA-seq methods. We believe there are three reasons, and that the summation of benefits imparted by the combination of these accounts for the outcome.

First, mouse brains are perfused with a solution emulating artificial CSF and then rapidly frozen over liquid nitrogen vapor in such a way that RNA integrity is highly preserved. The resulting bioanalyzer RIN scores of the starting brain tissues are routinely 9.8. Storage of the brains before dissection is at -80^{o}C in the presence of a hydration sink of 1ml of OCT compound pre-frozen into the bottom of a 5ml storage tube. This prevents sublimation and subsequent desiccation-dependent RNA fragmentation.

Second, we performed expeditious sample processing. We have a well-trained group of technicians who process the mouse brain (as above) and then perform the dissociation and FACS and 10X processing (as below) in one continuous protocol without pauses. For example, each mouse is perfused and ready for dissection within minutes (10), and we limit our sample size to 6 such that no sample is waiting to move through the process.

Third, the frozen tissue snRNA Seq protocol incorporates two main features that we believe are important to quality because they prevent the nuclei from "leaking" valuable signal and simultaneously contaminating the barcoded nuclei mixture with exogenous RNA signal. Feature one is a very low level of centrifugation, which we have found to cause both loss of signal and increased exogenous signal. Feature two is the inclusion of an excipient reagent, BASF Kollidon VA-64, as per the McCarroll Lab protocol [59].

1.5.3. Tissue collection and isolation of nuclei for epigenomic samples

The following methods apply to the datasets snmC-Seq and snATAC-seq generated at the Salk Institute and University of California, San Diego.

Tissue preparation for nuclei production: Procedures involving animals at The Salk Institute were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals under protocol number 18-00006 and approved by the Institutional Animal Care and Use Committee. Male C57BL/6J mice were purchased from Jackson laboratories at 8 weeks of age and maintained in the Salk animal barrier facility on 12-hr dark-light cycles with controlled temperature (20-22 Celcius range) and humidity (30-70% range), and food ad libitum for one week before dissection.

Brains were extracted from 56-63 day old mice and immediately sectioned into 0.6 mm coronal sections, starting at the frontal pole, in ice-cold dissection media[27]. The primary motor cortex (MOp) was dissected from slices 2 through 5 along the anterior-posterior axis according to the Allen Brain reference Atlas (Figure S1d). Slices were kept in ice-cold dissection media during dissection and immediately frozen in dry ice for subsequent pooling and nuclei production. For nuclei isolation, the MOp dissected regions from 15-23 animals were pooled for each biological replicate, and two replicates were processed for each region. Nuclei were isolated by flow cytometry as described in previous studies[27,28]. Briefly, nuclei were produced by homogenization in sucrose buffer as described[27], and the nuclei pellet produced was divided into two aliquots. One aliquot underwent sucrose gradient purification and NeuN labeling (snmC-Seq), and the second went directly to tagmentation (snATAC-seq).

Bisulfite conversion and library preparation for snmC-Seq2: Detailed methods for bisulfite conversion and library preparation are previously described for snmC-Seq2[32], and the protocol is available on protocols.io[60]. The snmC-Seq2 libraries were sequenced using an Illumina Novaseq 6000 instrument (RRID:SCR_016387) with S4 flowcells and 150 bp paired-end mode.

snATAC-seq data generation: Combinatorial barcoding single nucleus ATAC-seq was performed as described previously[45,61]. 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 cOmpleteTM, EDTA-free protease inhibitor cocktail (Roche) in PBS) and pelleted again (500 x g, 5 min, 4°C; 5920R, Eppendorf, RRID:SCR_018092). 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[61] were added using a BenchSmart™ 96 (Mettler Toledo, RRID:SCR_018093), mixed five times and incubated for 60 min at 37 °C with shaking (500 rpm). To inhibit the Tn5 reaction, 10 µL of 40 mM EDTA were added to each well with a BenchSmart™ 96 (Mettler Toledo) and the plate was incubated at 37 °C for 15 min with shaking (500 rpm). Next, 20 µL 2 x sort buffer (2 % BSA, 2 mM EDTA in PBS) were added using a BenchSmart™ 96 (Mettler Toledo). All wells were combined into a FACS tube and stained with 3 µM Draq7 (Cell Signaling). Using a SH800 (Sony), 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, RRID:SCR_018094). After 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 (Beckmann Coulter, 0.55x and 1.5x). Libraries were purified one

more time with SPRI Beads (Beckmann Coulter, 1.5x). Libraries were quantified using a Qubit fluorimeter (Life technologies, RRID:SCR_018095) and the nucleosomal pattern was verified using a Tapestation (High Sensitivity D1000, Agilent). The library was sequenced on a HiSeq2500 sequencer (Illumina, RRID:SCR_016383) using custom sequencing primers, 25% spike-in library and following read lengths: 50 + 43 + 37 + 50 (Read1 + Index1 + Index2 + Read2)[29].

1.5.4. Genomic library preparation, sequencing and data processing

Single cell and single nucleus RNA-Seq (Allen Institute)

For SMART-Seq processing, we performed the procedures with positive and negative controls as previously described[4]. The SMART-Seq v4 (SSv4) Ultra Low Input RNA Kit for Sequencing (Takara Cat# 634894) was used to reverse transcribe poly(A) RNA and amplify full-length cDNA. Samples were amplified for 18 cycles in 8-well strips, in sets of 12–24 strips at a time. All samples proceeded through Nextera XT DNA Library Preparation (Illumina Cat# FC-131-1096) using Nextera XT Index Kit V2 (Illumina Cat# FC-131-2001) and a custom index set (Integrated DNA Technologies). Nextera XT DNA Library prep was performed according to manufacturer's instructions, with a modification to reduce the volumes of all reagents and cDNA input to 0.4x or 0.5x of the original protocol.

For 10x v2 processing, we used Chromium Single Cell 3' Reagent Kit v2 (10x Genomics Cat# 120237). We followed the manufacturer's instructions for cell capture, barcoding, reverse transcription, cDNA amplification, and library construction. We targeted sequencing depth of 60,000 reads per cell.

For 10x v3 processing, we used the Chromium Single Cell 3' Reagent Kit v3 (10x Genomics Cat# 1000075). We followed the manufacturer's instructions for cell capture, barcoding, reverse transcription, cDNA amplification, and library construction. We targeted sequencing depth of 120,000 reads per cell.

RNA-Seq data processing and QC (Allen)

Processing of SMART-Seq v4 libraries was performed as described previously[4]. Briefly, libraries were sequenced on an Illumina HiSeq2500 platform (paired-end with read lengths of 50 bp) and Illumina sequencing reads were aligned to GRCm38.p3 (mm10) using a RefSeq annotation gff file retrieved from NCBI on 18 January 2016 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/). Sequence alignment was performed using STAR v2.5.3[62]. PCR duplicates were masked and removed using STAR option 'bamRemoveDuplicates'. Only uniquely aligned reads were used for gene quantification. Gene counts were computed using the R GenomicAlignments package (RRID:SCR_018096)[63] and summarizeOverlaps function in 'IntersectionNotEmpty' mode for exonic and intronic regions separately. For the SSv4 dataset, we only used exonic regions for gene quantification. Cells that met any one of the following criteria were removed: < 100,000 total reads, < 1,000 detected genes (CPM > 0), < 75% of reads aligned to the genome, or CG dinucleotide odds ratio > 0.5. Cells were classified into broad classes of excitatory, inhibitory, and non-neuronal based on known markers, and cells with ambiguous identities were removed as doublets[4].

10x v2 and 10x v3 libraries were sequenced on Illumina NovaSeq 6000 (RRID:SCR_016387) and sequencing reads were aligned to the mouse pre-mRNA reference transcriptome (mm10) using the 10x Genomics CellRanger pipeline (version 3.0.0, RRID:SCR_017344) with default parameters. Cells were classified into broad classes of excitatory, inhibitory, and non-neuronal based on known markers. Low-quality cells that fit the following criteria were filtered from clustering analysis. Different filtering criteria were used for neurons and non-neuronal cells as neurons are bigger than non-neuronal cells and contain more transcripts. For scRNA datasets, we excluded neurons with fewer than 2000 detected genes and non-neuronal cells with fewer than 1000 detected genes; for snRNA datasets, we excluded neurons with fewer than 1000 detected genes and non-neuronal cells with fewer than

500 detected genes. Doublets were identified using a modified version of the DoubletFinder algorithm[64] and removed when doublet score > 0.3.

Chromatin accessibility (snATAC-Seq) data pre-processing (UCSD)

Paired-end sequencing reads were demultiplexed and aligned to mm10 reference genome using bwa[65]. After alignment, we converted paired-end reads into fragments and for each fragment, we checked the following attributes: 1) mapping quality score MAPQ; 2) whether two ends are appropriately paired according to the alignment flag information; 3) fragment length. We only keep the properly paired fragments whose MAPQ (--min-mapq) is greater than 30 with fragment length less than 1000bp (--max-flen). Because the reads have been sorted based on the names, fragments belonging to the same cell (or barcode) are naturally grouped together which allows for removing PCR duplicates. After alignment and filtration, we used Snaptools (https://github.com/r3fang/SnapTools, RRID:SCR_018097) to generate a snap-format file that contains metadata, cell-by-bin count matrices of a variety of resolutions, cell-by-peak count matrix.

Filtering cells by TSS enrichment and unique fragments: The method for calculating enrichment at TSS was adapted from a previously described method [66]. TSS positions were obtained from the GENCODE database (RRID:SCR_014966). Briefly, Tn5 corrected insertions were aggregated +/-2,000 bp relative (TSS strand-corrected) to each unique TSS genome wide. Then this profile was normalized to the mean accessibility +/-1,900-2,000 bp from the TSS and smoothed every 11bp. The max of the smoothed profile was taken as the TSS enrichment. We excluded any single cells that had fewer than 1,000 unique fragments or a TSS enrichment of <10 for any sample sets.

Doublet removal: After filtering out low-quality nuclei, we used Scrublet (RRID:SCR_018098)[67] to remove potential doublets for every sample set. Cell-by-peak count matrices are used as input, with default parameters.

DNA methylation (snmC-Seq) data pre-processing (Salk)

Mapping and feature count pipeline for snmC-Seq2: We implemented a versatile mapping pipeline (cemba-data.rtfd.io) for all the single-cell methylome based technologies developed by our group[11,27,32]. The main steps of this pipeline included: 1) Demultiplexing FASTQ files into single-cell files; 2) Reads level QC; 3) Mapping; 4) BAM file processing and QC; 5) final molecular profile generation. The details of the five steps for snmC-seq2 were described previously[68]. We mapped all the reads onto the mouse mm10 genome. After mapping, we calculated the methyl-cytosine counts and total cytosine counts in two sets of genome regions for each cell: the non-overlapping 100kb bins tiling the mm10 genome, which was used for methylation-based clustering analysis, and gene body regions ± 2kb, which is used for cluster annotation and cross modality integration.

Quality control and cell filtering: We filtered the cells based on five quality metrics: 1) The rate of bisulfite non-conversion as estimated by the rate of methylation at CCC positions (mCCC) < 0.03. mCCC rate reliably estimates the upper bound of bisulfite non-conversion rate[27], 2) overall mCG rate > 0.5, 3) overall mCH rate < 0.2, 4) total final reads (combining R1 and R2) > 500,000, 5) Total mapping rate (using Bismark[69]) > 0.5.

Preprocessing and clustering: The clustering steps of snmC-seq2 data were described previously[11]. In brief, we calculated posterior mCH and mCG rate based on beta-binomial distribution for the non-overlapping 100kb bins matrix, we then selected top 3000 highly variable features to perform PCA and find dominant PCs for mCH and mCG separately. We concatenate PCs from both methylation types together to construct a KNN graph, and ran the Leiden community detection algorithm[70] repeatedly to get the consensus clustering results. The stopping criteria of clustering considered number of marker genes, accuracy of the reproducible supervised model based on the cluster assignments, and minimum cluster size. We performed the clustering in two iterations to get major types and fine-grained types for comparison with other modalities in further integration.

1.5.5. Estimation of library size

For each dataset, we estimated the total library size, i.e. the number of unique RNA or DNA fragments (F), based on the rate of duplicate sequence reads. The number of unique mapped reads is $N_{unique} = F(1 - Bin[0|S, 1/F]) = F[1 - (1 - 1/F)^S]$, where S is the total number of sequenced reads. Using this equation, we numerically solved for F using the median values of S, N_{unique}.

1.5.6. Transcriptome analysis

Clustering individual datasets. Clustering for each sc/snRNASeq dataset was performed independently using the R package *scrattch.hicat*[4] (RRID:SCR_018099, available at https://github.com/AllenInstitute/scrattch.hicat). This package supports iterative clustering by making successively finer splits while ensuring all pairs of clusters, even at the finest level, are separable by stringent differential gene expression criteria[4]. For the scRNA 10x datasets, we used q1.th = 0.4, q.diff.th=0.7, de.score.th=150, min.cells=10. For the snRNA 10x datasets, we used q1.th=0.3, q.diff.th=0.7, de.score.th=100, min.cells=10. For the scRNA SMART datasets, we used q1.th = 0.5, q.diff.th=0.7, de.score.th=150, min.cells=4. For the snRNA SMART dataset, we used q1.th=0.4, q.diff.th=0.7, de.score.th=100, min.cells=4. We further performed consensus clustering by repeating iterative clustering on a subsample of 80% of cells, resampled 100 times, followed by final clustering based on the co-clustering probability matrix. Using this procedure, we could fine tune cluster boundaries as well as assess cluster uncertainty.

Next, we removed low-quality and doublet-driven clusters. We performed differential gene expression analysis between every pair of clusters within each subclass. If any cluster had ≤2 up-regulated genes (fold-change>2, FDR<0.01, with additional dataset-specific parameters listed in the previous paragraph) compared to another cluster, and had a substantially lower average number of detected genes per cell, we flagged the cluster as low-quality and removed it

from further analysis. Next, if the up-regulated genes between any two clusters within a subclass were predominantly marker genes for a different subclass, and one of the clusters had significantly higher average genes detected per cell and UMI count, we flagged the cluster as a potential doublet cluster and removed it from further analysis. These criteria led to the exclusion of 8.3% of all cells, the vast majority of which came from the two 10x v3 datasets (scRNA 10X v3 A, snRNA 10X v3 B). While the 10X v3 platform boosts the gene detection for good cells, it does the same to damaged cells or debris, leading to an elevated number of clusters that were excluded for these datasets.

Joint clustering of multiple datasets. To provide a consensus cell-type taxonomy across all transcriptomic datasets, we developed a novel integrative clustering analysis across multiple data modalities. This procedure is available via the *harmonize* function of the *scrattch.hicat* package. Unlike Seurat/CCA[71], which aims to find aligned common reduced dimensions across multiple datasets, this method directly builds a common adjacency graph using the cells from all datasets, and then applies the Louvain community detection algorithm[72]. We extended the cluster merging algorithm in the *scrattch.hicat* package to ensure that all clusters can be separated by conserved DE genes across platforms. The *i_harmonize* function, similar to the *iter_clust* function in the single dataset clustering pipeline, applies integrative clustering across datasets iteratively while ensuring that all the clusters at each iteration are separable by conserved DE genes.

To build a common adjacency matrix incorporating samples from all the datasets, we first chose a subset of datasets which we used as “reference datasets.” For this study, we used 10x v2 single cell dataset from Allen (scRNA 10x v2 A) and 10x v3 single nucleus dataset from Broad (snRNA 10x v3 B) as the reference datasets, as both are large datasets that provide comprehensive cell type coverage and relatively sensitive gene detection.

The key steps of the pipeline are outlined below:

1 **Perform single-dataset clustering** (Methods described above).

2 **Select anchor cells for each reference dataset.** For each reference dataset (scRNA 10x v2 A or snRNA 10x v3 B), we randomly sampled up to $max(100, \frac{5000}{\#clusters})$ anchor cells per cluster to normalize coverage for each cell type. This is the only step that uses the dataset-specific clustering information.

3 **Select highly variable genes (HVGs).** Highly variable gene selection and dimensionality reduction by principal component analysis (PCA) were performed using the *scrattch.hicat* package. We removed PCs with a Pearson correlation coefficient of more than 0.7 with log2(Ngenes). This step was implemented to mitigate the effect of cell/nucleus quality on gene expression variability, and to select only biologically relevant PCs. For each remaining PC, Z-scores were calculated for gene loadings. The top 100 genes with absolute Z-score greater than 2 were selected as HVGs. The HVGs from each reference dataset were combined.

4 **Compute K nearest neighbors (KNN).** For each cell in each query dataset, we computed its K nearest neighbors (k=15) among anchor cells in each reference dataset (scRNA 10x v2 A or snRNA 10x v3 B), based on the highly variable genes selected above. The RANN package was used to compute KNN based on the Euclidean distance when the query and reference dataset is the same. To compute nearest neighbors across datasets, we used correlation as a similarity metric.

5 **Compute the Jaccard similarity**. For every pair of cells from all datasets, we compute their Jaccard similarity, defined as the ratio of the number of shared K nearest neighbors (among all anchors cells from all the reference datasets) divided by the number of combined K nearest neighbors.

6 **Perform Louvain clustering.**

7 **Merge clusters.** To ensure that every pair of clusters are separable by conserved differentially expressed (DE) genes across all datasets, for each cluster, we first

identified the top 3 most similar clusters. For each pair of such closely-related clusters, we computed the differentially expressed genes in each dataset. We focus on the conserved DE genes that are significant in at least one dataset, while also having more than two-fold change in the same direction in all but one datasets. We then compute the overall statistical significance based on such conserved DE genes for each dataset independently. If any of the datasets pass our DE gene criteria described in the "clustering" section, the pair of clusters remain separated; otherwise they are merged. DE genes were recomputed for the merged clusters, and the process was repeated until all clusters are separable by the conserved DE genes criteria. If one cluster has fewer than the minimal number of cells in a dataset (4 cells for SMART-Seq and 10 cells for 10x), then this dataset is not used for DE gene computation for all pairs involving the given cluster. This step allows detection of unique clusters absent in some platforms.

8 **Iterative clustering.** Repeat step 1-6 for cells within each cluster to gain finer resolution clusters until no more clusters can be found.

9 **Final compilation and merging of clusters.** Concatenate all the clusters from all the iterative clustering steps, and perform final merging as described in step **6.**

Marker gene selection. For each pair of clusters, we computed the conserved DE genes, i.e. those which are significantly DE in one at least dataset, with ≥2-fold change in expression in the same direction among 70% of datasets. To allow computation of DE genes involving cell types only present in a subset of datasets, only the datasets with enough cells (based on min.cells parameter) for both cell types under comparison were used for DE gene calculation. We selected the top 50 genes in each direction. After pooling genes from all pairwise comparisons, we identified a total of 3,792 marker genes (Table S6).

Imputation. To facilitate direct comparison, we projected gene expression of all datasets to the space of a given reference dataset. To do that, we leveraged the KNN matrices computed

during the iterative joint clustering step to adjust the expression values for systematic differences between datasets. During each iteration of the joint clustering, for cells in each dataset, we used the average gene expression of their k nearest neighbors among the anchor cells from the reference dataset as the adjusted expression in the reference space. At the top-level clustering, we imputed the expression for all genes. For each subsequent iteration, we only imputed the expression of the high-variance genes and the conserved DE genes for the clusters defined in that iteration. We used this iterative approach for imputation because the nearest neighbors based on the genes chosen at the top level may not reflect the distinction between the finer types, and the imputed values for the DE genes that define the finer types consequently are not accurate based on these nearest neighbors. Therefore, we deferred imputation of the DE genes between the finer types to the iteration when these types were defined. This method is provided in the **impute_knn_global** function in *scrattch.hicat* package[4]. We imputed the gene expression matrix for both reference datasets used in the integrative clustering.

Building a cell-type taxonomy tree. We first computed the average adjusted expression of marker genes for each cluster. This average was computed using each of the two reference datasets (scRNA 10x v2 A, snRNA 10x v3 B). Then, the two matrices were concatenated. We constructed a hierarchy (tree) using the **build_dend_harmonize** function in *scrattch.hicat* package[4].

Dimensionality reduction by UMAP. We performed principal component analysis (PCA) based on imputed gene expression matrices of 3,792 marker genes using 10x single-nucleus dataset from Broad as the reference, and selected the top 50 principal components (93% variance explained). We removed PCs with Pearson correlation coefficient >0.6 with the log2(Ngenes) to reduce bias related to the number of detected genes. Uniform Manifold Approximation and Projection (UMAP) was used to embed the cells in two dimensions with parameters nn.neighbors=25 and md=0.3[33].

1.5.7. MetaNeighbor analysis

To quantify replicability of clusters across the 7 transcriptomic datasets, we applied a modified version of unsupervised MetaNeighbor (RRID:SCR_016727)[42]. MetaNeighbor uses a neighbor voting algorithm and a cross-dataset validation scheme to quantify cluster similarity across multiple datasets. It requires a set of unnormalized datasets, a set of cluster labels and a set of highly variable genes. We used the raw count data for all cells passing QC criteria for the 7 single cell transcriptome datasets, as well as the labels obtained through independent clustering (Table S5). We used MetaNeighbor's *variableGenes* procedure to select 310 highly variable genes that were detected as highly variable across all datasets.

We defined replicable clusters in a two-step procedure: first we quantified the similarity between clusters across datasets, then we extracted groups of highly similar clusters, or "meta-clusters". We used the *MetaNeighborUS* function to obtain an initial similarity matrix between clusters. By default, cluster similarity is quantified as a one-vs-all area under the receiver-operator curve (AUROC): given a training cluster (in one dataset), we ask how similar cells from a test cluster (in another dataset) are to training cells, compared to all other cells in the test dataset. To make cluster matching more stringent, we transformed the one-vs-all AUROC matrix into a one-vs-best AUROC matrix: instead of ranking test cells among all cells from the test dataset, we only compare them to cells from the best matching cluster. This modification ensures that only the best match can have an AUROC > 0.5, facilitating identification of reciprocal best hits. For interpretability and computational efficiency, we adopted the following convention: the best matching cluster's AUROC was obtained by comparing it to the second best matching cluster, the second best cluster's AUROC was obtained by computing 1-AUROC of the best matching cluster, and all other clusters obtained an AUROC of 0, as we were only interested in finding best matches. To extract meta-clusters, we interpreted the one-vs-best AUROC as a graph where nodes are clusters and edges connect nodes if they are

reciprocal best hits. We define meta-clusters as connected components in this graph. We can obtain more robust meta-clusters by requiring that best hits exceed some AUROC threshold. In practice, we noted that one-vs-best AUROC > 0.7 offered a good balance between the number of meta-clusters and reproducibility strength.

For scalability, we modified MetaNeighbor in the following ways. In the *MetaNeighborUS* function, we removed the rank standardization of the cell-cell similarity network (by setting parameter *fast_version* to *TRUE*) and the node degree normalization of the neighbor voting, enabling analytical simplifications of the neighbor voting procedure. The *variableGenes* procedure was applied to a random subset of 50,000 cells for datasets exceeding that size. MetaNeighbor analysis further allowed us to examine the consistency of computational clustering procedures (Figure 2h). We ran three widely used single-cell analysis packages[20,43,44] to generate a fine-grained clustering of each dataset. These cluster analyses were not optimized or manually curated; instead, we used “off-the-shelf” computational procedures to test the robustness of the results from a relatively straightforward and automated analysis. These clusters are thus expected to be less biologically meaningful and robust compared with more customized procedures, such as our reference clustering that incorporates analysis of differential expression to validate the biological reality of cell types. Using the three off-the-shelf cluster analyses, we created a sequence of increasingly coarse-grained clusterings by iteratively merging pairs of clusters chosen to maximize the consistency across computational methods (ARI-merging; see Methods). Finally, at each level of resolution we used MetaNeighbor to calculate the number of clusters which were highly replicable (AUROC>0.7) across datasets. The result of this analysis showed that fine partitions of the data with >30-50 clusters have limited replicability.

1.5.8. Cluster analysis for snmC-Seq

We concatenate principal components from both methylation types (CG and CH) together, and use these to construct a KNN graph followed by Leiden community detection[70]. We repeat the cluster analysis several times to get consensus clustering results. The stopping criteria of clustering considered number of marker genes, accuracy of the reproducible supervised model based on the cluster assignments, and minimum cluster size. We performed the clustering in two iterations to get major types and fine-grained cell types for comparison with other modalities in further integration.

Two-dimensional embedding using t-distributed stochastic neighbor embedding[73] (tSNE; perplexity = 30) was calculated based on the top principal components using the implementation from the scanpy package[74].

1.5.9. Cluster analysis for snATAC-seq

We used the snapATAC pipeline[61] to identify cell clusters with binarized cell-by-bin matrix in 5kb resolution as the input. Cell clusters were annotated to cell type by checking chromatin accessibility along the body of marker genes. Then another round of clustering was performed on MGE- and CGE-derived inhibitory GABA-ergic interneurons, in order to identify sub-cell types.

1.5.10. Multimodality integration

Computational data integration with LIGER We used LIGER (RRID:SCR_018100) to integrate the single-cell transcriptomic and epigenomic data as previously described in the LIGER paper[21], with one modification. We used the *optimizeALS* function in the LIGER package to perform joint factorization on all datasets except methylation (7 RNA datasets and one ATAC dataset) to infer shared (W) and dataset-specific (V_i) metagene factors and cell factor loadings (H_i). We then used the resulting W to calculate cell factor loadings (H_i) for the methylation data using the *solveNNLS* function in the LIGER package. We found that this strategy yielded better

integration than jointly factorizing all 8 datasets, possibly because the inverse relationship and massive dataset size imbalance between methylation and all other datasets complicates the learning of shared metagenes. Our analysis used only the cells annotated by each data-generating group as passing quality control. We did not perform any data imputation or smoothing, but simply normalized and scaled the raw cell-by-gene count matrices from each dataset using the *normalize* and *scaleNotCenter* functions in the LIGER package. We next used the *quantileAlignSNF* function with default settings to perform quantile normalization of cell factor matrices (H_i) from all 8 datasets. Finally, we performed Louvain clustering on the normalized cell factor matrices (H_i) to obtain joint clusters. We performed two rounds of integration and joint clustering; in the first round, we separately integrated all neurons across datasets and all glia across datasets. We then performed a second round of integration and clustering separately for each of four neuronal subclasses: excitatory intratelencephalic (IT) neurons, excitatory non-IT neurons, medial ganglionic eminence (MGE) interneurons, and caudal ganglionic eminence (CGE) interneurons. We used k=40 factors for the non-neuron analysis, k=30 for the first-round neuron analysis, and k=20 for all of the second-round analyses.

Computational integration with SingleCellFusion SingleCellFusion[11] is designed to robustly integrate DNA methylation, ATAC-Seq and/or RNA-Seq data. We applied SingleCellFusion iteratively to integrate all neurons from 8 datasets (Table S1) and jointly call cell clusters. To integrate both the broad and fine-grained cell types, we performed 3 rounds of integration. For every cell cluster generated in the previous round, it is further split into smaller clusters by re-applying SCF on cells in that cluster only. In the first round, we run SCF on all neurons from 8 datasets and get 10 broad neuronal clusters. Rounds 2 and 3 generate 29 clusters and 56 more fine-grained clusters, respectively (Table S3).

The procedure comprises 4 major steps: preprocessing: within-modality smoothing, cross-modality imputation, and clustering and visualization.

1. **Preprocessing.** We define a gene-by-cell feature matrix for each dataset. Droplet-based RNA-seq features (10x) are $\log_{10}$(CPM+1) normalized; Full-length RNA-seq (SMART-seq) features are $\log_{10}$(TPM+1) normalized. snATAC-seq data is represented by read counts within gene body, normalized by $\log_{10}$(RPM+1), where CPM stands for counts per million reads mapped (counts normalized), TPM stands for transcripts per million reads mapped (length normalized), and RPM stands for reads per million reads mapped (length normalized), respectively. DNA methylation data is represented by the mean gene body mCH level, normalized by the global (genome-wide) mean mCH level for each cell. For each dataset, we only used high-quality cells (passed QC) and highly variable genes (n=4,000~6,300) for further analysis. To select highly variable genes, for RNA-seq and ATAC-seq datasets, we first remove genes that are expressed in < 1% of cells. We then divide the remaining genes into 10 bins according to their mean expression across cells (CPM). For each bin, except for the one with the most expressions, we select top 30% of genes with the most expression dispersion (variance/mean) as the highly variable genes. For the DNA methylation dataset, we first select genes that have > 20 cytosine coverage in more than 95% of cells, then divide the remaining genes into 10 bins according to their mean normalized mCH level--raw mCH level normalized by the global mCH for each cell. For each bin, we select top 30% of genes with the most variance as the highly variable 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. The procedure is adapted from [75] and described in detail in [11]. Here we exactly followed Ref[11] with [ndim=50, k=30, ka=5] for all datasets, and [p=0.7] for RNA-seq datasets, [p=0.9] for the DNA methylation dataset, and [p=0.1] for the ATAC-seq dataset.
3. **Cross-modality imputation by Restricted k-Partners (RKP)**. To integrate all 8 datasets, we impute the scRNA_10x_v2_A gene features for cells in all 7 other datasets.

The imputation is done in pairwise between the scRNA_10x_v2_A dataset and one other dataset. For each pairwise imputation, we followed the procedure described in [11] with 20 RKP and relaxation parameter 3 [k=20, z=3]. Instead of using Euclidean distance in a low-dimensional space, we here use the (flipped) spearman correlation coefficient across genes that are highly variable in both datasets as the distance metric between cells in 2 different modalities.

4. **Clustering and visualization.** We start from a cell-by-feature matrix, where cells include all cells from 8 datasets and features are highly variable genes of the scRNA_10x_v2_A dataset. We reduce the dimensionality of features into top 50 Principal Components. Next, we perform UMAP embedding on the PC matrix [n_neighbors=60, min_dist=0.5]. Finally, we perform Leiden clustering on the kNN graph (symmetrized, unweighted) generated from the final PC matrix [Euclidean distance, k=30, resolution=0.1].

Figure specific analysis (Figure S7e) We created the embedding of the cluster centroids using the imputed scRNA_10x_v2_A gene features (log10(CPM+1)) for all cells from the 8 different datasets generated from SingleCellFusion integration. Clusters are defined by individual dataset clusterings and by the joint clustering with SingleCellFusion. Cluster centroids are calculated by the mean imputed scRNA_10x_v2_A gene profiles across cells. After getting a gene-by-cluster matrix, we apply PCA to reduce to 50 feature dimensions, followed by applying a UMAP embedding with min_dist=0.7 and n_neighbors=10.

Figure specific analysis (Figure 3e) To compare molecular signals across data modalities, all signals are normalized to [0, 1]. This is achieved by first getting molecular signals by dataset-specific normalization (Step1), followed by a linear transformation (Step2). In Step1, for SMART-seq datasets, we show log10(TPM+1); for 10x RNA-seq datasets, we show log10(CPM+1); for the ATAC-seq dataset, we show log10(RPM+1) normalized gene body counts, and for DNA methylation we show gene body mCH normalized by global mCH level of

each cell. For Step2, we apply a linear transformation to map the range of the signal to [0, 1]. For datasets other than DNA methylation, we apply the following formula:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where x is the dataset-specific gene-level signal for a cell, x_{min} and x_{max} are defined as the bottom 2 percentile and top 2 percentile of x across all cells, respectively. For the DNA methylation dataset, we apply the following formula:

$$x_{normalized} = 1 - \frac{x - x_{min}}{x_{max} - x_{min}}$$

, with which signals are still mapped to [0, 1] but flipped--a high signal on the plot means a low DNA methylation level. We do this to align DNA methylation signals with gene expression (and open chromatin) signals, because DNA methylation is a repressive marker of gene expression and negatively correlates with it. Besides, x_{min} and x_{max} are defined as the bottom 2 percentile and top 50 percentile of x across all cells, respectively.

Figure specific analysis (Figure 3f) For each gene, cell-level signals are normalized the same way as described in Step1 of Figure 3e. Cluster level signals are the mean cell-level signals across cells in clusters. After getting gene-by-cluster matrices this way, for non-DNA methylation datasets, the matrices are further normalized by the maximum of each cluster (column); for DNA methylation datasets, no further normalization is done, for they are already normalized by cell.

Figure specific analysis (Figure S7g,h) The heatmaps show pairwise Spearman correlation coefficients between the centroids of cells from each cell type (SingleCellFusion) and each dataset, using the gene expression levels ($\log_{10}$(CPM+1); measured or imputed by SingleCellFusion) of the scRNA_10x_v2_A dataset as features. Centroid-level profiles are computed as the average of cell-level profiles across cells from the same cell type and the same dataset. The row and column orderings are the same, generated by a hierarchical clustering on

the above defined centroid-level features with average linkage and euclidean distance. 5f shows the correlations between broad-level joint clusterings (10 subclasses; SingleCellFusion L0; Table S8); 5g shows those between fine-level joint clusterings (56 clusters in total; not all are shown; SingleCellFusion L2; Table S8) for four example broad-level subclasses (MGE, CGE, L2/3 IT, L4/5 IT).

Figure specific analysis (Figure S7c; Agreement metric) We calculated dataset agreement metrics as described in the LIGER paper [21]. Briefly, we performed dimensionality reduction using either NMF (for LIGER) or PCA (for SingleCellFusion) and built a k-nearest neighbor graph for each individual dataset. Then we built a *k*-nearest neighbor graph using the joint latent space from either LIGER or SingleCellFusion and calculated what fraction of the nearest neighbors from individual datasets were still nearest neighbors in the joint space. This metric assesses how well the joint latent space preserves the structure of each individual dataset. An agreement metric close to zero indicates poor preservation of structure from individual datasets, while an agreement metric close to 1 ideally preserves the structure.

Figure specific analysis (Figure S7d; Alignment metric) We calculated dataset alignment metrics as described in the LIGER [21] and Seurat [71] papers, except that we first downsampled cells so that the cluster proportions and total number of cells were identical across all datasets. Then we built a *k*-nearest neighbor graph using the joint latent space from either LIGER or SingleCellFusion and calculated what fraction of the nearest neighbors around each point come from each dataset. We then normalized the metric to be between 0 (no alignment) and 1 (perfect mixing of datasets). This metric assesses how well the joint latent space aligns the datasets. Note that maximizing alignment and maximizing agreement are competing objectives. For example, it is possible to trivially maximize alignment by randomly mixing cells from all datasets according to a spherical Gaussian distribution; conversely, one could trivially maximize agreement by simply assigning non-overlapping latent representations

to all datasets. However, methods must balance these competing objectives to score highly on both alignment and agreement metrics.

Figure specific analysis (Figure S7f) To get cluster-level gene signals, we first get normalized cell-level signals the same way as Step 1 of Figure 3e, followed by taking the mean cell-level signals across cells in clusters.

1.5.11. Analysis of enhancers

Epigenome Cluster Level. Based on the cell-cell integration in Figure 7, in order to have enough whole-genome coverage of each cell type, we further merged the co-clusters into a higher level to increase the coverage of each cluster, which we termed as the epigenome cluster level.

Differentially methylated region (DMR) calling. For DMR calling in the snmC-seq2 data, we merged single-cell ALLC files into the pseudo-bulk level for each cluster, and then used methylpy[76] DMRfind function to calculate mCG DMRs across all clusters. The base call of each paired CpG sites was added up 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, and then merge the DMS within 250bp into DMR. Hypo-DMR and Hyper-DMR were then assigned to each sample by examining the residue of observed counts from the expected counts. We also filtered the DMRs by requiring the maximin difference of mCG rate between clusters larger than 0.3.

snATAC peak calling. We called peaks according to the ENCODE ATAC-seq pipeline (https://www.encodeproject.org/atac-seq/). For every cell cluster, we combined all properly paired reads to generate a pseudobulk ATAC-seq dataset for individual biological replicates. In addition, we generated two pseudo-replicates, each of which includes half of the reads from each biological replicate. We called peaks independently for each of these four datasets, as well as for a pool of the data from both biological replicates. Peak calling was performed on the

Tn5-corrected single-base insertions using MACS2[77] (RRID:SCR_013291) with parameters: --shift -75 --extsize 150 --nomodel --call-summits --SPMR --keep-dup all -q 0.01. We extended peak summits by 250 bp on either side to a final width of 501 bp for merging and downstream analysis. To generate a list of reproducible peaks, we kept peaks that 1) were detected in the pooled dataset and overlapped ≥50% of the peak length with a peak in both individual biological replicates or 2) were detected in the pooled dataset and overlapped ≥50% of peak length with a peak in both pseudo-replicates.

To account for differences in performance of MACS2 based on read depth and/or number of nuclei in individual clusters, we converted MACS2 peak scores (-log10(q-value)) to score per million (SPM)[78] and kept peaks with SPM>2. We only kept reproducible peaks on chromosome 1-19 and both sex chromosomes, and filtered ENCODE mm10 blacklist regions[79,80] http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/mm10.blacklist.bed.gz. Finally, since snATAC-seq data are relatively sparse, we selected only elements that were identified as open chromatin in a significant fraction of the cells in each cluster. To this end, we defined a set of background regions, matching the number of peak regions for each cell type, by randomly selecting regions from the genome while excluding accessible sites from the ENCODE registry of cis-regulatory elements (https://screen.encodeproject.org/). We calculated the fraction of nuclei for each cell type that had ATAC fragments mapping to the background regions. Next, we fitted a zero-inflated beta model and empirically identified a significance threshold of FDR < 0.01 to filter potential false positive peaks. Peak regions with FDR < 0.01 in at least one of the clusters were included into downstream analysis.

We used the *bedtools intersect* with the "-wa -u" parameter to calculate DMR and ATAC peak overlaps[81] (RRID:SCR_006646).

Saturation analysis. To investigate the efficiency of regulatory elements identification in terms of cell number in the epigenomic data, we did a saturation analysis using the two most

abundant cell types: the L2/3 IT and the L6 CT excitatory neurons, the total reads assigned to these two cell types were comparable to bulk-seq. We subsampled a different number of cells without replacement in each cluster three times when having enough cells, and used cells from each replicate separately when possible. In the last group, we used all the cells for each cell type as a maximum reference. For methylome data, We call DMRs between L2/3 IT and L6 CT within each cell number group. Peaks are called for each cell type group.

REPTILE enhancer prediction. We performed enhancer prediction using the REPTILE[82] algorithm. The REPTILE is a random-forest-based supervised method that incorporates different sources of 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 previous studies[82,83] using CG methylation, chromatin accessibility of each epigenome cluster and mouse embryonic stem cells (mESC). The model was first trained on mESC data and then predicted a quantitative score we termed enhancer score for each cell type's DMRs. The positives were 2kb regions centered at the summits of top 5,000 EP300 peaks in mESCs. Negatives include randomly chosen 5,000 promoters and 30,000 2kb genomic bins. The bins have no overlap with any positives or promoters[83]. Methylation and chromatin accessibility profiles in bigwig format for mESC were from the mouse ENCODE project[83]. The mCG rate bigwig file was generated from cell type-merged ALLC files using in-house python script. For chromatin accessibility of each cell type, we merged all fragments from snATAC-seq cells that were assigned to this cell type in the integration analysis and used "*deeptools bamcoverage*" to generate CPM normalized bigwig files. All bigwig files' bin size was 50bp.

Motif Enrichment Analysis. We used 724 motif PWMs from the JASPAR 2020 CORE vertebrates database[84], where each motif was able to assign corresponding mouse transcription factor genes. For each set of REPTILE predicted enhancers, we standardized the region length into center ± 250bp and used the FIMO tool from the MEME suite[85] to scan the motifs in each

enhancer with log odds p-value < 10^{-6} as the threshold of motif hit. To calculate motif enrichment, we use the adult non-neuronal mouse tissue DMRs[86] as background regions. 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. Transcription factors with significant motif enrichment were grouped by TFClass[87] classification. Genes within the same group share very similar motifs.

1.5.12. Cluster validation analysis

Downsampling analysis of cluster number (Figure 5a-d) We first preprocessed raw count matrices in the same way as described in the section of Computational integration with SingleCellFusion. After preprocessing, we get a gene-by-cell feature matrix for each dataset. Only neuronal cells passing QC (Table S1) and highly variable genes for each dataset are included. Then we do clustering, which takes 3 steps. We first reduce feature dimensions by PCA [n=50]. We then build a k-nearest neighbor graph [k=30] between cells using the Euclidean distance in the Principal Component space. We finally apply the Leiden clustering algorithm with a fixed resolution parameter [r=6]. For each dataset, we report the number of clusters as a function of the number of cells randomly downsampled from the full dataset. Error bars show the standard error of the mean of [n=10] repeats of downsampling.

Clustering with within-modality cross validation (Figure 5c) This analysis aims to estimate the "optimal" number of clusters of a dataset, by testing which clustering granularity best preserves the gene-level features of cells. For a given dataset--a gene-by-cell matrix, we first randomly split gene features into 2 sets, for clustering and validation, respectively. To avoid any potential linkage, the split is done by separating chromosomes into 2 sets, such that genes from the same chromosomes are always in the same set. We then perform Leiden clustering (as described in methods related to Figure 5a) on all cells using the clustering feature set only with

different clustering resolutions. After clustering, every cell in the dataset gets a cluster label. We next randomly separate those cells into 2 sets--for training and testing, respectively. Using training-set cells, we train a supervised model to predict the validation set gene features based on cluster assignments. The model is trained by minimizing the mean squared error between model prediction and data. This is equivalent to predicting a cell's gene features as its cluster centroid. Assuming a R2 loss, this is equivalent to calculating the cluster centroid of each cluster in the space of validation gene set using training-set cells only. Finally, we evaluate the model performance by calculating apply the model to cells in the test set, and evaluate the mean squared error for the test-set cells. of model performance. This is equivalent to estimating the mean squared distance between individual cells in the test set to its cluster centroid calculated by cells in the training set. As a function of number of clusters (by varying the resolution parameter in Leiden clustering), we observe a U-shaped curve of mean squared error, because both under-splitting and over-splitting results in high mean squared error. The minimum point of the curve represents the most plausible clustering resolution. Applying this scheme to each dataset and different downsampling levels of cells, we report the number of clusters as a function of the number of cells, for each dataset. For robustness, random split of gene features are repeated n=5 times; random split of cells are repeated n=5 times with k=5 fold cross validations each time.

Clustering with cross-modality cross validation (Figure 5d) Extending the within-modality clustering cross validation scheme used in Figure 5c, we developed a cross-modality cross validation method, by combining the previously described within-dataset cross-validation method with a joint clustering method--SingleCellFusion. First of all, similar to within-dataset cross validation, we first randomly split gene features into clustering and validation set for all datasets. We then generate integrated clusterings across data modalities by applying SingleCellFusion on all cells and half of the gene features (the clustering feature set). After clustering, we estimate the mean squared error of clustering on the validation feature set

as described above for each dataset on its own. Applying this scheme to different downsampling levels of cells, we report in Figure 5d the number of clusters as a function of the number of cells from each dataset.

1.5.13. Integrated analyses: trade-off between replicability and resolution and cluster consistency

We collected the clusters obtained with the 4 integrative clustering methods described previously (Conos, LIGER, RNA consensus clustering from Figure 2, SingleCellFusion), as well as the "subclass" level from the independent clustering of the RNA datasets. Each integrative method returned clusters at two granularity levels. We named the coarser level of clustering L1 and the finer level of clustering L2 clusters. We focused our analyses on the neuron clusters of the transcriptomic data, as we wished to investigate the agreement of neuron cluster hierarchies.

To quantify replicability, we used the same modified version of MetaNeighbor, same datasets and same variable genes as defined above (see "MetaNeighbor analysis"). We used the one-vs-best AUROC to obtain cluster similarity scores, then computed an average AUROC score per integrated cluster (averaged over every pair of datasets in which the cluster is present). For every method, we reported the median AUROC across integrated clusters as the final reproducibility score. To quantify the overall similarity of the clustering results, we computed the Adjusted Rand Index (ARI). When necessary, we restricted the ARI computation to the intersection of labeled cells (the intersection being recomputed for every pair of methods).

1.5.14. Clustering on network of samples (CONOS) analysis

To evaluate the extent to which different cell subpopulations were supported by different platforms, we assessed the difference in the ability to recover the corresponding cell with and without within-platform comparisons. The clustering of cells was performed using Conos[52], using

walktrap community detection to identify hierarchical cell populations. The stability of the hierarchical clusters was estimated as follows: 20 random cell subsampling rounds were performed, each sampling 95% of cells from each dataset, and repeating the walktrap hierarchical clustering procedure. For each node in the original walktrap tree, we evaluated stability as a minimum of specificity and sensitivity relative to the ensemble of subsampled trees by finding the best matching subtree. To evaluate the ability to recover subpopulations based on cross-platform comparisons only, we removed within-platform edges (those connecting datasets generated by the same platform) in the joint graph (generated by Conos). In this way, the subpopulation was detected only based on mapping to the other platform. The modified approach facilitates grouping of cell populations that are common in the different platforms, as it removes the platform-specific information in the joint graph.

To assess similarity of expression profiles detected by different platforms for a given cell type (Figure 5h), we used Jensen-Shannon divergence to assess the overall similarity of gene expression patterns between the four RNA-Seq platforms (scRNA 10x v3 A, snRNA 10x v3 A, scRNA SMART and snRNA SMART). Specifically, 1000 cells were sampled from each cell type for each platform. If the number of cells from a cell type is smaller than 1000 cells, sampling with replacement was performed. Cell types that accounted for less than 1% (<300 cells) in any specific platform were omitted. The molecules detected for each gene were then aggregated across all sampled cells for each cell type in each platform. The counts were normalized by the total number of molecules for each cell type / platform, and Jensen-Shannon divergence was calculated.

1.6. Acknowledgments

Chapter 1, in full, is a reprint of the material as it appears in *Nature* 2021. Z. Yao, H. Liu, **F. Xie**, S. Fischer, R. S. Adkins, A. I. Aldridge, S. A. Ament, A. Bartlett, M. M. Behrens, K. Van den Berge, D. Bertagnolli, H. R. de Bézieux, T. Biancalani, A. S. Booeshaghi, H. C. Bravo, T.

Casper, C. Colantuoni, J. Crabtree, H. Creasy, K. Crichton, M. Crow, N. Dee, E. L. Dougherty, W. I. Doyle, S. Dudoit, R. Fang, V. Felix, O. Fong, M. Giglio, J. Goldy, M. Hawrylycz, B. R. Herb, R. Hertzano, X. Hou, Q. Hu, J. Kancherla, M. Kroll, K. Lathia, Y. E. Li, J. D. Lucero, C. Luo, A. Mahurkar, D. McMillen, N. M. Nadaf, J. R. Nery, T. N. Nguyen, S.-Y. Niu, V. Ntranos, J. Orvis, J. K. Osteen, T. Pham, A. Pinto-Duarte, O. Poirion, S. Preissl, E. Purdom, C. Rimorin, D. Risso, A. C. Rivkin, K. Smith, K. Street, J. Sulc, V. Svensson, M. Tieu, A. Torkelson, H. Tung, E. D. Vaishnav, C. R. Vanderburg, C. van Velthoven, X. Wang, O. R. White, Z. J. Huang, P. V. Kharchenko, L. Pachter, J. Ngai, A. Regev, B. Tasic, J. D. Welch, J. Gillis, E. Z. Macosko, B. Ren, J. R. Ecker, H. Zeng, E. A. Mukamel, A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. *Nature*. **598**, 103–110 (2021). The dissertation author was a co-first author of this paper.

These authors contributed to RNA data generation: A.R., A.T., B.T., C.R., C.R.V., D.B., D.M., E.L.D., E.Z.M., H.T., H.Z., J.G., J.S., K.C., K.L., K.S., M.K., M.T., N.D., N.M.N., O.F., T.C., T.N.N., T.P. These authors contributed to DNA methylation (snmC-Seq2) data generation: A.B., A.C.R., A.I.A., A.P-D., C.L., H.L., J.D.L., J.K.O., J.R.E., J.R.N., M.M.B., S.N., Y.E.L. These authors contributed to snATAC data generation: A.P., B.R., J.D.L., J.K.O., M.M.B., S.P., X.H., X.W., Y.E.L. These authors contributed to data archive/infrastructure: A.M., B.R.H., C.C., C.V.V., E.A.M., F.X., H.C., H.C.B., J.C., J.G., J.K., J.O., M.G., M.H., O.R.W., R.F., R.H., R.S.A., S.A.A., S.N., V.F., W.I.D., Z.Y. These authors contributed to data analysis: A.R., A.S.B., B.T., D.R., E.A.M., E.D.V., E.P., E.Z.M., F.X., H.L., H.R.D.B., H.Z., J.D.W., J.G., J.G., J.O., K.S., K.S., K.V.D.B., L.P., M.C., O.F., O.P., P.V.K., Q.H., R.F., S.D., S.F., S.N., T.B., V.N., V.S., W.I.D., Y.E.L., Z.Y. These authors contributed to data interpretation: A.R., B.R., B.T., C.L., E.A.M., E.D.V., E.Z.M., F.X., H.L., H.Z., J.D.W., J.G., J.N., M.C., M.M.B., P.V.K., Q.H., R.F., S.F., T.B., Y.E.L., Z.Y. These authors contributed to writing manuscript: A.S.B., E.A.M., F.X., H.L., H.Z., J.D.W., J.G., L.P., M.C., Q.H., S.F., Z.J.H., Z.Y.

We are grateful to Anita Bandrowski and Yong Yao for their insightful comments. This work was funded by the NIH BRAIN Initiative (U19MH114830 to H.Z.; U19MH121282 to J.R.E.; U19MH114821 to Z.J.H.; R24MH114788 to O.R.W.; U24MH114827 to M.H.; R24MH114815 to R.H./O.R.W.; NIH NIDCD DC013817 to R.H.), the Hearing Restoration project Hearing Health Foundation (R.H.), and NIH NIGMS (GM114267 to H.C.B.).

Chapter 2. Validation of computational integration using single-cell multiomic sequencing data

2.1. Abstract

Single-cell technologies enable a measure of unique cellular signatures but are typically limited to a single modality. Computational approaches allow integration of diverse single-cell datasets, but their efficacy is difficult to validate in the absence of authentic multi-omic measurements. To comprehensively assess the molecular phenotypes of single cells in tissues, 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 integration methods.

2.2. Introduction

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 [88,89]. 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 [4,27,29,57,90], 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 between modalities. Recently, computational methods based on Canonical Correlation Analysis [20], mutual nearest neighbors [23] or matrix factorization [21] have been developed to integrate molecular data types. However, validating the results of computational integration 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 [88]. 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 [30]. 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 [91–93]. 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 postmortem human frontal cortex tissues. 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 integrate multiple single-cell data types.

2.3. Experimental 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 (Figure 6a). We treated single cells/nuclei with Smart-seq or Smart-seq2 reactions for *in situ* cDNA synthesis and amplification of full-length cDNA [94,95]. 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 [27,68]. 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 [96]. 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 [97]. 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 [57]) were misclassified as methylome reads (Figure S9). 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, Figure 6a) [93,98–100]. 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 (Figure 6a).

2.4. Results

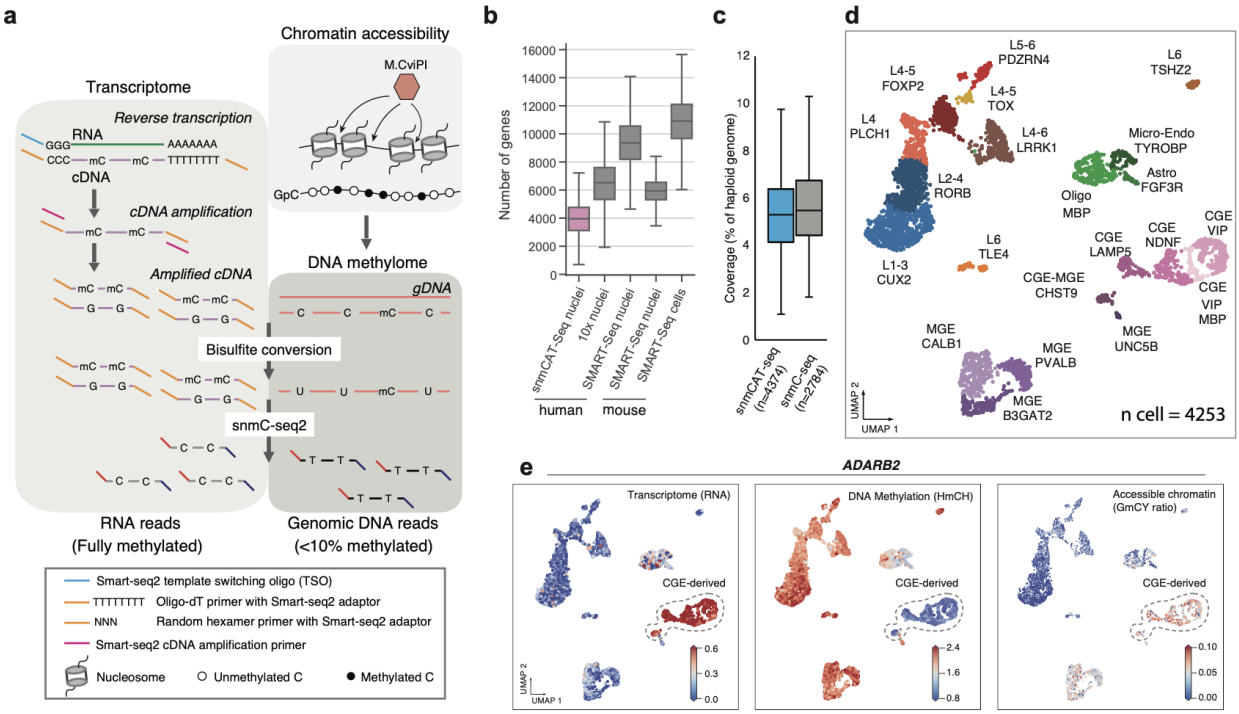
2.4.1. Multi-omic profiling of postmortem human brain tissue with snmCAT-seq

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). 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 (Figure S10a), the fraction of transcriptome reads mapped to introns and exons (Figure S10b) and the number of genes detected (Figure 6b and Figure S10c). Compared with snmC-seq and snmC-seq2 data generated from human single nuclei [27,68], the DNA methylome component of snmCAT-seq had comparable genomic coverage (Figure 6c), mapping efficiency (Figure S10d), and showed only moderately reduced library complexity (Figure S10e) with similar coverage uniformity (Figure S10f-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). We identified highly variable features and calculated principal components separately for each modality, then concatenated the principal components together as the input features for multi-modal clustering and visualization using Uniform Manifold Approximation and Projection

(UMAP) [33] of the three data types. 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 (Figure S10h). These cell types were effectively separated by performing dimensionality reduction using each data type (Figure S10i-k). The comparison of homologous clusters between snmCAT-seq transcriptome and snRNA-seq 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, $p = 3x10^{-301}$) (Figure S10l-m). In summary, snmCAT-seq can simultaneously profile transcriptome and methylome in single nuclei, accurately recapitulating cell-type signatures for each data type.

Figure 6: 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**) UMAP embedding of human frontal cortex snmCAT-seq profiles. (**e**) UMAP embedding of transcriptome, methylome and chromatin accessibility profiled by snmCAT-seq for *ADARB2*. From left to right, the cells are colored by gene expression (CPM, counts per million), non-CG DNA methylation (HmCH ratio normalized per cell) and chromatin accessibility (MAGIC imputed GmCY ratio, see methods).


a
Chromatin accessibility
M.CviPI
GpC
Transcriptome
Reverse transcription
RNA
GGG
AAAAAAA
CCC
mC
TTTTTTTT
cDNA
cDNA amplification
Amplified cDNA
DNA methylome
gDNA
Bisulfite conversion
snmC-seq2
RNA reads
(Fully methylated)
Genomic DNA reads
(<10% methylated)
Smart-seq2 template switching oligo (TSO)
Oligo-dT primer with Smart-seq2 adaptor
NNN
Random hexamer primer with Smart-seq2 adaptor
Smart-seq2 cDNA amplification primer
Nucleosome
Unmethylated C
Methylated C
b
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
L5-6 PDZRN4
L6 TSHZ2
L4-5 FOXP2
L4-5 TOX
L4 PLCH1
L4-6 LRRK1
Micro-Endo TYROBP
Astro FGF3R
L2-4 RORB
Oligo MBP
CGE VIP
CGE NDNF
CGE LAMP5
L6 TLE4
L1-3 CUX2
CGE-MGE CHST9
CGE VIP MBP
MGE UNC5B
MGE CALB1
MGE PVALB
MGE B3GAT2
n cell = 4253
UMAP 2
UMAP 1
e
ADARB2
Transcriptome (RNA)
DNA Methylation (HmCH)
Accessible chromatin (GmCY ratio)
CGE-derived

2.4.2. 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 [30]. Cross-dataset integration can be used to assess cluster robustness, but it may be limited by systematic differences between the datasets or modalities used [42]. 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, Figure 7a). 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 (Figure 7b-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, Figure 7b-c and Figure S11a). 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 (Figure 7b-c). The same approach can also be applied to each individual modality separately, to identify the number of

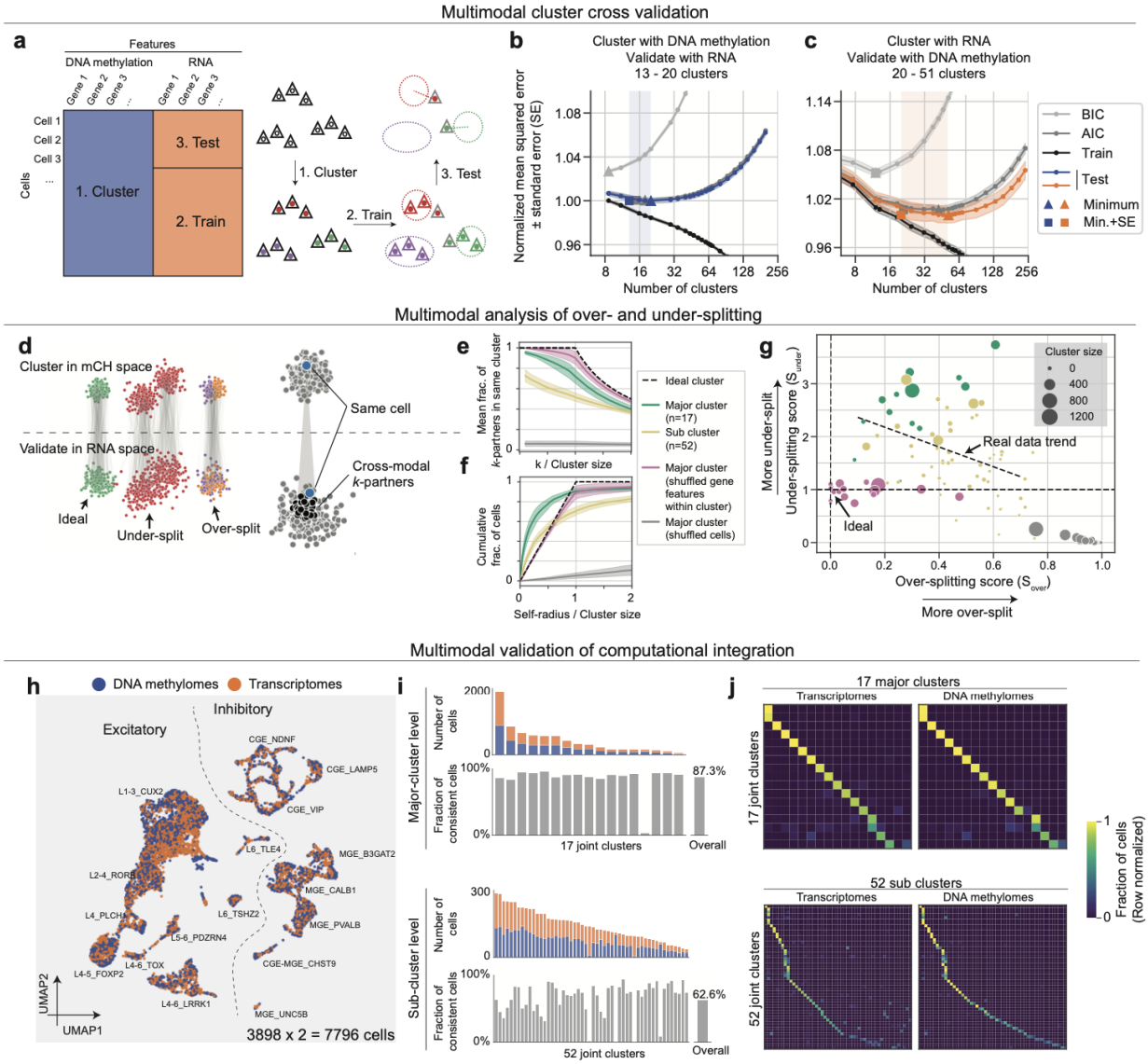
clusters supported by DNA methylation features and by RNA features, respectively (Figure S11a).

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; Figure 7D, Figure S11b). After jointly embedding mC and RNA data in a common low-dimensional space [71], 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; Figure 7D, E). We assessed the over-splitting of 17 major neuronal clusters and 52 neuronal sub-clusters identified by single-cell methylomes and found major clusters resemble ideal, homogeneous clusters (simulated by shuffling gene features) with low over-splitting scores (Figure 7E, Figure S11c, 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 (Figure S11c, 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 (Figure 7F, Figure S11d, 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 (Figure 7G).

Integration of single-cell genomic data has been a focus of recent computational studies, yet existing methods lack validation on ground truth from experimental single-cell multi-omic datasets [18]. By treating snmCAT-seq transcriptome and mC profiles as if they were generated from different single cells, we could test the performance of computational integration using Seurat [20], Harmony [101], Scanorama [23], LIGER [21] and Single Cell Fusion (Methods; Figure 7H, Figure S11g-k, first row). To evaluate the integration at the cell-level, we calculated the self-radius as mentioned above and determined miss integrated events by normalized self-radius > 0.3 (Figure S11g-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 (Figure 7I, Figure S11g-k, third row). Overall, the Single Cell Fusion and Seurat outperform the other tools, with the Single Cell Fusion achieving the lowest miss integrated ratio (5.7%) and highest overall major cell-type level accuracy (87.3%) (Figure 7I-J, Figure S11g). We also tested the Single Cell Fusion accuracy at the subtype level. As expected, computational integration of fine-grain clusters was less accurate (62.6%) and more variable across clusters (Figure 7I), potentially because of the greater degree of over-clustering (Figure 7E).

Figure 7: 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 integration using the SingleCellFusion method, assuming snmCAT-seq transcriptomes and methylome were derived from independent datasets. **i.** Accuracy of computational integration 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. 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
Ideal
Under-split
Over-split
Same cell
Cross-modal k-partners
e
Mean frac. of k-partners in same cluster
k / Cluster size
f
Cumulative frac. of cells
Self-radius / Cluster size
Ideal cluster
Major cluster (n=17)
Sub cluster (n=52)
Major cluster (shuffled gene features within cluster)
Major cluster (shuffled cells)
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)

2.5. 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 [68] and can be readily scaled to analyze thousands of cells/nuclei.

The continuous development of multi-omic profiling techniques such as scNMT-seq [93] and snmCAT-seq, and several methods for joint RNA and chromatin accessibility profiling sci-CAR [46], SNARE-seq [102], Paired-seq [103] and SHARE-seq [104] 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, [27]) and current (20 major types and 63 subtypes) studies. Using snmCAT-seq as a “ground-truth”, we determined that computational multi-modal integration tools perform well at the major cell-type level but show variable accuracy for the integration 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 [105,106].

2.6. Methods

2.6.1. 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.

2.6.2. 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.

2.6.3. 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.

2.6.4. 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.

2.6.5. 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 [107]. 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).

2.6.6. Bisulfite conversion and library preparation

Detailed methods for bisulfite conversion and library preparation are previously described for snmC-seq2 [27,68]. 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 [27]. The rest of the snmCAT-seq libraries were generated using the snmC-seq2 method as described in Luo et al., 2018 [68]. 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.

2.6.7. The mapping pipeline for snmCAT-seq

We implemented a versatile mapping pipeline (cemba-data.rtfd.io) for all the methylome based technologies developed by our group [27,68]. 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 [27,68]. 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 [69] 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 [62] 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) [68], which contains no transcriptomic reads. Similarly, the criteria for identifying snmCAT-seq methylome reads were applied to Smart-seq data (SRR944317, SRR944318, SRR944319, SRR944320) [95], 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 [76] 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 [108] 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”.

2.6.8. 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 integration between methylome and transcriptome.

2.6.9. Preprocessing snmCAT-seq data methylome for clustering analysis

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.

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 [27], 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 [79] 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 $(\alpha,\ \beta)$ 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: $\hat{mc} = \frac{\alpha+mc}{\alpha+\beta+cov}$., We normalized this rate by the cell's global mean methylation, $m = \alpha/(\alpha + \beta)$. Thus, all the posterior $\hat{mc}$ 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 [74]. 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.

2.6.10. Preprocessing snmCAT-seq data transcriptome for clustering analysis

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.

2.6.11. Preprocessing snmCAT-seq data chromatin accessibility (NOMe-seq) for clustering analysis

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 $C_{ij} = log(B_{ij} + 1) \times log(1 + \frac{\#cells}{\sum_{i'=1}^{\#cells} A_{i'j}})$ 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.

2.6.12. General strategies for clustering and manifold learning

Consensus clustering on concatenated PCs

We used a consensus clustering approach based on multiple Leiden-clustering [109] 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) [110] 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* [111] as an evaluation score. The multiclass prediction model is based on BalancedRandomForestClassifier from the imblearn package that accounts for imbalanced classification problems [112]. 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

t-SNE and UMAP embeddings are run on the PC matrix the same as the clustering input using the scanpy package.

2.6.13. snmCAT-seq baseline clustering

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 [27,35]. We also calculated the UMAP coordinates based on concatenated PCs (Figure 6D) and PCs from every single modality separately (Figure S10i-k).

2.6.14. Cross-validation of cell clusters

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.

2.6.15. AIC and BIC metrics in the cluster cross-validation analysis

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].

2.6.16. Quantification of over-splitting and under-splitting of cell clusters

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 [113]. 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 [114]. 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 [71]:

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

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:

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

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 (Figure 7F). 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 Figure 7F). 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 Figure 7F). 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 (Figure 7F). 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.

2.6.17. Computational data integration with SingleCellFusion (Figure 9)

Several computational methods have been proposed for integrating multiple single-cell sequencing datasets across batches, sequencing technologies, and modalities [21–23,71,101]. 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 integrated 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 integration 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 [22,23], while other approaches including canonical correlation analysis (CCA) and non-negative matrix factorization (NMF) have previously been used for integrating transcriptomic and epigenomic data [21,71]. Single Cell Fusion is designed to robustly integrate 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 [75]. 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) [22]. 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 [22] on the PC matrix [n_neighbors=30, min_dist=0.5]. Finally, we perform Leiden clustering (Traag [22] on the kNN graph (symmetrized, unweighted) generated from the final PC matrix [Euclidean distance, k=30, resolution=0.3, 1, 2, 4].

2.6.18. Evaluation of Computational Integration Methods (Figure S11)

We tested five data integration tools: 1) Scanorama [23]; 2) Harmony [101]; 3) Seurat [20]; 4) LIGER [21,115] 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 integration to account for the negative correlation of mCH fraction and RNA expression.

Below we describe the integration 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 integration 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 integration and dimension reduction using Scanorama V1.7 on the scaled (via scanpy.pp.scale) mC and RNA matrix. We used the top 20 integrated PCs (n_components = 20) for integration 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. Harmony integrated PCs were then used for evaluation.

3) For Seurat, we followed the Seurat (v4.0.0) vignette steps to perform integration (https://satijalab.org/seurat/articles/integration_introduction.html). When calculating integration anchors (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 integration and used the normalized matrix H (the cells' decomposed matrix from the online iNMF algorithm) for integration evaluation.

Finally, the SingleCellFusion analysis was described in the manuscript, we used the integrated PCs for evaluation.

2.6.19. Metrics for integration evaluation

We used three different approaches to evaluate the integration results. First, We ran UMAP on the decomposed matrix from each tool to provide an overview of the integrated 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 integration 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 integration and larger values indicating inadequate integration 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 integration, and a low accuracy indicates inadequate integration.

2.7. Acknowledgements

Chapter 2, in part, has been submitted for publication of the material. The preprint of this manuscript is posted on *bioRxiv*. C. Luo, H. Liu, **F. Xie**, E. J. Armand, K. Siletti, T. Bakken, R. Fang, W. I. Doyle, R. D. Hodge, L. Hu, B.-A. Wang, Z. Zhang, S. Preissl, D.-S. Lee, J. Zhou, S.-Y. Niu, R. Castanon, A. Bartlett, A. Rivkin, X. Wang, J. Lucero, J. R. Nery, D. A. Davis, D. C. Mash, J. R. Dixon, S. Linnarsson, E. Lein, M. Margarita Behrens, B. Ren, E. A. Mukamel, J. R.

Ecker, Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv* (2019), p. 2019.12.11.873398. The dissertation author was a co-first author of this paper.

This work was supported by NIH grants: 5R21HG009274, 5R21MH112161 and 5U19MH114831 to J.R.E; R01HG010634 to J.R.E and J.E.D; U01MH114812 to E.L. J.R.E. is an Investigator of the Howard Hughes Medical Institute. W.D. 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. We thank the tissue donors and their families for their invaluable contributions to the advancement of science. Wethank the QB3 Macrolab at UC Berkeley for purification of Tn5 transposase. Work at the Center for Epigenomics was supported in part by the UC San Diego School of Medicine.

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.

Chapter 3. Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes

3.1. Abstract

Integrating single-cell transcriptomes and epigenomes across diverse cell types can link genes with the *cis*-regulatory elements (CREs) that control expression. Gene co-expression across cell types confounds simple correlation-based analysis and results in high false prediction rates. We developed a procedure that controls for co-expression between genes and integrates multiple molecular modalities, and used it to identify >10,000 gene-CRE pairs that contribute to gene expression programs in different cell types in the mouse brain.

3.2. Introduction

Single-cell epigenome sequencing techniques, including snATAC-seq and snmC-seq, can identify cell-type-specific candidate *cis*-regulatory elements (cCREs), such as enhancers[9,15]. To validate putative enhancers and elucidate their function, it is important to identify the genes they directly regulate[116]. This can be accomplished by simultaneously perturbing enhancer activity and measuring gene expression in the same cells[117,118]. However, perturbation experiments are complex and to date have been used to screen pre-selected enhancers in cell types that could be cultured *in vitro*[117,118]. By contrast, single-cell transcriptomes and epigenomes from complex tissues, such as the brain, contain distinct genome-wide profiles from several hundred cell types[7,10]. Correlating enhancer epigenetic profiles with transcription across cell types can identify potential cell-type-specific enhancer-gene links[9,15,119]. However, genes with related functions often have correlated expression patterns, leading to incidental associations that could confound co-expression analyses with false-positives that do not reflect genuine enhancer-target gene interactions[9,15,103,119,120].

To separate spurious from genuine associations, *trans* enhancer-gene correlations can be used as a negative control[78,104,121–123]. However, a principled analysis and validation of the most appropriate null model has not been performed. Moreover, different epigenetic assays, such as snATAC-seq and snmC-seq, measure distinct aspects of enhancer activity. It is unclear how the differences between these data modalities affect the sensitivity and specificity for detecting enhancer-gene correlations. Furthermore, correlation results may be strongly influenced by clustering analysis of single cell data, which in turn depends on multiple unconstrained parameters and algorithmic choices.

To address these gaps, we identify high-confidence, robust enhancer-gene links using a non-parametric permutation-based procedure to control for gene co-expression (Figure 8a, Figure S12a). We first integrate single-cell transcriptomes (scRNA-seq) and epigenomes (open chromatin, snATAC-seq, and DNA methylation, snmC-Seq) to generate multi-modality profiles using a dataset with over 200,000 single cells from the mouse primary motor cortex[10]. We correlate the epigenetic state of putative enhancers with expression of nearby genes, and compare the observed correlation with two null distributions. A conventional shuffling procedure that randomly permutes cell labels effectively controls for noise present in single-cell sequencing measurements[9,15,103]. However, as we discuss below, this null distribution is confounded by gene co-expression and leads to spurious enhancer-gene associations. This challenge can be addressed statistically using generalized least squares regression[124] (GLS), which transforms data matrices to decorrelate observations. We used a more general non-parametric approach, shuffling genomic regions to create an appropriate null distribution[78,104,121–123]. Moreover, we leveraged three complementary data modalities to cross-validate enhancer-gene links with independent data. Finally, we validated the predicted links with multimodal 3D chromatin conformation (snm3C-seq) data[125].

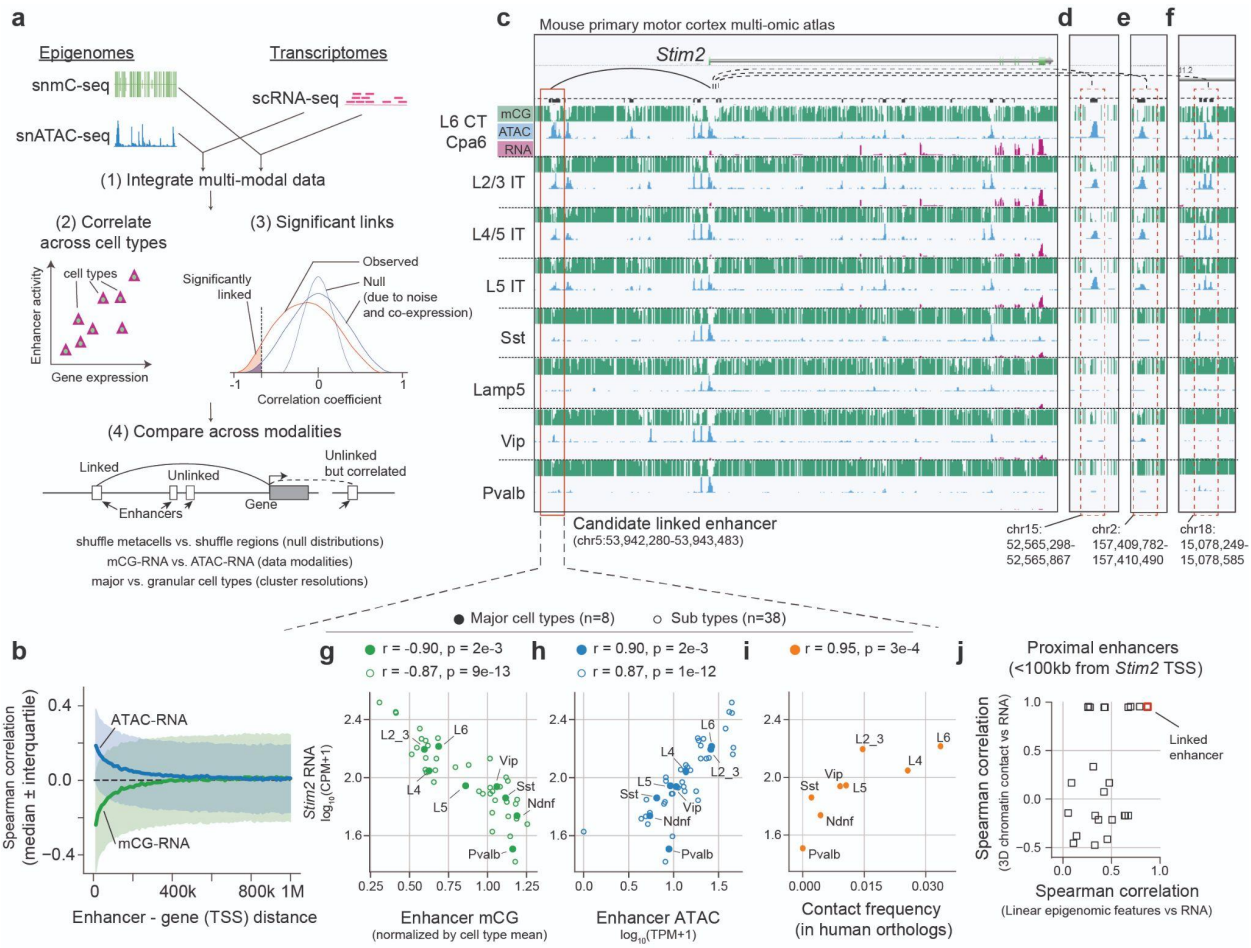


Figure 8: Identifying enhancer-gene links through integrated analysis of single-cell transcriptomes and epigenomes. a. Our proposed method links enhancers with target genes by (1) integrating single-cell transcriptomes (scRNA-seq) and epigenomes (snmC-seq and snATAC-seq), (2) correlating enhancer activity with gene expression across metacells, (3) identifying significant links compared with a shuffled null distribution, and (4) evaluating predicted links across null models, data modalities, and metacell resolutions. **b.** Strength of enhancer-gene association as a function of genomic distance. The wide interquartile range (shading) indicates high variability in enhancer-gene associations. **c-f.** Correlation of the gene *Stim2* with nearby (**c**) and distal (**d-f**) enhancer regions. **g-i.** Scatter plots of *Stim2* expression versus enhancer mCG (**g**), ATAC-seq signal (**h**), and enhancer-TSS chromatin contact frequency in human orthologs (**i**). **j.** Enhancer-gene association from linear-genome features (mCG, ATAC) versus 3D-genome features (chromatin contact frequency) for *Stim2* proximal enhancers. The x-axis shows the minimum absolute correlation value between mCG-RNA and ATAC-RNA. Enhancer mCG level is normalized by the global mean mCG level of each cell type; RNA is $\log_{10}$(CPM+1) normalized; ATAC is $\log_{10}$(TPM+1) normalized.

3.3. Results

To illustrate the risk of false associations due to gene co-expression, we analyzed a large set of single-cell transcriptome and epigenome data from the mouse primary motor cortex[10]. Putative enhancers (see Methods; Table S10, Figure S12b) within ~100 kb of a gene promoter were enriched in associations with gene expression, including positive correlations for chromatin accessibility and negative correlations for enhancer DNA methylation (mCG) (Figure 8b, Figure S12c,d). However, these associations were highly variable: We observed many weak correlations for proximal enhancers (<100 kb), and relatively strong correlations for some distal enhancers (>500kb) (Figure 8b, interquartile range ~0.4). The broad distribution of correlation strength makes it difficult to reliably link specific enhancers with their target genes.

A representative example is the gene *Stim2,* encoding a calcium sensor that helps maintain basal Ca^{2+} levels in pyramidal neurons[126]. In cortical neurons, we identified 33 enhancers within 100 kb of the *Stim2* promoter. *Stim2* expression correlates with low mCG (r = –0.87, p=9e-13, n=38 cell types) and high chromatin accessibility (r=0.87, p=1e-12) at a nearby enhancer (Figure 8c,g,h). By contrast, 15 other nearby enhancers have weaker, though still significant (FDR<0.05), correlation with *Stim2* expression (|r|=0.46~0.85). Moreover, *Stim2* expression also correlated significantly with 25,027 other enhancers located throughout the genome (FDR<0.05; both mCG-RNA and ATAC-RNA), most of which (n = 23,526) were on different chromosomes (Figure 8d-f). Such numerous correlations with *trans*-enhancers likely reflect gene co-expression, rather than direct causal links with the *Stim2* gene. For example, these *trans*-enhancers might directly regulate nearby genes whose expression patterns across cell types are similar to *Stim2* (Figure S12e-h,j,k).

Next, we used three-dimensional genome conformation data to test whether putative enhancer-gene links correspond to bona fide physical interactions[127]. We analyzed the 3D chromatin contact frequency of the predicted enhancer-gene pair (Figure 8c) across

homologous human brain cell types, using multi-omic snm3C-seq data[125]. Chromatin contact frequency for this enhancer was strongly correlated with *Stim2* expression (r=0.95, p=3e-4; Figure 8i; Figure S12i). By contrast, other proximal enhancers were less correlated (Figure 8j).

In addition to the challenge of widespread spurious correlations, the case of *Stim2* also illustrates the challenges associated with defining cell types[113]. For example, the same set of cells can be grouped into either 8 major types or 38 fine-grained sub-types, leading to different correlation values (Figure 8g,h; Figure S12j,k).

To address these issues, we developed a procedure that controls the risk of false positives from gene co-expression, and compares predicted links across data modalities and cell type resolutions (Figure 9a, Figure S13). We first integrate single-cell transcriptomes (RNA) and epigenomes (DNA methylation or chromatin accessibility) using correlated gene-level features across data modalities (SingleCellFusion)[10,11,21]. This allows us to build a neighbor graph connecting cells within and across data modalities (see Methods). Next, we define metacells[128], which aggregate the transcriptomic and epigenomic profiles from groups of similar cells. Each metacell has a complete bi-modal (transcriptomic and epigenomic) profile, which then allows us to correlate enhancer epigenetic features with gene expression. These metacells represent cells with an adjustable resolution, capturing both discrete and continuous patterns of variation.

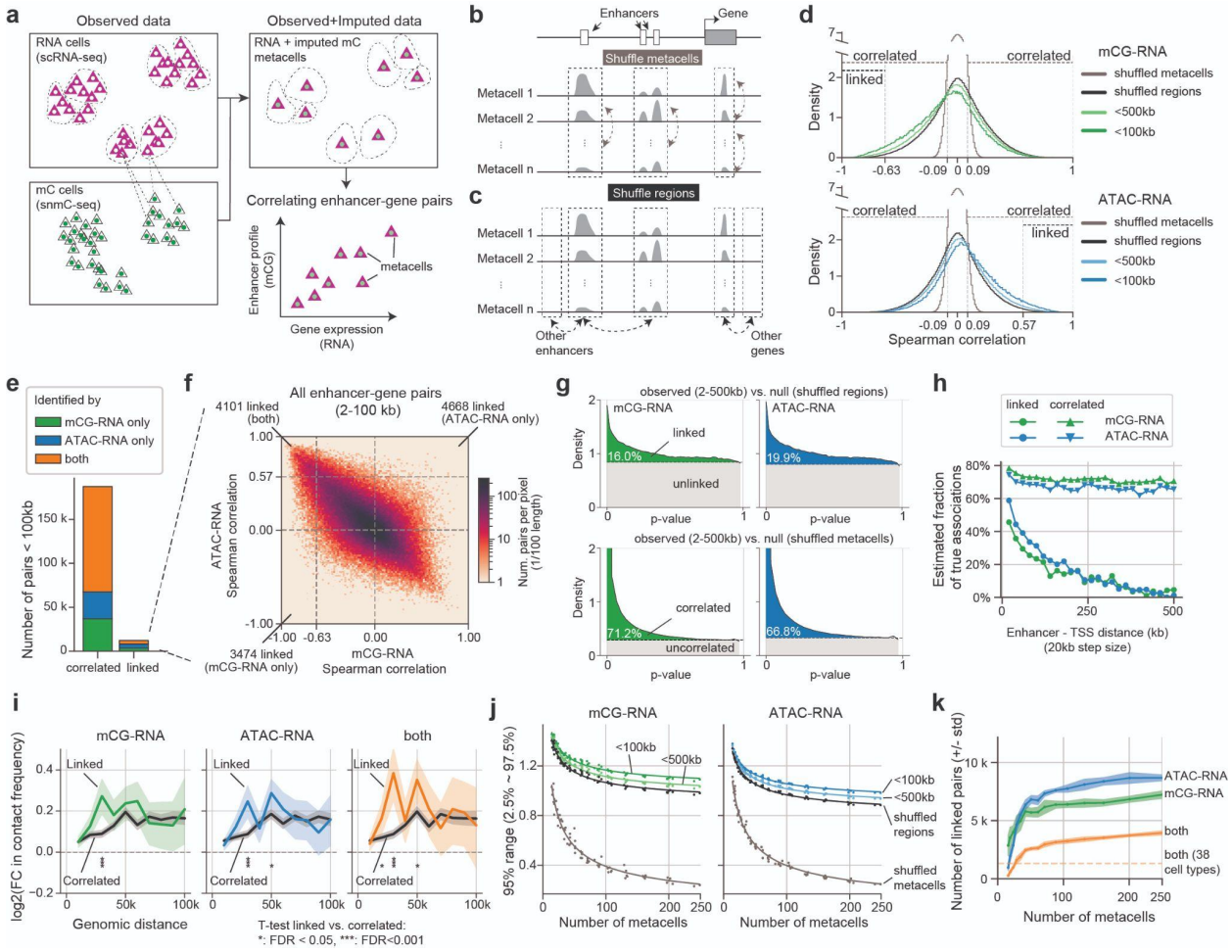


Figure 9: Stringent statistical criteria capture enhancer-gene links with consistent signatures across data modalities and cell type resolutions. a. Method for linking enhancers to target genes using metacells with bi-modality profiles. **b-c.** Null distributions derived from shuffling metacells (**b**) or shuffling regions (**c**). **d.** Distribution of enhancer-gene correlations. Bars indicate regions of statistical significance (FDR=0.2 for pairs <100kb). Two null models induce two different types of significance: linked (black bar; shuffle regions) and correlated (gray bar; shuffle metacells). **e.** The number of significantly linked or correlated pairs using mCG-RNA, ATAC-RNA, or both. **f.** Joint distribution of mCG-RNA correlation versus ATAC-RNA correlation for enhancer-gene pairs (2-100 kb). **g.** P-value histograms of enhancer-gene pairs (2-500 kb), using shuffled regions (top panels) or shuffled metacells (bottom panels). The estimated fraction of true positives is shown[129]. **h.** Estimated fraction of true associations vs. enhancer-TSS distance. **i.** Enrichment of chromatin contact frequency of linked and correlated enhancer-gene pairs compared with random genomic region pairs (mean ±95% confidence interval). Tracks are aggregated across all contacts from 8 neuronal cell types. **j.** The spread (95% range) of correlation coefficients as a function of the number of metacells. Dots represent observed data; lines represent inverse square root fit ($y \sim a/\sqrt{x} + b$). **k.** Number of linked pairs as a function of the number of metacells (FDR=0.2; mean ± standard deviation across 5 bootstrap samples with 80% of cells.)

We reasoned that genuine enhancer-gene interactions should correspond to stronger correlations than the background induced by co-expression. Correlations mediated by co-expression are inherently limited in their strength by the magnitude of gene-gene correlations, whereas direct enhancer-gene interactions can produce stronger associations. Importantly, this assumption applies to the strongest enhancer-gene interactions; weak interactions that don't exceed the background of gene co-expression cannot be detected by correlation-based methods.

To test whether the observed correlations exceed what is expected due to noise and gene co-expression, we compared the observed correlation coefficients with two null distributions: shuffling metacells[9,15,103] and shuffling regions[78,104,121] (Figure 9b-d). Shuffling metacells decouples epigenetic and transcriptomic signatures across metacells, removing both enhancer-gene correlation and gene co-expression (Figure 9b). The significance arising from this distribution is inflated by gene co-expression, potentially leading to false positives in which an enhancer-gene pair may be correlated due to shared upstream regulation rather than direct interaction. Shuffling regions retains the gene co-expression structure imposed by the hierarchical organization of cell types, but it correlates each gene's expression with distant, randomly selected enhancers (Figure 9c)[78,104,121].

As expected, the distribution obtained by shuffling regions was wider than that derived from shuffling metacells (Figure 9d), reflecting incidental correlations due to gene co-expression. Enhancer-gene pairs within 500kb of the TSS are significantly enriched in both positive and negative correlations when compared with shuffling metacells. However, when compared with shuffling regions, enrichment is only present in positive correlation for ATAC-RNA, and in negative correlation for mC-RNA. Thus, shuffling regions is a more stringent null distribution for calling significant enhancer-gene links, as it effectively controls for spurious enhancer-gene correlations due to gene co-expression.

We call an enhancer-gene pair significantly “correlated” if it passes an FDR-adjusted threshold based on shuffling metacells, whereas we reserve the term significantly “linked” for pairs that pass the criteria set by shuffling regions. We used a relatively lenient FDR threshold of 0.2 to reduce the risk of false negatives from our stringent null distribution. Linked pairs (n=12,243 within 100kb, FDR<0.2) are a subset of correlated pairs (187,343 within 100kb, FDR<0.2) (Figure 9e,f), but they have a stronger association that rises above the background from gene co-expression. Lowering the FDR threshold to 0.1 or 0.05 reduced the number of linked pairs to 3,142 and 489, respectively.

Notably, we found that removing sample covariance using GLS abolished the difference between shuffling regions and shuffling cells (Figure S14a-b). This manipulation thus removes the distinction between correlated pairs and linked pairs (Figure S14c). In addition, the shuffling-regions null distribution was robust with respect to differences in enhancer GC content and an enhancer’s distance to its nearest gene (Figure S15a-d).

We compared our results with two alternative strategies for estimating enhancer-gene interactions using single-cell epigenomes. Using open chromatin data, CICERO[119] identified 1,869 significant enhancer-gene associations located within 100kb. These significantly overlap with a subset of the correlated pairs we identified, and to a lesser degree with linked pairs (Figure S16a,b). Notably, the mean CICERO co-accessibility scores are 4.8-5.9 fold higher ($p<2e-8$) for linked pairs than for correlated pairs (Figure S16c). A second strategy, the activity-by-contact (ABC) model[118], identified enhancer-gene links using both chromatin accessibility and chromatin conformation data. This model identified enhancer-gene links for each cell type independently, without considering correlated variability in expression across cells. The ABC model identified 150,228 associations within 100kb, which significantly overlap with our correlated and linked pairs (Figure S16d,e). In addition, the ABC scores are 1.09-1.22 fold higher ($p<1e-8$) for linked pairs than for correlated pairs (Figure S16f). These results show

that linked pairs have stronger associations than correlated pairs, and are more likely to capture genuine enhancer-gene associations.

A potential pitfall of our stringent enhancer-gene linking procedure is a higher risk of false-negatives, i.e. failure to detect true interactions. We next empirically compared correlated versus linked pairs from several biological and statistical perspectives, to test whether the correlations filtered out by our method are likely false positives arising from gene co-expression.

First, we observed that correlated pairs include many enhancer-gene links with a non-canonical direction of association (Figure 9d; Figure S17a). For example, we found about a third (47,137/150,285) of these pairs had a negative correlation of gene expression with chromatin accessibility, and a similar proportion (53,687/156,932) had a positive correlation with mCG. Non-canonical associations were also reported in recent large-scale studies of brain cell epigenomes[9,15]. These correlations could suggest novel biological mechanisms such as methylcytosine-preferring transcription factors[130]. However, they may also include false-positive associations due to gene co-expression. Indeed, none of the non-canonical associations passed our threshold for linked pairs (Figure 9d). This is consistent with the canonical understanding of enhancer activity associating with low DNA methylation and open chromatin.

Second, as enhancer-gene interactions are mostly concentrated within ~100-500 kb around gene promoters[117,118], we compared the distance dependence of linked and correlated pairs. Using a p-value histogram method[129], we estimated 16.0-19.9% of enhancers that are 2-500kb away from a promoter are linked (Figure 9g, Figure S17b). A much larger fraction (66.8-71.2%) were correlated. Notably, the proportion of correlated pairs remains high even for distal pairs (e.g. >60% for pairs >1 Mb or on other chromosomes), whereas <5% of these pairs are linked (Figure 9h, Figure S17c). These correlated pairs contradict the biological understanding that most enhancers activate genes in *cis*; the linked pairs are more coherent with this canonical framework.

Third, we validated our predicted links with independent chromatin conformation data from the human brain[125]. We reasoned that linked enhancer-gene pairs which are conserved across species should have higher chromatin contact frequency compared with random regions. Indeed, we found enrichment of contact frequency for both linked (mean fold change (FC) = 1.15, p=2e-4) and correlated pairs (mean FC = 1.10, p=1e-5). Moreover, linked pairs located 10-30 kb apart have higher levels of contact enrichment than correlated pairs (FDR<0.05; Figure 9i, Figure S17d,e).

A key parameter for our analysis is the cell type granularity, as determined by the number of metacells. The sparse genomic coverage of single-cell sequencing and the limited number of profiled cells create a tradeoff between the number of metacells and the quality of each metacell--i.e. between fine-grained resolution and signal/noise ratio. As the number of metacells (N) increases, the width of the null distribution for the shuffled metacells approaches zero as $\frac{1}{\sqrt{N}}$, which is consistent with independent random signals for each metacell (see Methods; Figure 9j; Figure S18a-c). By contrast, the range of the null distribution for shuffled regions does not vanish for large N, but instead asymptotes at a non-zero value that reflects gene co-expression (Figure S18c). Notably, the shuffling-regions null distribution is less sensitive to the number of metacells, and more closely reflects the behavior of the observed correlations. This suggests enhancer-gene link calling using shuffling-regions is less sensitive to the choice of cell type granularity than using shuffling-metacells. We found more linked pairs as the number of metacells increases, but with diminishing returns after $N > 50$. (Figure 9k; Figure S18d).

We used our procedure to comprehensively examine regulatory interactions in neurons of the mouse primary motor cortex[10]. Linked enhancer-gene pairs formed 15 modules that capture diverse cell-type-specific signatures (Figure 10a,b). For example, genes in module 13 are specifically expressed in pan-inhibitory neurons, with corresponding low CG methylation

level and open chromatin at linked enhancers. Module 9 is most active in caudal ganglionic eminence (CGE) derived inhibitory neurons (Lamp5, Sncg, and Vip) and in superficial-layer excitatory neurons (L2/3 IT and L4/5 IT). These consistent gene- and enhancer-level signals integrated from three data modalities provide strong support for our identified enhancer-gene associations.

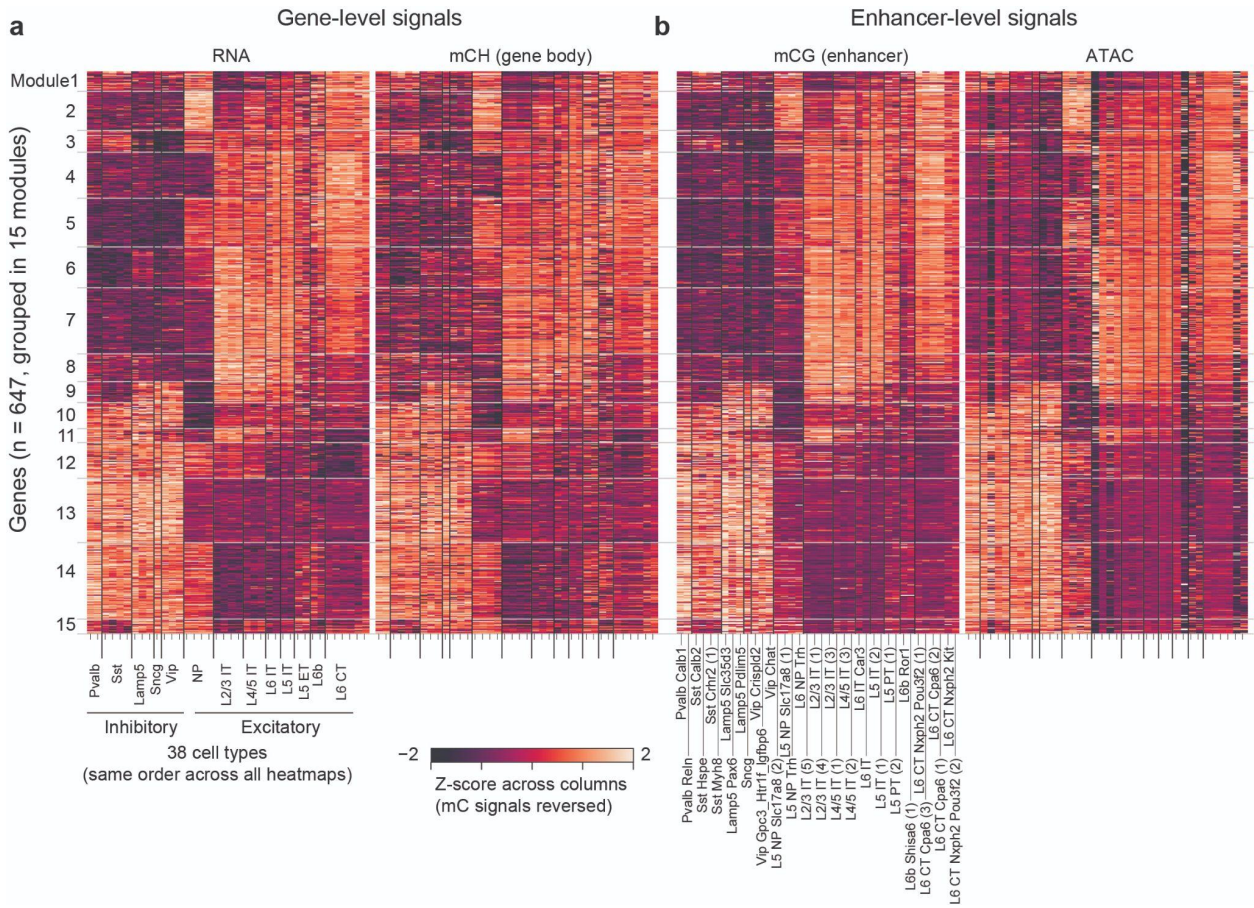


Figure 10: Consistent gene- and enhancer-level signatures for hundreds of enhancer-gene links. a-b. Gene expression, gene body DNA methylation (**a**), and enhancer mCG and ATAC signal (**b**) across cell types. Genes are organized into 15 modules by K-means clustering. Enhancers are ordered according to the genes they are linked to (FDR < 0.2 for both mCG-RNA and ATAC-RNA across n=38 cell types). Signals from multiple enhancers linked to the same gene were averaged. The colormap for the mC modalities (gene body mCH and enhancer mCG) are reversed.

Our analyses highlight the challenge of distinguishing genuine enhancer-gene interactions from spurious correlations due to gene co-expression. We addressed this by empirically estimating the expected correlations for unlinked enhancer-gene pairs under co-expression, and comparing results across different epigenetic assays and cell type granularities. Notably, mCG-RNA and ATAC-RNA associations show striking similarities (Figure 9d,f-k; Figure 10), despite measuring distinct epigenetic features with opposite effects on gene expression. Predicted enhancer-gene links are robust with respect to a wide range of cell type granularities (Figure 9k). We identified hundreds of genes and thousands of linked cCREs with highly coordinated gene- and enhancer-level activities (Figure 10a,b).

Correlation-based analysis has notable limitations. First, this approach cannot identify constitutive enhancer-gene links that are present in all cell types. Larger datasets including more diverse tissues or cell types may partly address this limitation. Second, rigorous control for spurious correlations limits the power of detecting genuine but weak enhancer-gene interactions. Finally, true causal interactions cannot be inferred from correlational analysis alone. The links we identified (Figure 10a,b) are strong candidates for causal enhancer-gene interactions, which must be tested by perturbative experiments[131,132]. Future experimental validation, including large-scale assays[117,118,133], will be needed to test correlation-based predictions. By bringing together multiple data modalities to define robust enhancer-gene links, these analyses can reveal the regulatory principles of cell-type-specific gene expression.

3.4. Methods

3.4.1. Datasets

We used three single-cell sequencing datasets from the mouse primary motor cortex (MOp)[10]. They are scRNA-seq (single cell; 10x genomics V3; Allen Institute for Brain Science), snmC-seq (single nucleus; DNA methylation; Ecker lab from the Salk Institute), and snATAC-seq (single nucleus; chromatin accessibility; Ren lab from UCSD). Only high-quality neuronal cells, as determined in Ref[10] (from its Supplementary Table 2; column SCF/SingleCellFusion), are retained for our analysis. These datasets are publicly available and provided by a previous study (Ref[10]; https://assets.nemoarchive.org/dat-ch1nqb7). The starting point of all analyses are gene-by-cell matrices and/or enhancer-by-cell matrices depending on the data modality. For the scRNA-seq dataset, we start from the gene-by-cell count matrix. For the snATAC-seq dataset, we quantified both enhancer-by-cell and gene-by-cell count matrices. For the snmC-seq dataset, we quantified enhancer-by-cell CG DNA methylation profiles and gene-by-cell non-CG (CH) DNA methylation profiles. The DNA methylation profile for a particular region and cell can be summarized by two numbers: the number of methylated cytosines (mC) and the total number of cytosines covered (C). The DNA methylation level is the ratio of mC to C (mC/C). Please see sections below for dataset specific procedures of normalizations. The mouse gene annotation file is downloaded from gencode (vM16). The enhancer list is adapted from the putative enhancer list from Ref[10] (see below).

3.4.2. Calling putative enhancers

We constructed our putative enhancer list based on the mouse MOp neuronal cell type-specific putative enhancers from Ref[10] (from its Supplementary Table 7). In that study, the enhancers are called using REPTILE[82], an algorithm that uses the DNA methylation and ATAC-seq profiles of 13 mouse neuronal cell types, as well as mouse embryonic stem cells, as input. Starting from this list, we first selected regions with enhancer score >0.5 and merged

overlapping regions using bedtools[134]. We subsequently removed regions overlapping any gene promoter regions (transcription start site +/- 2kb; all transcripts from gencode vM16), exons (vM16), and ENCODE blacklist[79]. This leaves us with 233,524 enhancers in total, with a median size of ~250 bp (Figure S12b; Table S10).

3.4.3. Curated cell types

For analyses related to Figure 8, we curated a list of 38 neuronal cell clusters based on the SingleCellFusion clusters (L1 and L2, with n=29 to 56 cell types respectively) in Ref[10]. We aimed to merge small clusters to increase pseudo bulk coverage at enhancers, while retaining as much cell type diversity as possible. To achieve this, we first call an enhancer *covered* in a cluster if it has at least 20 sequenced CpG sites in that cluster, where the cluster-level coverage is the sum of cell-level coverages. Next, we call an enhancer *common*, if it is covered in more than half of the L2 clusters. We call a cluster *covered*, if more than half of the common enhancers are covered in that cluster. For each L1 cluster we then evaluate 3 cases:

1. If the cluster itself is not covered, we drop it along with all its child (L2) clusters.
2. Else if less than 2 (n<2) of its child (L2) clusters are covered, we retain the L1 cluster itself, but drop all its child (L2) clusters.
3. Else if at least 2 (n≥2) of its child (L2) clusters are covered, we retain the covered L2 clusters, but drop the uncovered L2 clusters and the L1 cluster.

This procedure resulted in 38 clusters with adequate coverage. Table S13 summarized the correspondence between the 38 clusters we get from this procedure and the cell types defined in Ref[10].

To compare with the cell types in snm3C-seq data[125], we further merged these 38 fine-grained clusters into 8 major clusters based on the well-established neuronal cell type taxonomy[2]. Table S13 summarized the correspondence between the 38 fine grained and the 8 major cell clusters defined in this study and those defined in Ref[10,125].

3.4.4. Clustering and defining metacells

For analyses related to Figure 9, we generated cell clusterings with a range of cluster resolutions. We start by normalizing the scRNA-seq count matrix with $\log_{10}$(CPM+1), where CPM stands for counts per million mapped reads. We then calculated the top 50 principal components (PCs), and built a k-nearest neighbor graph (k = 30) connecting cells according to the Euclidean distance in the PC space. We used Leiden community detection to generate clusters[70]. Different resolution parameters (r = 1 ~ 794) were chosen to generate clusters with different granularity (n = 13 ~ 8850 metacells). The pseudo bulk profiles from each of the individual clusters were used as metacells.

3.4.5. Feature selection and normalization

We preprocessed the data matrices separately for each data modality. The starting point is always cell-level matrices containing counts (RNA and ATAC) or methylation level (mC). To get cluster-level (metacell) matrices, we summed counts from cells in the same clusters (metacells) to create pseudo-bulk samples. For methylation data, we summed methylated counts and total counts (coverage) separately. Next, we normalized matrices as follows:

- For an RNA matrix (gene-by-cluster/metacell), we normalize the raw count matrix with $\log_{10}$(CPM+1).
- For an ATAC matrix (enhancer-by-cluster/metacell), we normalize the raw count matrix with $\log_{10}$(TPM+1), where TPM stands for transcripts per million mapped reads. Enhancers that are covered in <50% of clusters are removed.
- For a gene body mCH matrix (gene-by-cluster/metacell), we first removed low coverage genes if the gene has <50% clusters surpassing 1000 counts in the gene body (or < 80% metacells surpassing 20 counts). We then take the ratio of the number of

methylated to the number of coverage to get the methylation fraction. All the steps here consider cytosines in non-CG (CH) dinucleotide context only.

- For an enhancer mCG matrix (enhancer-by-cluster/metacell), we first removed low coverage enhancers if the gene has <50% clusters surpassing 20 counts (or <80% metacells surpassing 5 counts) in the enhancer region. We then take the ratio of the number of methylated to the number of coverage to get the methylation fraction. All the steps here consider cytosines in CG dinucleotide context only.

After normalization and filtering of individual matrices, we then consider only enhancers that are shared in both ATAC and mCG matrices for downstream analyses.

3.4.6. Correlating enhancer-gene pairs across cell types

We calculate the Spearman correlation coefficient between any pair of enhancer and gene that are within 1 Mbp (enhancer center to gene TSS) across curated cell types (n=38 or n=8). This was done separately for enhancer mCG vs. RNA and enhancer ATAC vs. RNA. Enhancer mCG signals are normalized by the global mean mCG levels of each cell type; enhancer ATAC signals are $\log_{10}$(TPM+1) normalized; RNA expression levels are $\log_{10}$(CPM+1) normalized.

To assess the statistical significance of the enhancer-gene correlations, we repeated the correlation analysis with 2 types of data shuffling control, as explained in the main text. To control for random noise, we shuffled cell cluster labels of the gene-by-cluster RNA matrix, followed by calculating correlation coefficients. To control for background co-expression across enhancer-gene pairs, we shuffled gene labels of the gene-by-cluster RNA matrix, followed by calculating correlation coefficients.

3.4.7. Correlating enhancer-gene pairs across metacells

Given a transcriptomic dataset (scRNA-seq) and an epigenetic dataset (e.g. snmC-seq) collected from the same tissue, we first generate a constrained k-nearest neighbor network linking cells across the two modalities (SingleCellFusion; Ref[10,11]). This network allows us to impute the DNA methylation profiles (mC) for each RNA cell. We then cluster scRNA-seq cells using Leiden community detection [70] (see section **Clustering/Generating metacells**). We call these clusters *metacells*, to emphasize that they do not necessarily correspond to discrete cell types, but could also capture continuous changes among cell populations. These preparations allow us to construct bi-modal profiles for each metacell, by aggregating counts--either observed or imputed--from cells in the same metacells. Finally, we evaluate the correlations between enhancer-gene pairs across metacells.

To be specific, the starting point of this analysis involves 4 matrices: an enhancer-by-cell mCG (or ATAC) matrix E_{ec}, a gene-by-cell RNA matrix $R_{gc'}$, a cross-modal cell-to-cell k nearest neighbor matrix: $K_{cc'}$, and a metacell assignment matrix of RNA cells $K_{c'z}$. Here we use c, $c'$and z to denote an mC cell, an RNA cell, and a metacell, respectively. A metacell is a group of RNA cells generated by Leiden clustering. We use g and e to denote an enhancer and a gene, respectively. All matrices contain unnormalized raw counts. $K_{cc'}$ is generated by SingleCellFusion[10,11] with default settings and cross-modal k=30. $K_{c'z}$ is generated by Leiden clustering on the RNA-seq dataset as mentioned in previous sections.

To get bi-modal profiles for a metacell, we aggregate counts from the cells belonging to that metacell: $R_{gz} = \sum_{c'} R_{gc'}K_{c'z}$, and $E_{ez} = \sum_{c} E_{ec}K_{cc'}K_{c'z}$. The metacell profiles are then normalized as mentioned in previous sections to adjust for metacell size, library size, and gene length. Finally, normalized R_{gz} and E_{ez} allow us to correlate a specific pair of gene $g^{(i)}$and

enhancer $e^{(i)}$ across metacells (z). We calculated Spearman correlation coefficients for all enhancer-gene pairs with distance between 2kb to 1Mb (enhancer center - TSS).

3.4.8. Estimating the statistical significance of enhancer-gene links

To assess the statistical significance of a correlation coefficient r, we constructed two null distributions by shuffling metacells (Figure 9b) and shuffling regions (Figure 9c). In the first case, we shuffle metacell labels independently for transcriptomic and epigenetic data, such that the two data modalities become independent of each other. In the second case, we permute and enhancers randomly from their original genomic location to the locations of other genes and enhancers, while retaining the linked bi-modal profiles of each metacell.

Either null distribution can be used to get empirical p-values and false discovery rate (FDR). The empirical p-value of a correlation coefficient r is defined as the cumulative fraction of the null distribution that has more extreme (stronger) correlation coefficients than r. We calculated two-sided p-values when using the shuffled metacells distribution, and single-sided p-values when using the shuffled regions distribution. FDRs are then calculated using the Benjamini-Hochberg procedure[135]. We call an enhancer-gene pair significantly *linked* (*correlated)* if its empirical FDR is <0.2 using shuffling regions (metacells) as the null.

To see if the shuffled regions distribution depends on enhancer properties such as its sequence GC content and distance to the nearest gene, we also performed stratified shuffling analyses (Figure S15). We first grouped enhancers into 10 bins (deciles) according to their GC content or distance to the nearest gene. We then shuffled enhancers within each bin and compared observed enhancer-gene correlations with shuffled ones for each bin separately.

3.4.9. Enrichment of 3D chromatin contact frequencies

We validated the predicted enhancer-gene links using single-cell measurements of 3D-chromatin contact frequency in human prefrontal cortex [125]. Raw contact matrices of 8

neuronal cell types were downloaded as mcool files [125]. We calculated contact frequencies from raw counts using matrix balancing using Cooler[136,137]. We then focused on analyzing these contact frequency matrices at a resolution of 10kb non-overlapping genomic bins across the genome.

To compare our enhancer-gene links predicted in the mouse brain with the chromatin contact data from human brain, we lifted genes (gencode vM16 whole genes) and putative enhancers from mm10 to hg38 using LiftOver[138] with parameters -minMatch=0.8 and -minBlocks=1.00.

To calculate enrichment, we first assigned enhancers (center) and genes (TSS) to their corresponding genomic bins (non-overlapping 10kb bins genomewide). We compared the contact frequencies of the predicted enhancer-gene pairs with random genomic region pairs with similar genomic distance. We separately tested the enrichment of contact frequencies of 6 groups of predicted enhancer-gene pairs: mCG-RNA linked, ATAC-RNA linked, pairs linked by both modalities, mCG-RNA correlated, ATAC-RNA correlated, and pairs correlated in both modalities. For each of the 8 neuronal cell types, we only include pairs that are active in the specific cell type, i.e. whose gene expression is greater than the median across all 8 cell types.

Comparison with CICERO.

We installed the R package CICERO[119] from the Bioconductor following the instructions from the authors' tutorial (https://cole-trapnell-lab.github.io/cicero-release/docs_m3/#constructing-cis-regulatory-networks). We ran CICERO on MOp ATAC-seq data using default parameters. The program takes as input a peak-by-cell ATAC-seq matrix, where peaks include both putative enhancers we specified and gene promoters (500 bp upstream of TSS). The program returns co-accessibility scores for peak pairs. We filtered the output down to enhancer-promoter pairs only, removing enhancer-enhancer and promoter-promoter pairs. We also focused on analyzing enhancer-gene

pairs that are within 100kb apart, to compare with our correlation-based analysis. We used a threshold = 0.2 following Ref[120] to call positive enhancer-gene pairs.

Comparison with the ABC model.

We downloaded code from the github repository of the ABC model[118] (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction) and followed instructions. We ran ABC for each MOp cell type (n=38) using our identified putative enhancer list (n=233,524) and pseudo-bulk ATAC-seq and RNA-seq data as input. We used genomic-distance based power law estimation to model chromatin contacts (--score_column powerlaw.Score). The software returns a score (ABC score) for each enhancer-gene pair and cell type. We excluded the expressed genes from the results, as suggested by the authors. We also focused on analyzing enhancer-gene pairs that are within 100kb. We used a threshold = 0.022 as recommended by the authors to call positive enhancer-gene pairs.

Generalized least squares (GLS) analysis to decouple covariance across metacells.

We used GLS [124] to test the association between gene expression and enhancer activity across cell types (metacells). We will focus on only one given enhancer-gene pair $(g,\ e)$, as the same procedure applies to all enhancer-gene pairs independently. Given an enhancer e and gene g, Let y_{cg} be the mRNA expression in cell type c, x_{ce} be the enhancer activity (e.g., mC or ATAC). Let C be the number of cell types. A linear model associating g and e can be written as:

$$y_c = a + \beta x_c + \varepsilon_c \qquad \text{(eq. 1)}$$

where c is the index for cell types, β is the association strength, and ε is a noise term. In addition, a is an intercept term that can be omitted after data centering (x and y can be pre-centered to ensure $E[y_c] = E[x_c] = 0$). In matrix notation, (eq. 1) can be simply noted as $y = \beta x + \varepsilon$.

In ordinary least squares (OLS), we assume ε is uncorrelated across cell types: $E[\varepsilon_c] = 0,\ E[\varepsilon_c \varepsilon_{c'}] = \sigma^2 \delta_{c,c'}$. The correlation coefficient $r = E[xy]/\sigma_x \sigma_y$ is then a measure of the linear association, and it has an associated p-value calculated using the t distribution. Alternatively, inference can be performed by permutation analysis to get an empirical p-value.

However, in our case we have correlated noise: $E[\varepsilon_c \varepsilon_{c'}] = \Omega_{c,c'}$, which reflects the correlation between cell types due to gene co-expression. That is, $\Omega_{c,c'}$ represents the background of correlated variability in gene expression due to the hierarchical structure of cell types in complex tissues. We can estimate the correlation using the genome-wide covariance, $\hat{\Omega}_{c,c'} = Cov[y]_{c,c'}$. In this case, generalized least squares[124] (GLS) can be used to give an estimate of the coefficient β. This corresponds to transforming the variables x, y from the original basis (cell types/metacells, denoted c) to an decorrelated basis (denoted r), and then performing OLS on the decorrelated variables.

We first use singular value decomposition (SVD) to decompose the mean-subtracted gene expression matrix, $y_{cg} = \sum_r U_{cr} S_{rr} V^T_{rg}$, where $r = min(c, g)$. Defining $Z = US$, we have $\Omega = ZZ^T$. Multiplying both sides of (eq. 1) by $Z^{-1} = S^{-1}U^T$ corresponds to a transformation from correlated to decorrelated (or whitened) basis:

$$y' = \beta x' + \varepsilon' \qquad \text{(eq. 2)}$$

where $y' = Z^{-1}y$, $x' = Z^{-1}x$, and $\varepsilon' = Z^{-1}\varepsilon$. The noise term is now uncorrelated, because

$$Cov[\varepsilon'] = E[\varepsilon'\varepsilon'^T] = E[Z^{-1}\varepsilon\ \varepsilon^T (Z^{-1})^T] = Z^{-1}\Omega(Z^{-1})^T = Z^{-1}ZZ^T(Z^{-1})^T = I$$

where I is the identity matrix. We can therefore use the correlation coefficient and its associated test statistics on transformed data $y'$and x', as in the case of OLS.

3.4.10. Expected range of correlation coefficients for independent variables

Here we provide theoretical justification on why we expect the range of correlation coefficients ($\hat{r}$) to scale as $\frac{1}{\sqrt{N}}$, as seen in Figure 9j and Figure S18b, where N is the number of metacells.

Let X and Y be two independent random variables. Let x_i and y_i be independent and identically distributed samples of X and Y, where $i \in \{1, 2, \ldots, N\}$. In our case, N represents the number of metacells, and x_i and y_i are the transcriptomic and epigenetic signals for a given enhancer-gene pair for metacell i. We require X and Y to be independent of each other as they are unlinked, and x_i and y_i be independent samples as different metacells are also independent observations of X and Y, such as in the case of null distribution created by shuffling cells.

To simplify the notation, we assume $E[X] = E[Y] = 0$, as the mean does not affect correlation coefficient r. We also assume X and Y are symmetric, as in the case of normal distribution. It is obvious that $r(X, Y) = 0$. However, we are interested in how the variance of $\hat{r}$ depends on N, where $\hat{r}$ is the sample estimate of r by $\{x_i\}$ and $\{y_i\}$.

$$var[\hat{r}] \sim E(\hat{r}^2) \sim E\left[\frac{(\sum_{i=1}^{N} x_i y_i)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = E\left[\frac{\sum_{a=1}^{N}\sum_{b=1}^{N} x_a y_a x_b y_b}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = \sum_{a=1}^{N}\sum_{b=1}^{N} E\left[\frac{x_a y_a x_b y_b}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = \sum_{a=1}^{N} E\left[\frac{(x_a y_a)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right]$$

(eq.3)

The last equality holds, as only non-interaction terms ($a = b$) are nonzero. Moreover, as $(x_a y_a)^2$ are equivalent for different $a = \{1 \ldots N\}$, the above summation can be further simplified as:

$$\sum_{a=1}^{N} E\left[\frac{(x_a y_a)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = N \cdot E\left[\frac{(x_1 y_1)^2}{\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i^2}\right] = N \cdot E\left[\frac{x_1^2}{\sum_{i=1}^{N} x_i^2}\right] \cdot E\left[\frac{y_1^2}{\sum_{i=1}^{N} y_i^2}\right], \quad \text{(eq. 4)}$$

where $E[\frac{x_1^2}{\sum_{i=1}^{N} x_i^2}] = \frac{1}{N} E[\frac{\sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} x_i^2}] = \frac{1}{N}$, due to the symmetry among indices. Therefore, we finally arrive at

$$var(\hat{r}) \propto N \cdot E[\frac{x_1^2}{\sum_{i=1}^{N} x_i^2}] \cdot E[\frac{y_1^2}{\sum_{i=1}^{N} y_i^2}] = N \cdot \frac{1}{N} \cdot \frac{1}{N} = \frac{1}{N}, \quad \text{(eq. 5)}$$

and thus the range of the distribution goes as $\frac{1}{\sqrt{N}}$.

3.5. Acknowledgements

Chapter 3, in full, has been submitted for publication of the material. The preprint of this manuscript is posted on *bioRxiv*. **F. Xie**, E. J. Armand, Z. Yao, H. Liu, A. Bartlett, M. Margarita Behrens, Y. E. Li, J. D. Lucero, C. Luo, J. R. Nery, A. Pinto-Duarte, O. Poirion, S. Preissl, A. C. Rivkin, B. Tasic, H. Zeng, B. Ren, J. R. Ecker, E. A. Mukamel, Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes. *bioRxiv* (2021), p. 2021.10.25.465795. The dissertation author was a co-first author of this paper.

EAM and FX designed the study. ZY, BT, and HZ generated scRNA-seq data. HL, AB, MMB, JDL, CL, JRN, APD, ACR and JRE generated DNA methylation (snmC-Seq) data. HL, MMB, YEL, JDL, APD, OP, SP, and BR generated snATAC-Seq data. FX led the computational analysis. FX and EA developed code and performed analysis. FX, EA, and EAM wrote and edited the manuscript. All authors approved the manuscript.

We gratefully acknowledge members of the Mukamel, Ecker, Ren, and Zeng laboratories collaborators within the BRAIN Initiative Cell Census Network (BICCN). This work was funded by the NIH BRAIN Initiative (RF1 MH120015 to E.A.M.; U19MH114830 to H.Z.; U19MH121282 to J.R.E.; J.R.E is an Investigator of the Howard Hughes Medical Institute) and by CZI Collaborative Computational Tools for the Human Cell Atlas (to E.A.M.).

APPENDIX

Competing interests

B.R. is a shareholder of Arima Genomics, Inc. P.V.K. serves on the Scientific Advisory Board to Celsius Therapeutics Inc. A.R. is an equity holder and founder of Celsius Therapeutics, an equity holder in Immunitas, and an SAB member in Syros Pharmaceuticals, Neogene Therapeutics, Asimov, and Thermo Fisher Scientific.

Data Availability

Chapter 1

The BICCN MOp data (RRID:SCR_015820) can be accessed via the NeMO archive (RRID:SCR_016152) at accession: https://assets.nemoarchive.org/dat-ch1nqb7. Visualization and analysis resources: NeMO analytics: https://nemoanalytics.org/, Genome browser: https://brainome.ucsd.edu/BICCN_MOp, Epiviz browser: https://epiviz.nemoanalytics.org/biccn_mop.

Chapter 2

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).

Chapter 3

The scRNA-seq, snmC-seq, and snATAC-seq datasets from the mouse primary motor cortex are generated by BICCN (RRID:SCR_015820) as reported previously[10]. The data can be

accessed via the NeMO archive (RRID:SCR_002001) at accession: https://assets.nemoarchive.org/dat-ch1nqb7. Genome browser: https://brainome.ucsd.edu/BICCN_MOp. The chromatin contact data generated by snm3C-seq is downloaded from publicly available files (Ref[125]; https://salkinstitute.app.box.com/s/fp63a4j36m5k255dhje3zcj5kfuzkyj1).

Code availability

Chapter 1

Tool	Purpose	Reference
scrattch.hicat: Hierarchical, Iterative Clustering for Analysis of Transcriptomics	RNA clustering	https://github.com/AllenInstitute/scrattch.hicat,
SnapTools	ATAC-seq analysis	https://github.com/r3fang/SnapTools
YAP (Yet Another Pipeline)	DNA methylation (snmC-seq) mapping and cluster-level aggregation.	https://github.com/lhqing/cemba_data documentation: cemba-data.rtfd.io
MetaNeighbor	Cluster reproducibility analysis	https://github.com/gillislab/MetaNeighbor-BICCN
LIGER (Linked Inference of Genomic Experimental Relationships)	Multi-modal integration, embedding, and clustering	https://github.com/welch-lab/liger
SingleCellFusion	Multi-modal integration, embedding, and clustering	https://github.com/mukamel-lab/SingleCellFusion
Conos: Clustering on Network of Samples	Cluster reproducibility analysis	https://github.com/kharchenkolab/conos
STAR v2.5.3	RNA-seq alignment	[62]
Bismark	DNA methylation (snmC-seq) alignment	[69]

Chapter 2

- SingleCellFusion: https://github.com/mukamel-lab/SingleCellFusion.
- LIGER: https://github.com/welch-lab/liger.
- Bismark v0.14.4: http://www.bioinformatics.babraham.ac.uk/projects/bismark/ ; RRID:SCR_005604.
- STAR 2.5.2b: https://github.com/alexdobin/STAR; RRID:SCR_015899
- YAP: https://hq-1.gitbook.io/mc/.
- ALLCools: https://github.com/lhqing/ALLCools.
- Methylpy: https://github.com/yupenghe/methylpy.
- Seurat v4.0.0: https://satijalab.org/seurat/; RRID:SCR_016341
- Scanorama v1.7: https://github.com/brianhie/scanorama.
- Harmony (pyharmony): https://github.com/iandday/pyharmony**.**

Chapter 3

- Customized code for this study: https://github.com/FangmingXie/scf_enhancer_paper.
- SingleCellFusion: https://github.com/mukamel-lab/SingleCellFusion.
- ABC model: https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction; Ref[118].
- CICERO: https://www.bioconductor.org/packages/release/bioc/html/cicero.html; Ref[119].

Supplemental tables

Table S1:
List of datasets, number of cells, and other parameters of each dataset. Data from this study are available via the Neuroscience Multi-omics Archive (NEMO, RRID:SCR_016152) at https://assets.nemoarchive.org/dat-ch1nqb7.

Table S2:
List of all cells with cluster assignments from 3 computational methods (RNA consensus, SingleCellFusion, LIGER).

Table S3:
Proportion of each cell type within major cell class for each transcriptomic modality.

Table S4:
Cluster analysis and metadata for each dataset on its own. Eight individual files:

1. S4a - scRNA SMART
2. S4b - scRNA 10x v3 A
3. S4c - scRNA 10x v2 A
4. S4d - snRNA SMART
5. S4e - snRNA 10x v3 B
6. S4f - snRNA 10x v3 A
7. S4g - Open chromatin (ATAC-seq). Note that this table includes 3 columns describing the Class, Major Type, and Subtype of each cell as described in a related paper analyzing snATAC data from the whole mouse brain[9].
8. S4h - DNA methylation (snmC-seq2)

Table S5:
Full gene-by-cluster tables for each dataset, using the SingleCellFusion Level 2 clusters. Eight individual files for each dataset:

1. S5a - scRNA SMART
2. S5b - scRNA 10x v3 A
3. S5c - scRNA 10x v2 A
4. S5d - snRNA SMART
5. S5e - snRNA 10x v3 B
6. S5f - snRNA 10x v3 A
7. S5g - Open chromatin (ATAC-seq)
8. S5h - DNA methylation (snmC-seq2)

Table S6:
For each of the 116 consensus transcriptomic cell types, we performed differential expression (DE) analysis with respect to each of the other cell types. The table reports the top 50 conserved DE genes in each direction for each comparison. Conserved DE genes are significant in at least one dataset, while also having more than two-fold change in the same direction in all but one datasets.

Table S7:
Enhancers predicted for each cell type based on integrated DNA methylation and ATAC-Seq data using REPTILE.

Table S8:
List of SingleCellFusion clusters at three levels of cluster resolutions (L0, L1, L2).

Table S9:
Cluster annotations and unique accession IDs.

Table S10: A list of putative enhancers (cCREs; n=233,524 in total)

Table S11: Significant linked enhancer-gene pairs by mCG-RNA correlation

Table S12: Significant linked enhancer-gene pairs by ATAC-RNA correlation

Table S13: Cell type correspondence between this study and related studies

Supplemental figures

Figure S1: A multimodal molecular cell type atlas of mouse primary motor cortex (MOp). **a,** Anatomical location of mouse MOp in the Allen Mouse Brain Common Coordinate Framework (CCFv3) in 3D and in representative sagittal and coronal sections. **b-d,** Documentation of MOp samples collected at the Allen Institute (b), the Broad Institute (c), and the Salk Institute (d). Each panel shows a diagram of coronal brain slices and dissected regions for transcriptomic (sc/snRNA-seq) and epigenomic (snATAC and snmC-Seq) data samples based on the Allen Mouse Brain Common Coordinate Framework (CCF). Nissl-stained images in (d) show the posterior face of tissue slices (600 µm thickness). **e,** Number of cells and median number of unique sequenced DNA or RNA fragments per cell in each of 9 single-cell transcriptome and epigenome datasets. Squares show the extrapolated total library size based on sequence duplication rate. **f,** Number of cells in each of the major cell classes (glutamatergic excitatory, GABAergic inhibitory neurons, non-neurons) of each dataset. Differences in cell type sampling strategy, including the use of cell sorting to enrich neurons, affect the relative number of neurons and non-neuronal cells. Datasets include cells from the following numbers of animals (see Table S1): scRNA SMART: n=28 male, 17 female; scRNA 10x v3 A: n=3 male, 3 female; scRNA 10x v2 A: n=3 male; snRNA SMART: n=8 male, 2 female; snRNA 10x v3 B: n=5 male, 6 female; snRNA 10x v2: n=2 male, 1 female; snRNA 10x v3 A: n=1 female; snmC-seq and snATAC-seq: n=2 replicates, each pooled from 6-30 male animals. **g,** NeMO Analytics (nemoanalytics.org) visualization and analysis environment for the BICCN mouse molecular mini-atlas. Screenshot of NeMO Analytics showing multi-omic results for glutamate decarboxylase 2 (*Gad2*), a marker gene in inhibitory neurons. The web portal has the following features: (1) Search box for gene names; (2) Indicator of gene viewed; (3) Expandable species-specific functional annotation; (4) Link-outs to additional resources for the selected gene; (5,6,7) interactive visualizations of each BICCN dataset, displayed in a ‘standalone’ box showing gene expression and cell clustering on integrated UMAP coordinates. Additional data exploration options for each of the datasets are available via the drop-down menu at the upper right corner of the NeMO Analytics dataset titles. (8) An embedded Epiviz interactive workspace to visualize scATAC-seq and sncMethyl-seq datasets in a linear browser view (8a), here showing the average ATAC and % CG methylation at the *Gad2* locus (8c,8d) as well as in each major cluster of glutamatergic and GABAergic neurons (8b,8e,8f). Epigenomic data are also available at http://epiviz.nemoanalytics.org/biccn_mop, and instructions for setting up and extending the Epiviz workspaces are available at http://github.com/epiviz/miniatlas. **h,** Brainome epigenomics portal (https://brainome.ucsd.edu/BICCN_MOp). The portal shows single base resolution epigenomic and transcriptomic data (snmC-Seq, snATAC-Seq, sc/snRNA-Seq) using the AnnoJ browser. Drop-down menus allow the user to select groups of cells (e.g. Excitatory, Inhibitory, MGE-Derived, etc.), modalities (mCG, mCA, ATAC, scRNA, snRNA, enhancers), and display options. A Cell Browser allows visualizing scatter plots and heatmaps of groups of genes across data modalities.

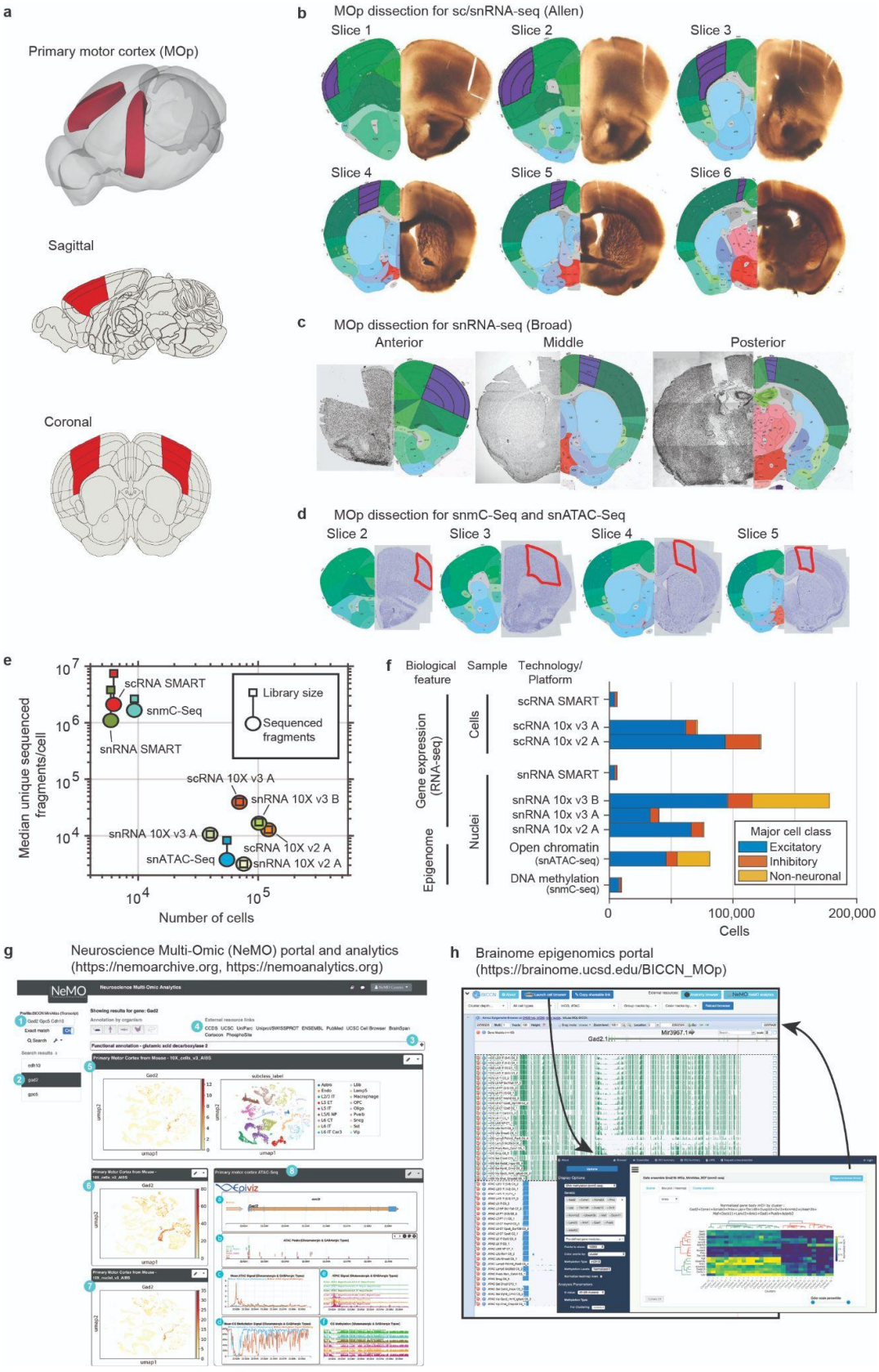

a
Primary motor cortex (MOp)
Sagittal
Coronal
b
MOp dissection for sc/snRNA-seq (Allen)
Slice 1
Slice 2
Slice 3
Slice 4
Slice 5
Slice 6
c
MOp dissection for snRNA-seq (Broad)
Anterior
Middle
Posterior
d
MOp dissection for snmC-Seq and snATAC-Seq
Slice 2
Slice 3
Slice 4
Slice 5
e
Median unique sequenced fragments/cell
Number of cells
scRNA SMART
snmC-Seq
snRNA SMART
Library size
Sequenced fragments
scRNA 10X v3 A
snRNA 10X v3 B
snRNA 10X v3 A
scRNA 10X v2 A
snATAC-Seq
snRNA 10X v2 A
f
Biological feature
Sample
Technology/ Platform
Gene expression (RNA-seq)
Epigenome
Cells
Nuclei
scRNA SMART
scRNA 10x v3 A
scRNA 10x v2 A
snRNA SMART
snRNA 10x v3 B
snRNA 10x v3 A
snRNA 10x v2 A
Open chromatin (snATAC-seq)
DNA methylation (snmC-seq)
Major cell class
Excitatory
Inhibitory
Non-neuronal
Cells
g
Neuroscience Multi-Omic (NeMO) portal and analytics (https://nemoarchive.org, https://nemoanalytics.org)
h
Brainome epigenomics portal (https://brainome.ucsd.edu/BICCN_MOp)

Figure S2: Cluster membership and gene expression consistency across sc/snRNA-Seq datasets. a, Pearson correlation of gene expression of 3,792 cell type-specific marker genes across cell types between every pair of datasets. Each violin plot shows the distribution of correlation values for all genes between a pair of datasets. Most genes have highly conserved gene expression patterns at cell type level among all datasets (average correlation 0.856 across all pairs of comparisons). The most consistent datasets are scRNA 10x v2 and v3 (average correlation 0.95), while snRNA 10x v3 B is also highly similar to both scRNA 10x v2 and v3 datasets. Overall, we found the differences between single cell and single nucleus datasets to be more significant than SMART-seq versus 10x platform differences. **b,** Number of genes detected per cell or nucleus by each transcriptomic assay as a function of sequencing depth, as determined by down-sampling analysis (n=79 independent biological samples, see Table S1). **c,** Gene detection frequency (sensitivity) at each gene expression range for each dataset (n=79 independent biological samples, see Table S1). Expression of all genes in each cell type was binned based on the average logCPM in scRNA 10x v2 and snRNA 10x v3 B datasets. Single cell datasets overall have higher sensitivity for gene expression than single nucleus datasets, with the exception of snRNA 10x v3 B dataset, which was more sensitive than scRNA 10x v2 A dataset. For weakly expressed genes, the gene detection frequency can vary dramatically between datasets. For these genes, scRNA SMART was the most sensitive, followed by 10x v3 datasets, all of which showed very robust gene detection. Note that sequencing depth was not considered for this analysis. For (b,c): Box-and-whisker plots show the median, inter-quartile range (IQR: 25-75th percentile), and whiskers show the smaller of the data range (min-max), or 1.5 times the IQR. **d,** Comparisons between clustering analysis of individual datasets with the consensus clusters derived from seven transcriptome datasets. The size of the dot indicates the number of overlapping cells, and the color of the dot indicates the Jaccard index (number of cells in intersection/number of cells in union) between the independent and joint clusters. **e,** Comparison of the relative gene expression of marker genes across all cell types between corresponding SMART-seq and 10x v2 datasets. To compare gene expression directly between SMART-seq and 10x datasets, which differ in experimental platforms, gene expression quantification software and gene annotation reference, for each gene, we normalized the average log2(CPM+1) values at the cluster level in the range [0,1] by subtracting the minimum value and then dividing by the maximum value for that gene. The smooth scatter plot corresponds to the normalized gene expression for all marker genes across all types in two datasets, with their overall Pearson correlation (across all marker genes and cell types) highlighted. **f,** Differential enrichment of transcripts in single cells (x-axis) vs. single nuclei (y-axis) across four platforms. Non-coding RNAs such as *Malat1* are enriched in nuclei. **g,** Distribution of the estimated nuclear localization fraction for all mRNAs based on comparison of the sn/scRNA 10x v2 datasets[40]. To calibrate the differences among cell types, we sampled the same number of cells in each cluster for both datasets, and aggregated all the cells for estimation. We plot the empirical cumulative density function for the marker genes and all other genes separately. The fraction of nuclear mRNAs for five selected genes are shown along the X axis. As expected, mitochondrial genes such as *mt-Nd3* have almost no nuclear localization, while *Vip* is significantly enriched in the nucleus. A selected set of 3,792 cell type-specific marker genes (see Methods section “Marker gene selection”) have lower nuclear fraction relative to the other genes (median 16.6%, compared with 21.9% for non-marker genes). **h,** Cluster resolution analysis, showing the number of clusters identified in each transcriptomic dataset with a fixed cluster procedure and resolution (r=6) as a function of the number of sequenced reads, and using the same number of cells for each of the 10x or SMART-Seq datasets. Shaded region shows standard error of the mean (SEM) from cross-validation with n=5 independent data partitions.

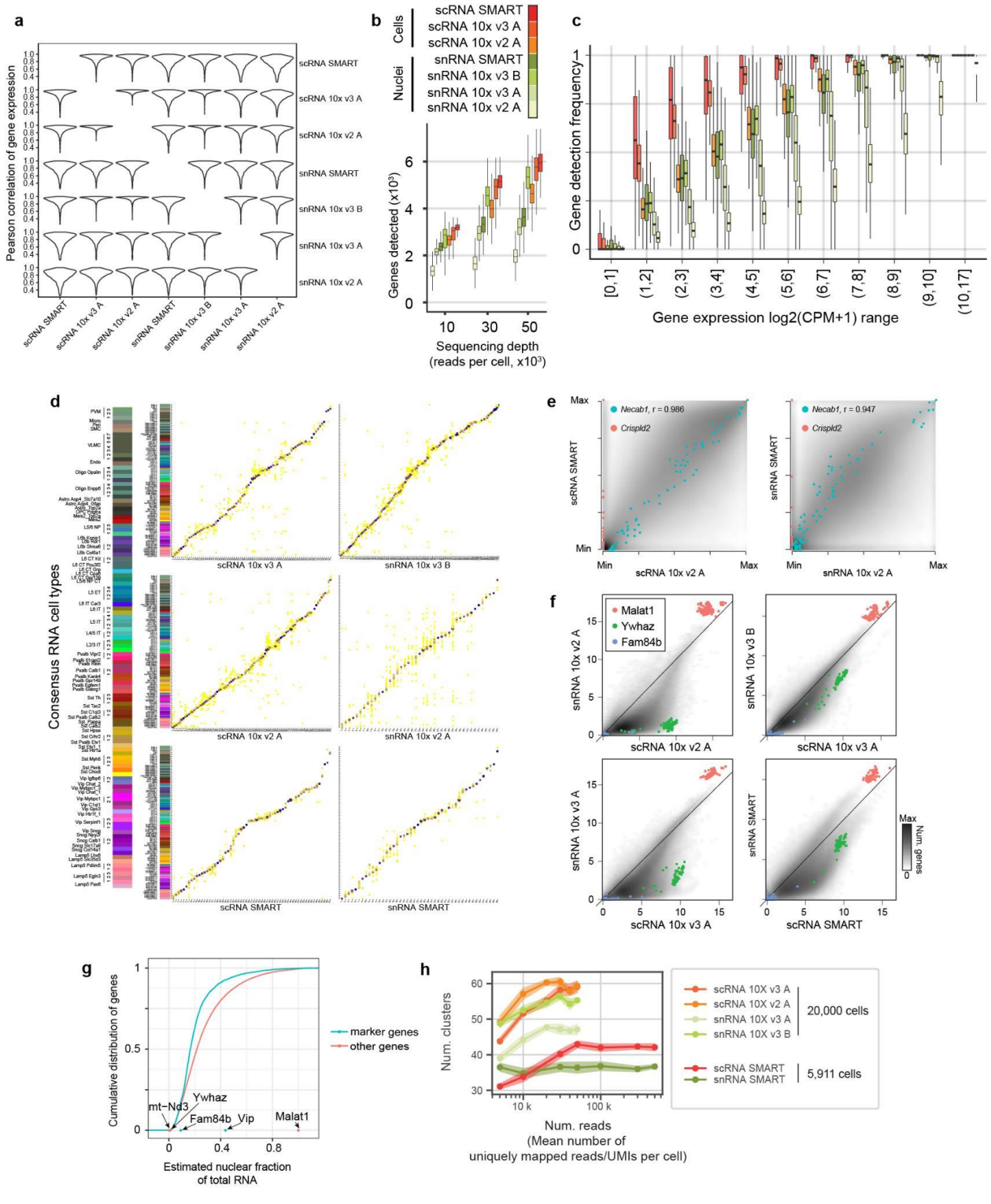

a
Pearson correlation of gene expression
scRNA SMART
scRNA 10x v3 A
scRNA 10x v2 A
snRNA SMART
snRNA 10x v3 B
snRNA 10x v3 A
snRNA 10x v2 A
b
Cells
Nuclei
Genes detected (x10³)
Sequencing depth
(reads per cell, x10³)
c
Gene detection frequency
Gene expression log2(CPM+1) range
[0,1]
(1,2]
(2,3]
(3,4]
(4,5]
(5,6]
(6,7]
(7,8]
(8,9]
(9,10]
(10,17]
d
Consensus RNA cell types
scRNA 10x v3 A
snRNA 10x v3 B
scRNA 10x v2 A
snRNA 10x v2 A
scRNA SMART
snRNA SMART
e
Necab1, r = 0.986
Crispld2
Necab1, r = 0.947
Max
Min
f
Malat1
Ywhaz
Fam84b
Num. genes
g
Cumulative distribution of genes
Estimated nuclear fraction
of total RNA
marker genes
other genes
mt-Nd3
Ywhaz
Fam84b
Vip
Malat1
h
Num. clusters
Num. reads
(Mean number of
uniquely mapped reads/UMIs per cell)
scRNA 10X v3 A
scRNA 10X v2 A
snRNA 10X v3 A
snRNA 10X v3 B
20,000 cells
scRNA SMART
snRNA SMART
5,911 cells

Figure S3: Correspondence between MOp consensus RNA-Seq cell type taxonomy and previously published VISp/ALM cell type taxonomy[4]. a, Cells from all sc/snRNA MOp datasets were mapped to the most correlated VISp/ALM cell types based on VISp/ALM cell type markers. The size of dots indicates the number of overlapping cells, and the color indicates the Jaccard index (number of cells in intersection/number of cells in union). MOp L5 ET types are mapped predominantly to L5 pyramidal tract (PT) ALM types in the VISp/ALM study. Note that we have adopted the nomenclature “extratelencephalic (ET)” for these neurons, instead of the previously used “pyramidal tract (PT),” due to the fact that not all of these neurons project to the pyramidal tract leading to spinal cord. **b,** Three L5 PT ALM types can be divided into two groups with distinct projection patterns. Cells in the pink group project to medulla and have been functionally associated with movement initiation, while the cells in the green group project to thalamus, associated with movement planning. Adapted from (Economo, et al. 2018)[37]. **c,** Enlarged view of the correspondence between MOp L5 ET types and VISp/ALM L5 PT types. Two subsets of medulla-projecting (pink) and thalamus-projecting (green) L5 PT cells are highlighted.

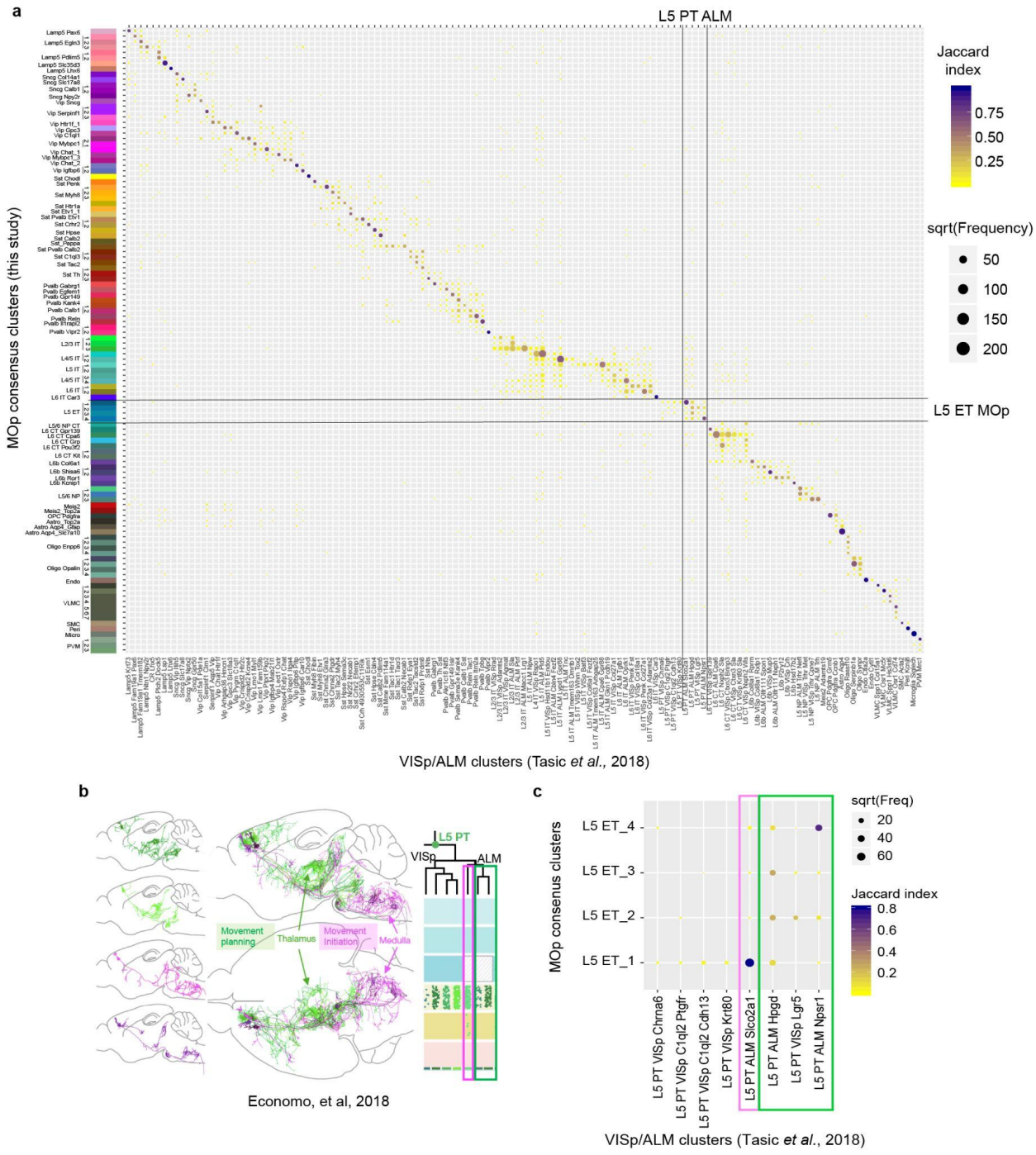

a
L5 PT ALM
Jaccard index
0.75
0.50
0.25
sqrt(Frequency)
50
100
150
200
L5 ET MOp
MOp consensus clusters (this study)
VISp/ALM clusters (Tasic et al., 2018)
b
L5 PT
VISp
ALM
Movement planning
Thalamus
Movement Initiation
Medulla
Economo, et al, 2018
c
sqrt(Freq)
20
40
60
Jaccard index
0.8
0.6
0.4
0.2
MOp consenus clusters
L5 ET_4
L5 ET_3
L5 ET_2
L5 ET_1
L5 PT VISp Chrna6
L5 PT VISp C1ql2 Ptgfr
L5 PT VISp C1ql2 Cdh13
L5 PT VISp Krt80
L5 PT ALM Slco2a1
L5 PT ALM Hpgd
L5 PT VISp Lgr5
L5 PT ALM Npsr1
VISp/ALM clusters (Tasic et al., 2018)

Figure S4: Marker genes for L5 ET cell types. a, Heatmap showing expression of a combination of marker genes of L5 PT ALM types in previously published dataset[4], and marker genes for MOp L5 ET types. The color bars on the top indicate the cell type and projection class. **b,** Heatmap for MOp L5 ET types in multiple sc/snRNA datasets using the same marker genes in the same order as in a. Cell types are divided into the pink and green groups based on correspondence in Figure S3c.

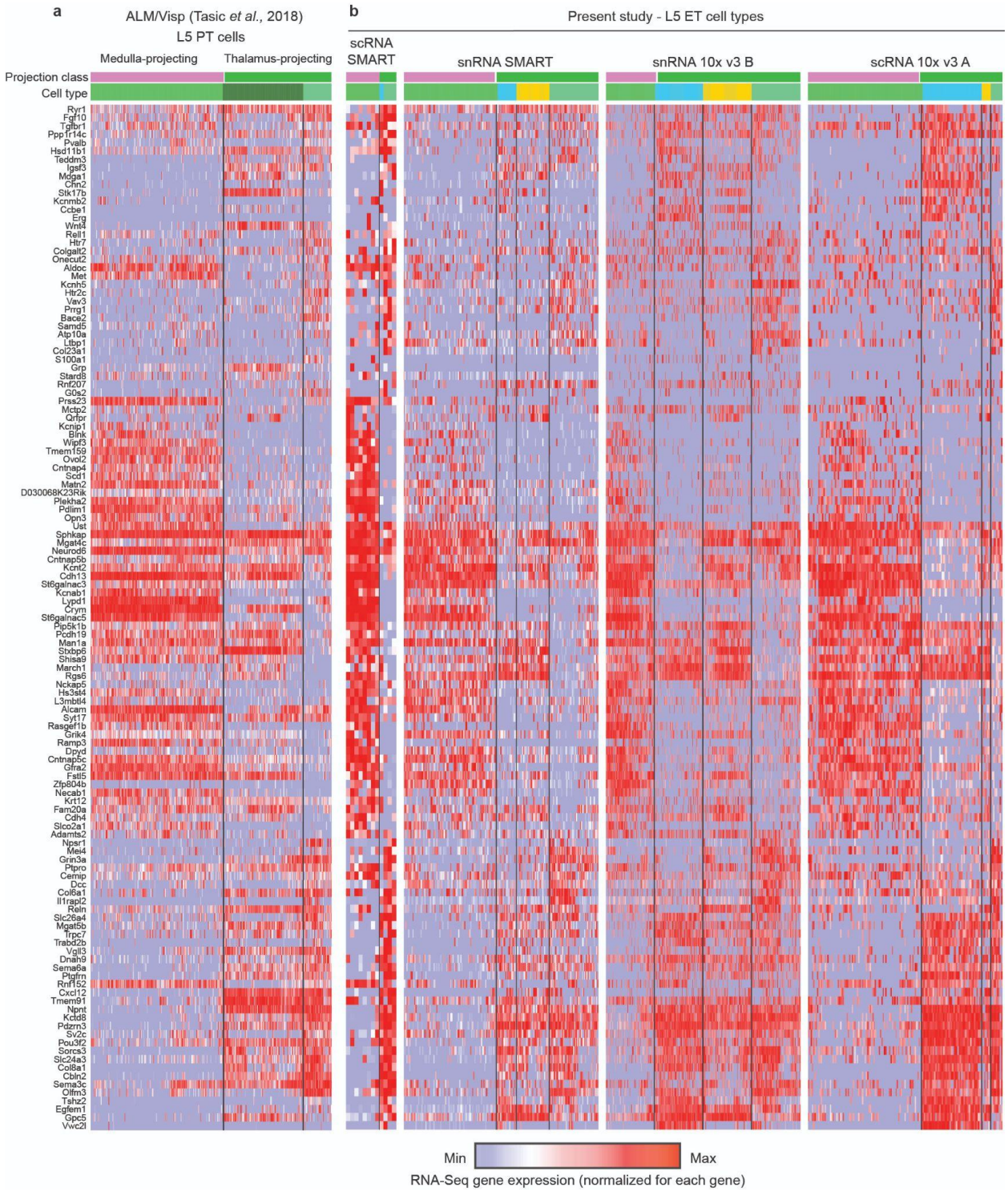

a
ALM/Visp (Tasic et al., 2018)
L5 PT cells
Medulla-projecting
Thalamus-projecting
b
Present study - L5 ET cell types
scRNA SMART
snRNA SMART
snRNA 10x v3 B
scRNA 10x v3 A
Projection class
Cell type
Min
Max
RNA-Seq gene expression (normalized for each gene)

Figure S5: Marker genes for L4/5 IT and L5 IT cell types. a, Heatmap of marker genes for MOp L4/5 IT and L5 IT types in multiple sc/snRNA datasets. **b**, *In situ* hybridization (ISH) showing validation of layer 4 marker genes (*Rspo1, Rorb*) and layer 5 (*Fezf2*) in mouse MOp. Note that *Rorb* labels both L4 and a subset of L5 neurons.

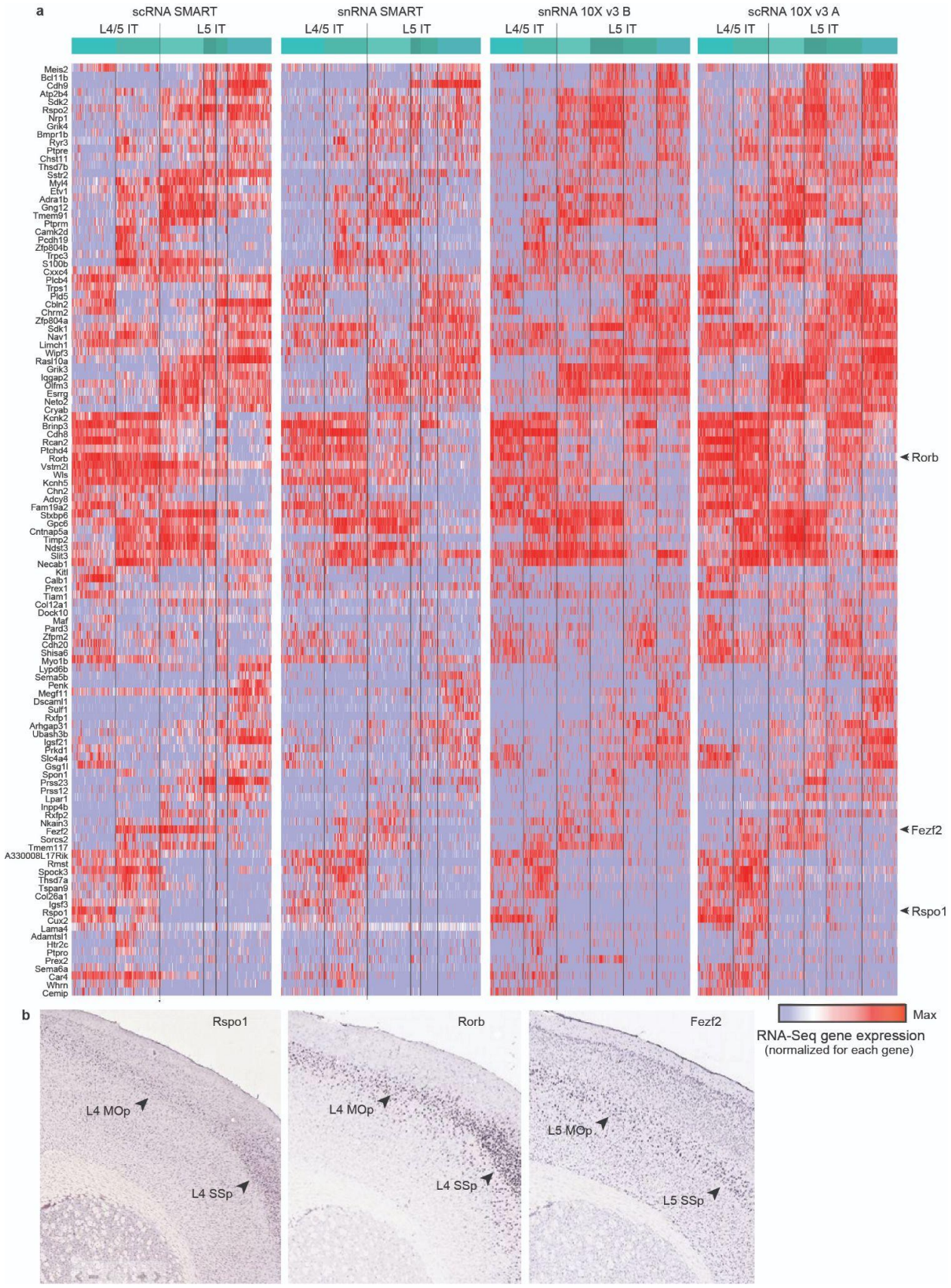

a
scRNA SMART
snRNA SMART
snRNA 10X v3 B
scRNA 10X v3 A
L4/5 IT
L5 IT
Rorb
Fezf2
Rspo1
b
Rspo1
Rorb
Fezf2
L4 MOp
L4 SSp
L5 MOp
L5 SSp
Max
RNA-Seq gene expression
(normalized for each gene)

Figure S6: Epigenomic cell types and multimodal integration. a, Cell type clusters from single-nucleus methyl-C-Seq (snmC-Seq2[27,32]) for 9,876 MOp nuclei are represented in a two-dimensional projection. Labels indicate broad cell types, colors show finest cluster resolution. **b,** Non-CG DNA methylation level (normalized mCH) for each cell at gene bodies of markers of major cell types. Actively expressed genes have low mCH, indicated by colored bars extending downward. Highly methylated (repressed) genes appear white in this plot. **c,** Two-dimensional projection of cell type clusters from single-nucleus ATAC-Seq (snATAC-Seq[29]) profiles for 81,196 cells. **d,** Gene body chromatin accessibility (total snATAC-Seq read density, log(CPM+1)) for marker genes. For b and d, each bar represents one cell. Cell type abbreviations as in Figure 3. CGE/MGE - Caudal/Medial ganglionic eminence derived inhibitory cells. **e,f,** Integrated, multimodal UMAP embeddings (e: SingleCellFusion; f: LIGER) colored by the clusters assigned in separate analysis of each dataset. Each panel shows the cells from a single dataset. **g,** Integrated analysis of major cell classes by LIGER. Cells in each of 5 cell classes are separately integrated, illustrating fine-grained resolution of integrated data. **h**, Number of cells in each of 56 multimodality cell types (SingleCellFusion; L2), ranked by cluster size. **i,j,** The number of cells for 56 integrated clusters (i: SingleCellFusion L2; j: LIGER L2), as well as the corresponding coarser clusters (L1, L0). Cluster order and color scheme are the same as shown in Figure 3.

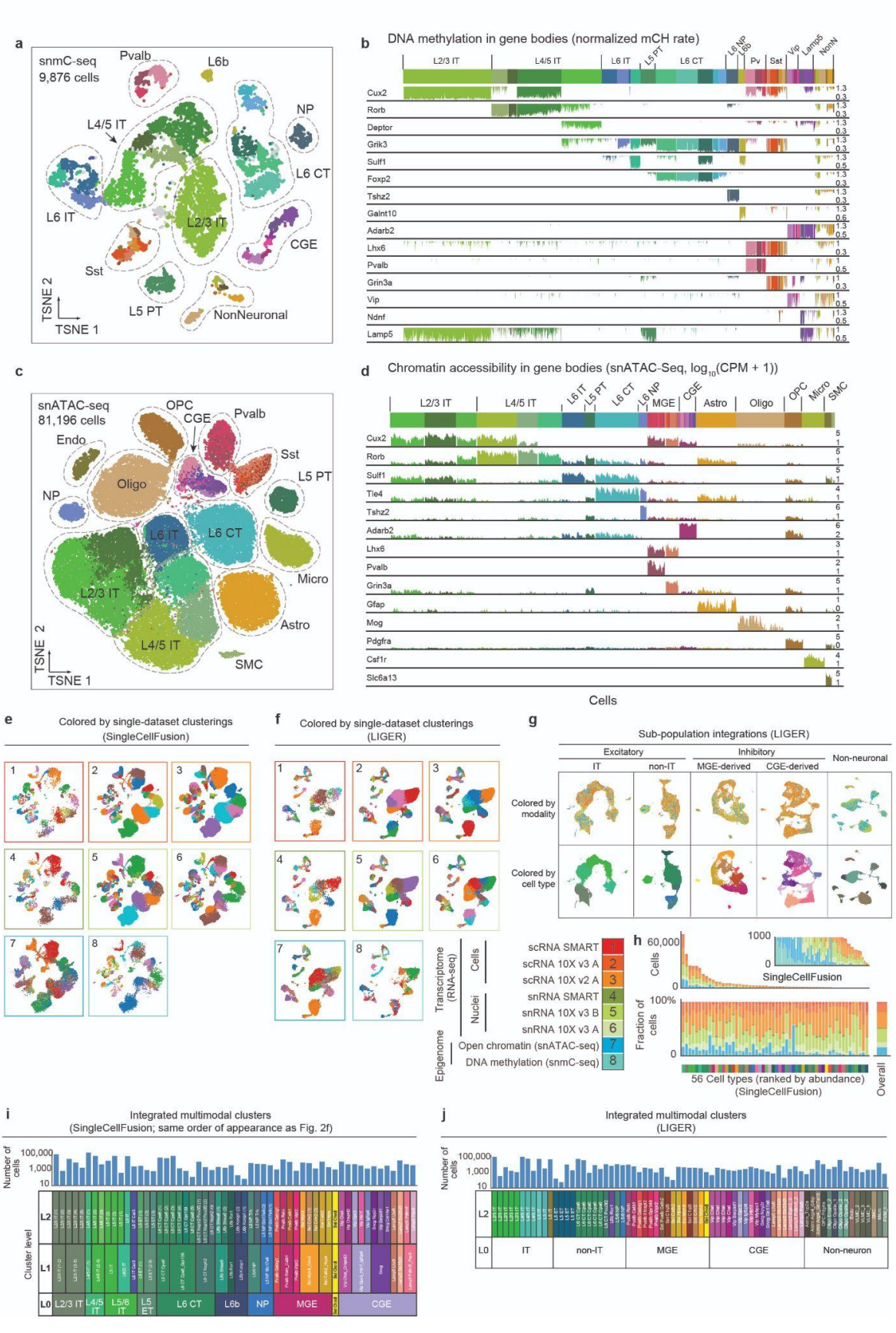

a
snmC-seq
9,876 cells
Pvalb
L6b
NP
L4/5 IT
L6 CT
L6 IT
L2/3 IT
CGE
Sst
L5 PT
NonNeuronal
TSNE 2
TSNE 1
b
DNA methylation in gene bodies (normalized mCH rate)
L2/3 IT
L4/5 IT
L6 IT
L5 PT
L6 CT
L6 NP
L6b
Pv
Sst
Vip
Lamp5
NonN
Cux2
Rorb
Deptor
Grik3
Sulf1
Foxp2
Tshz2
Galnt10
Adarb2
Lhx6
Pvalb
Grin3a
Vip
Ndnf
Lamp5
c
snATAC-seq
81,196 cells
OPC
CGE
Pvalb
Endo
Sst
L5 PT
Oligo
NP
L6 IT
L6 CT
Micro
L2/3 IT
Astro
L4/5 IT
SMC
TSNE 2
TSNE 1
d
Chromatin accessibility in gene bodies (snATAC-Seq, log10(CPM + 1))
L2/3 IT
L4/5 IT
L6 IT
L5 PT
L6 CT
L6 NP
MGE
CGE
Astro
Oligo
OPC
Micro
SMC
Cux2
Rorb
Sulf1
Tle4
Tshz2
Adarb2
Lhx6
Pvalb
Grin3a
Gfap
Mog
Pdgfra
Csf1r
Slc6a13
Cells
e
Colored by single-dataset clusterings
(SingleCellFusion)
f
Colored by single-dataset clusterings
(LIGER)
g
Sub-population integrations (LIGER)
Excitatory
Inhibitory
Non-neuronal
IT
non-IT
MGE-derived
CGE-derived
Colored by modality
Colored by cell type
scRNA SMART
scRNA 10X v3 A
scRNA 10X v2 A
snRNA SMART
snRNA 10X v3 B
snRNA 10X v3 A
Open chromatin (snATAC-seq)
DNA methylation (snmC-seq)
Transcriptome (RNA-seq)
Cells
Nuclei
Epigenome
h
60,000
Cells
0
1000
0
SingleCellFusion
100%
Fraction of cells
0
56 Cell types (ranked by abundance)
(SingleCellFusion)
Overall
i
Integrated multimodal clusters
(SingleCellFusion; same order of appearance as Fig. 2f)
Number of cells
100,000
1,000
10
Cluster level
L2
L1
L0
L2/3 IT
L4/5 IT
L5/6 IT
L5 ET
L6 CT
L6b
NP
MGE
CGE
j
Integrated multimodal clusters
(LIGER)
Number of cells
100,000
1,000
10
L2
L0
IT
non-IT
MGE
CGE
Non-neuron

Figure S7: Validation of multimodal integration of transcriptomic and epigenomic data. a, Confusion matrices comparing integrated clusters generated by SingleCellFusion versus clusters generated by LIGER (left), and comparing SingleCellFusion versus consensus transcriptomic taxonomy (right). **b**, Confusion matrix comparing integrated clusters (SingleCellFusion L2) with single-modality clustering for every dataset. **c,d,** Agreement and alignment metrics[21] characterize the fidelity of the joint low-dimensional embedding for LIGER and SingleCellFusion. Agreement measures the fraction of k-nearest neighbors for each dataset are still nearest neighbors in the low-dimensional embedding. A high value of the agreement metric thus indicates preservation of each dataset's internal structure in the joint embedding. Alignment measures the mixing of datasets in the joint low-dimensional space, and is a normalized measure of the mean number of k-nearest neighbors that come from each of the datasets. **e,** Embedding of multimodality cluster centroids. Black dots are cluster centroids of integrated clusters (SingleCellFusion), colored dots are cluster centroids of individual datasets. **f,** Molecular signatures at the gene body of *Lhx9*, a developmentally expressed transcription factor, across cell types (n=29; SingleCellFusion L1). We found enrichment of mCG and mCH in L6b neurons with no corresponding RNA or ATAC-Seq signal. **g**, Spearman correlation matrix for cluster centroid gene expression (measured or imputed) across major cell subclasses for each dataset (SingleCellFusion L0). **h**, Correlation for subsets of inhibitory (CGE, MGE) and excitatory (L4/5 IT, L2/3 IT) neuron types using fine-grained integrated clusters (SingleCellFusion L2).

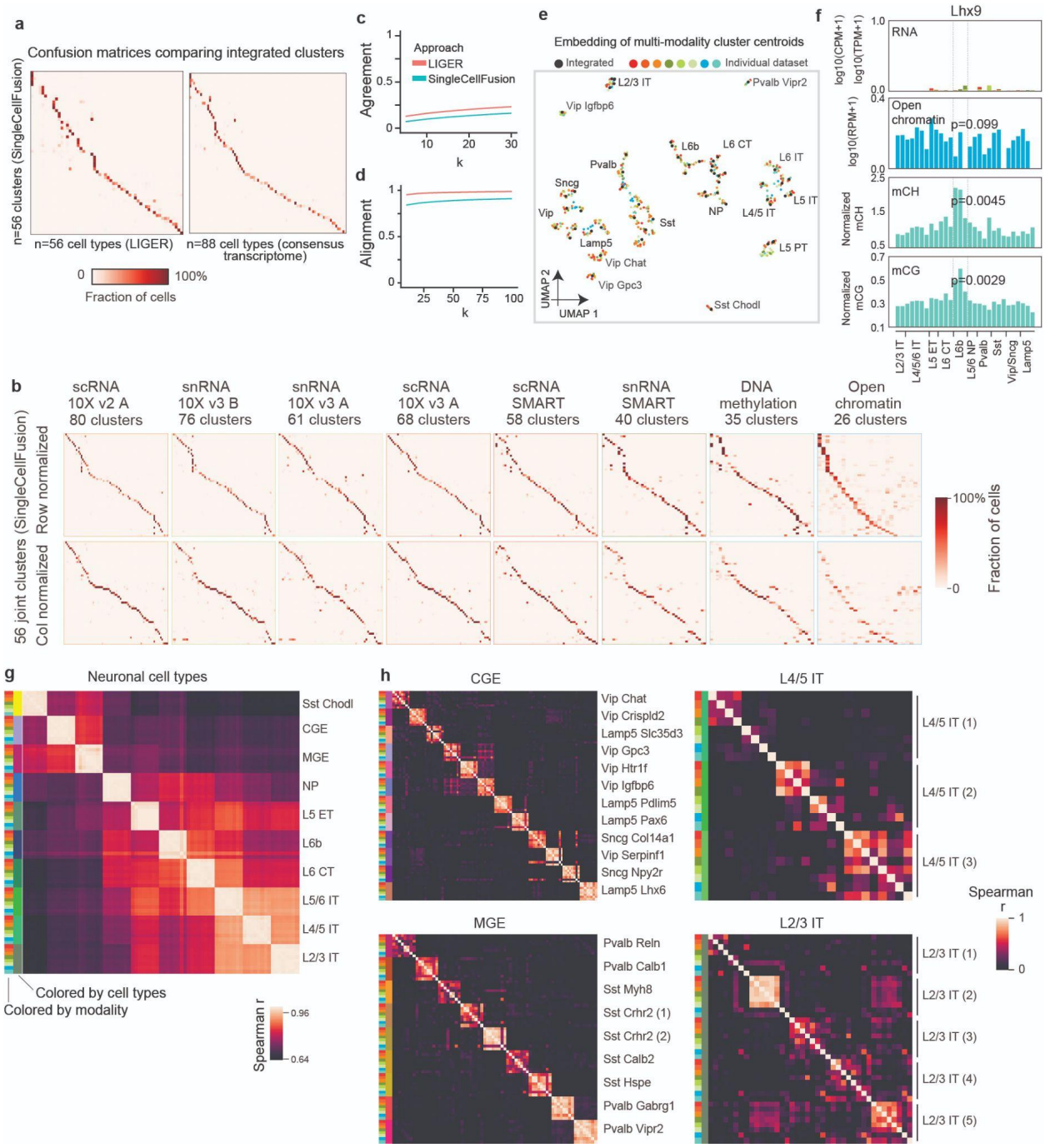
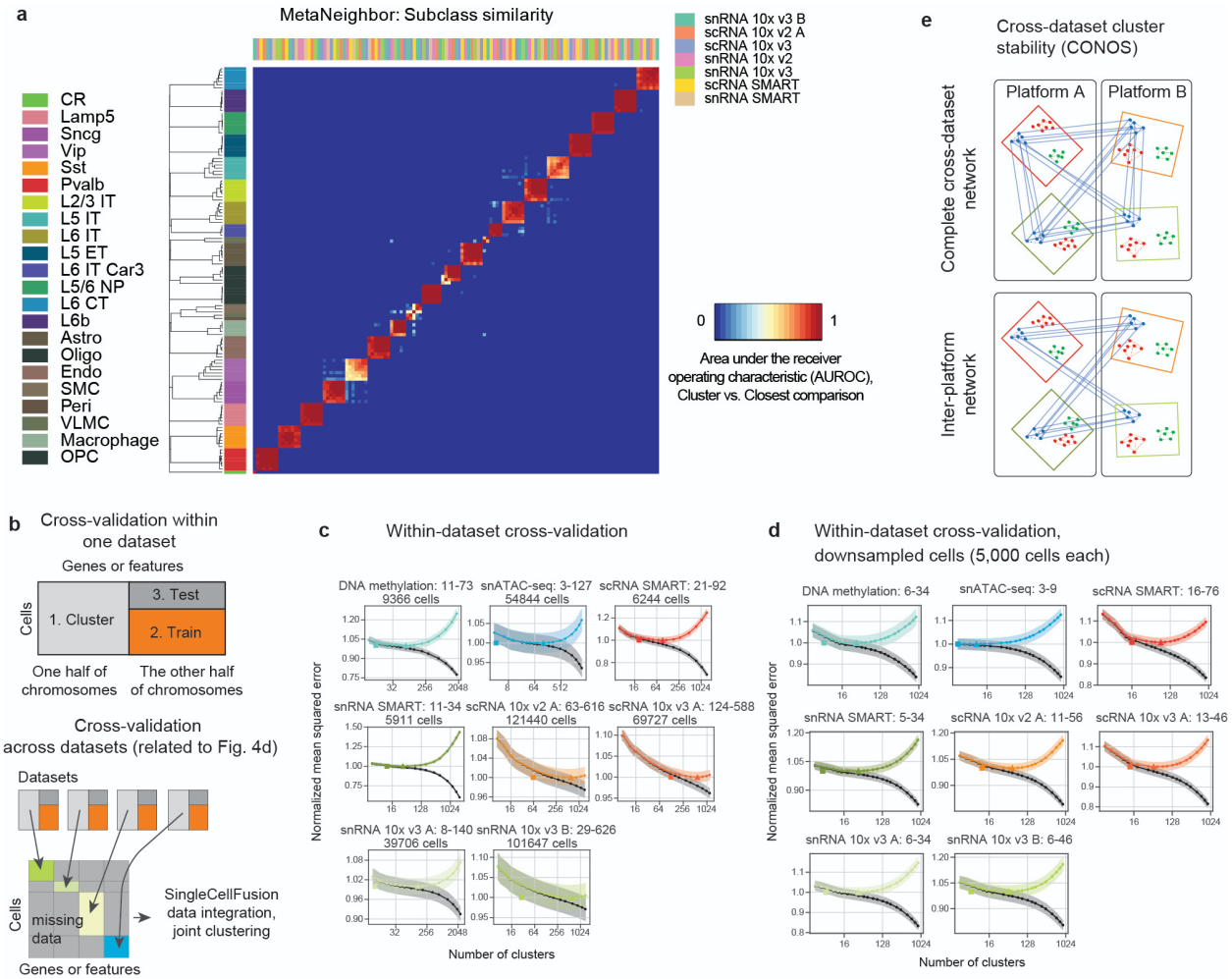

a
Confusion matrices comparing integrated clusters
n=56 clusters (SingleCellFusion)
n=56 cell types (LIGER)
n=88 cell types (consensus transcriptome)
0
100%
Fraction of cells
c
Agreement
Approach
LIGER
SingleCellFusion
k
d
Alignment
k
e
Embedding of multi-modality cluster centroids
Integrated
Individual dataset
L2/3 IT
Pvalb Vipr2
Vip Igfbp6
L6b
L6 CT
L6 IT
Pvalb
Sncg
Vip
Sst
NP
L4/5 IT
L5 IT
Lamp5
L5 PT
Vip Chat
Vip Gpc3
Sst Chodl
UMAP 2
UMAP 1
f
Lhx9
log10(CPM+1)
RNA
log10(TPM+1)
Open chromatin
p=0.099
log10(RPM+1)
mCH
p=0.0045
Normalized mCH
mCG
p=0.0029
Normalized mCG
L2/3 IT
L4/5/6 IT
L5 ET
L6 CT
L6b
L5/6 NP
Pvalb
Sst
Vip/Sncg
Lamp5
b
scRNA 10X v2 A 80 clusters
snRNA 10X v3 B 76 clusters
snRNA 10X v3 A 61 clusters
scRNA 10X v3 A 68 clusters
scRNA SMART 58 clusters
snRNA SMART 40 clusters
DNA methylation 35 clusters
Open chromatin 26 clusters
56 joint clusters (SingleCellFusion)
Row normalized
Col normalized
100%
0
Fraction of cells
g
Neuronal cell types
Sst Chodl
CGE
MGE
NP
L5 ET
L6b
L6 CT
L5/6 IT
L4/5 IT
L2/3 IT
Colored by cell types
Colored by modality
Spearman r
0.96
0.64
h
CGE
Vip Chat
Vip Crispld2
Lamp5 Slc35d3
Vip Gpc3
Vip Htr1f
Vip Igfbp6
Lamp5 Pdlim5
Lamp5 Pax6
Sncg Col14a1
Vip Serpinf1
Sncg Npy2r
Lamp5 Lhx6
L4/5 IT
L4/5 IT (1)
L4/5 IT (2)
L4/5 IT (3)
Spearman r
1
0
MGE
Pvalb Reln
Pvalb Calb1
Sst Myh8
Sst Crhr2 (1)
Sst Crhr2 (2)
Sst Calb2
Sst Hspe
Pvalb Gabrg1
Pvalb Vipr2
L2/3 IT
L2/3 IT (1)
L2/3 IT (2)
L2/3 IT (3)
L2/3 IT (4)
L2/3 IT (5)

Figure S8: MetaNeighbor and cross-validation analysis of cluster reproducibility. a, Heatmap showing replicability scores (MetaNeighbor AUROC) at the subclass level of the independent clusterings of seven RNA-Seq datasets. High AUROC indicates that the cell type labels in one dataset can be reliably predicted based on the nearest neighbors of those cells in another dataset, together with the independent cluster analysis of that dataset. **b**, Scheme for within- and across-dataset cross-validation. **c,d** Within-dataset cross-validation analysis for each dataset, either using the full set of cells (c) or using a random sample of 5000 cells (d)**.** In each plot, the black curve shows training error while the colored U-shaped curve shows the test set error, with a minimum at the cluster resolution that balances over- and under-fitting. Shaded region shows standard error of the mean based on cross-validation with n=5 data partitions. **e,** Transcriptomic platform consistency is assessed by cross-dataset cluster stability analysis (Conos[52]).


a
MetaNeighbor: Subclass similarity
CR
Lamp5
Sncg
Vip
Sst
Pvalb
L2/3 IT
L5 IT
L6 IT
L5 ET
L6 IT Car3
L5/6 NP
L6 CT
L6b
Astro
Oligo
Endo
SMC
Peri
VLMC
Macrophage
OPC
snRNA 10x v3 B
scRNA 10x v2 A
scRNA 10x v3
snRNA 10x v2
snRNA 10x v3
scRNA SMART
snRNA SMART
0
1
Area under the receiver operating characteristic (AUROC), Cluster vs. Closest comparison
e
Cross-dataset cluster stability (CONOS)
Platform A
Platform B
Complete cross-dataset network
Inter-platform network
b
Cross-validation within one dataset
Genes or features
Cells
1. Cluster
3. Test
2. Train
One half of chromosomes
The other half of chromosomes
Cross-validation across datasets (related to Fig. 4d)
Datasets
Cells
missing data
Genes or features
SingleCellFusion data integration, joint clustering
c
Within-dataset cross-validation
DNA methylation: 11-73 9366 cells
snATAC-seq: 3-127 54844 cells
scRNA SMART: 21-92 6244 cells
snRNA SMART: 11-34 5911 cells
scRNA 10x v2 A: 63-616 121440 cells
scRNA 10x v3 A: 124-588 69727 cells
snRNA 10x v3 A: 8-140 39706 cells
snRNA 10x v3 B: 29-626 101647 cells
Normalized mean squared error
Number of clusters
d
Within-dataset cross-validation, downsampled cells (5,000 cells each)
DNA methylation: 6-34
snATAC-seq: 3-9
scRNA SMART: 16-76
snRNA SMART: 5-34
scRNA 10x v2 A: 11-56
scRNA 10x v3 A: 13-46
snRNA 10x v3 A: 6-34
snRNA 10x v3 B: 6-46
Normalized mean squared error
Number of clusters

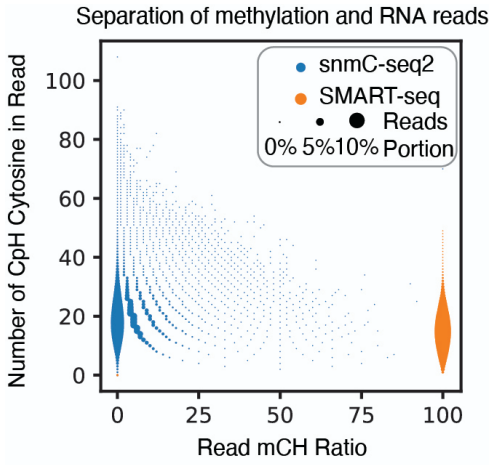


Figure S9: The specificity for classifying methylation (snmC-seq2) and transcriptome (snRNA-seq) reads plotted as a function of the number of CpH cytosine in the reads.

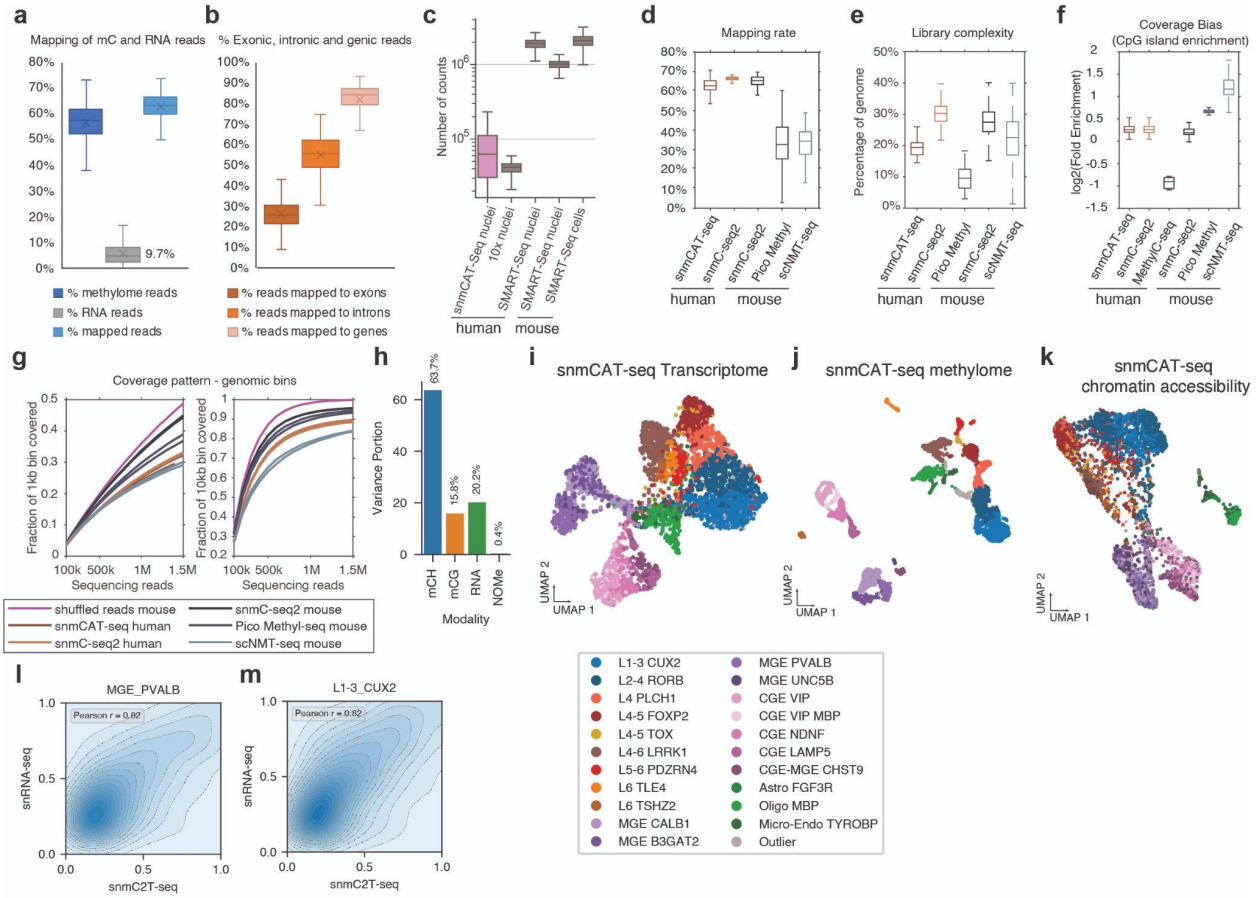


Figure S10: snmCAT-seq generates single-nucleus multi-omic profiles from human brain tissues. a. The fraction of total snmCAT-seq reads derived from methylome and transcriptome. **b.** The fraction of snmCAT-seq transcriptome reads mapped to exons, introns or gene bodies. **c.** Boxplot comparing the number of reads detected in each cell/nucleus by different single-cell or single-nucleus RNA-seq technologies. **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.** The portion of variance explained by each data modality. **i-k.** UMAP embedding of 4253 snmCAT-seq cells using single modality information: transcriptome (**i**), methylome (mCH and mCG, **j**) and chromatin accessibility (**k**). **l-m.** Pearson correlation of gene expression quantified by snmCAT-seq transcriptome and snRNA-seq in MGE PVALB (**l**) and L1-3 CUX2 (**m**) cells.

Figure S11: 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. **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.** Integration 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 integration method, from top to bottom the first panel shows mC and RNA modalities on the joint UMAP embedding after integration. The second panel shows the normalized self-radius. The third and fourth panels show the co-cluster level cell composition and clustering accuracy.

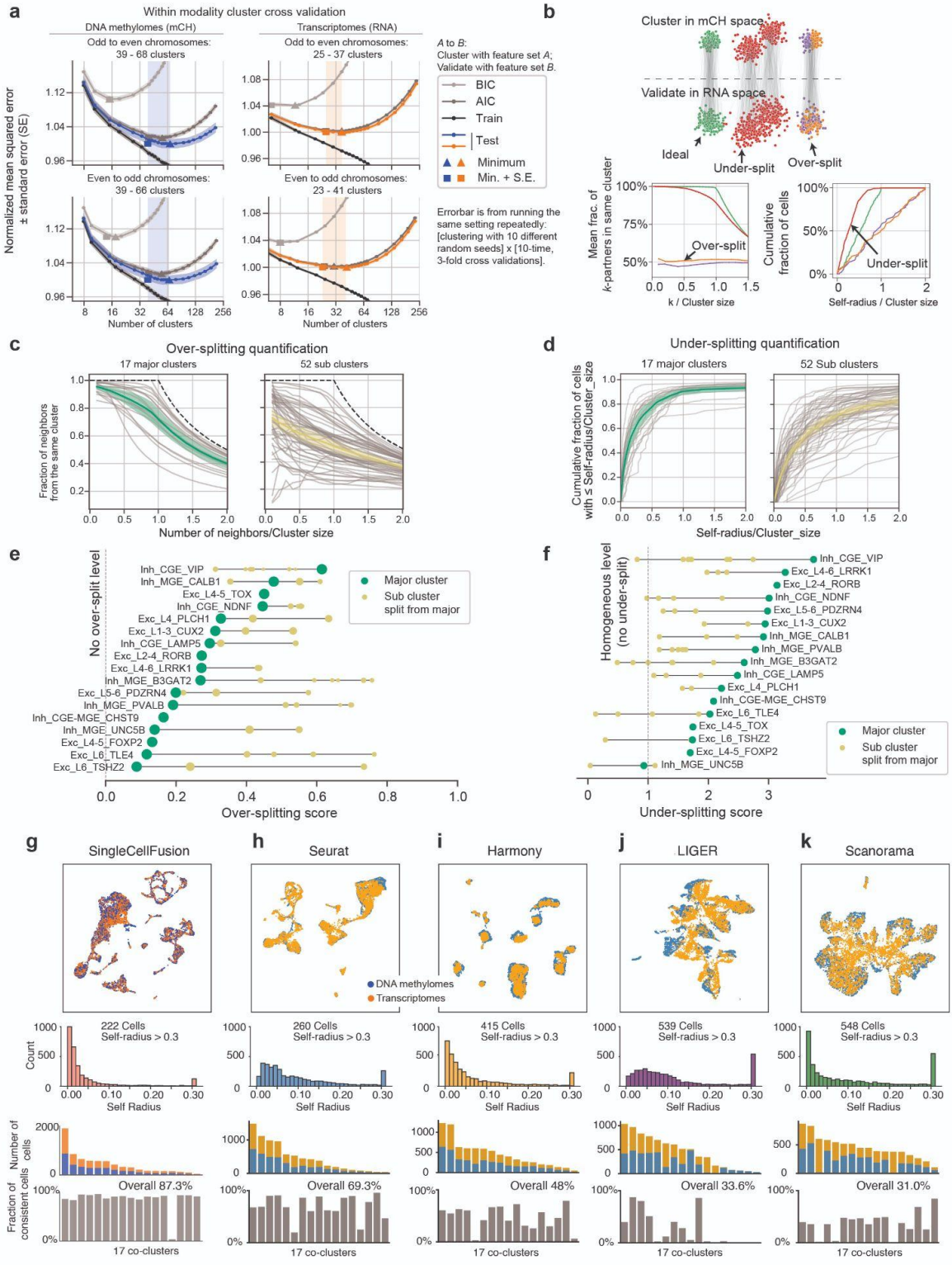
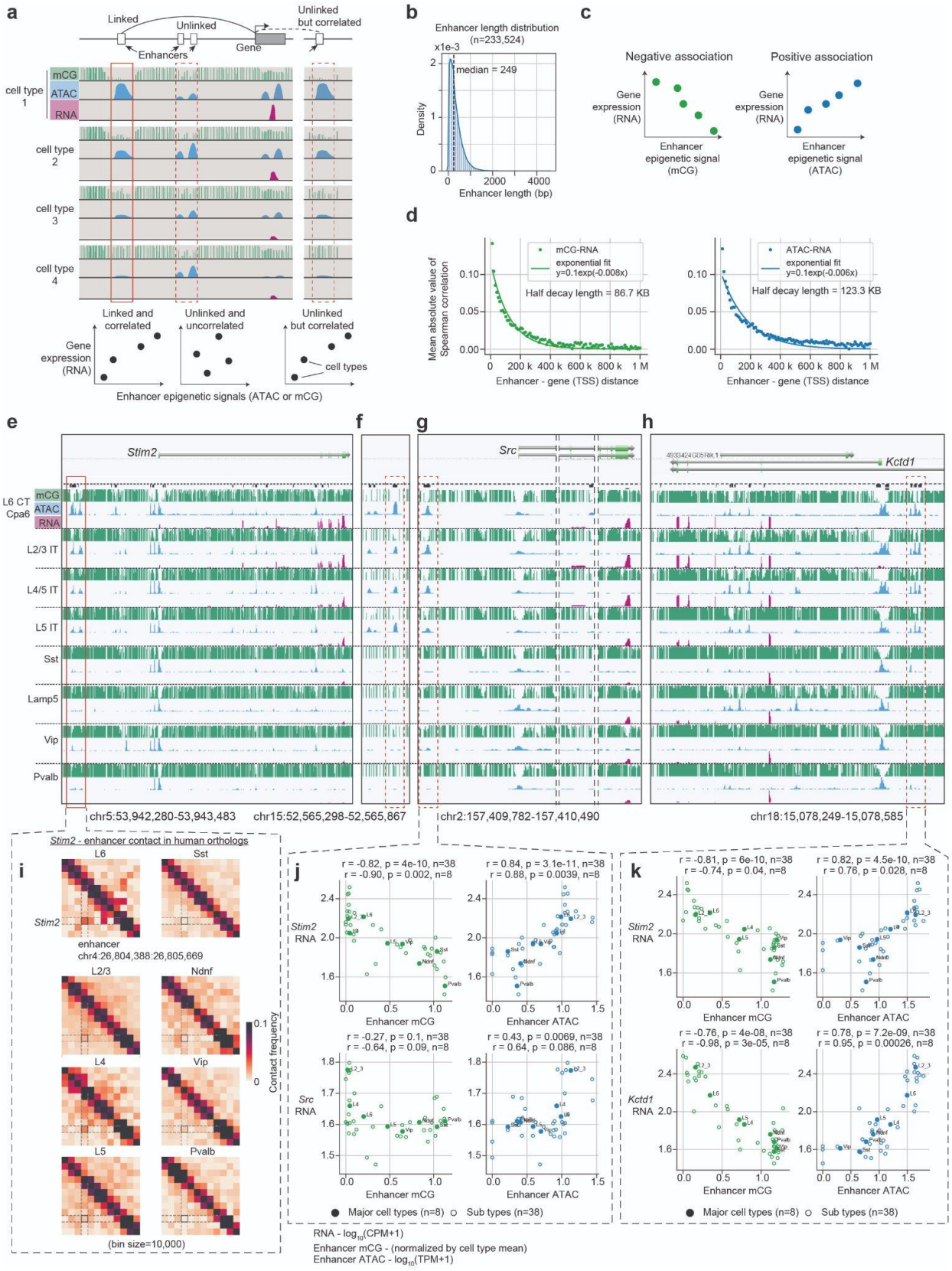

a
Within modality cluster cross validation
DNA methylomes (mCH)
Transcriptomes (RNA)
Odd to even chromosomes:
39 - 68 clusters
25 - 37 clusters
Even to odd chromosomes:
39 - 66 clusters
23 - 41 clusters
Normalized mean squared error
± standard error (SE)
Number of clusters
A to B:
Cluster with feature 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

Figure S12: Examples of enhancer-gene links. a. Approach for linking enhancers to target gene(s) by correlating enhancer activities and gene expression across cell types. Statistically significant correlation alone may not distinguish genuine vs. spurious links. **b.** Distribution of putative enhancer length (list adapted from Ref[10]; see Methods). **c.** Illustration of two modes of enhancer-gene associations: enhancer mCG typically have positive correlation with gene expression, while enhancer ATAC-seq signals typically have negative correlation. **d.** Median Spearman correlation as a function of enhancer-TSS distance. Both mCG-RNA and ATAC-RNA decay exponentially, with a half decay length of 86.7 kb and 123.3 kb, respectively. **e-h.** Genome browser views across cell types and data modalities near the gene *Stim2* (**e**), as well as other regions (**f-h**) with strongly correlated enhancer signals. Note that the highlighted enhancers in (**g-h**) are also correlated with the expression of their nearby genes (*Src* and *Kctd1*). **i.** Heatmaps of chromatin contact frequency in human brain cells near *Stim2* and the human ortholog of the highlighted enhancer across 8 human neuronal cell types. **j-k.** Scatter plot of *Stim2* expression (upper row) / local genes expression (lower row) versus the highlighted enhancers. Enhancer mCG level is normalized by the global mean mCG level of each cell type; RNA is log10(CPM+1) normalized; ATAC is log10(TPM+1) normalized.


a
Linked
Unlinked
Unlinked but correlated
Enhancers
Gene
mCG
ATAC
RNA
cell type 1
cell type 2
cell type 3
cell type 4
Linked and correlated
Unlinked and uncorrelated
Unlinked but correlated
Gene expression (RNA)
cell types
Enhancer epigenetic signals (ATAC or mCG)
b
Enhancer length distribution (n=233,524)
x1e-3
median = 249
Density
Enhancer length (bp)
c
Negative association
Positive association
Gene expression (RNA)
Enhancer epigenetic signal (mCG)
Enhancer epigenetic signal (ATAC)
d
mCG-RNA
exponential fit
y=0.1exp(-0.008x)
Half decay length = 86.7 KB
ATAC-RNA
exponential fit
y=0.1exp(-0.006x)
Half decay length = 123.3 KB
Mean absolute value of Spearman correlation
Enhancer - gene (TSS) distance
e
Stim2
f
g
Src
h
Kctd1
L6 CT Cpa6
L2/3 IT
L4/5 IT
L5 IT
Sst
Lamp5
Vip
Pvalb
chr5:53,942,280-53,943,483
chr15:52,565,298-52,565,867
chr2:157,409,782-157,410,490
chr18:15,078,249-15,078,585
i
Stim2 - enhancer contact in human orthologs
L6
Sst
Stim2
enhancer
chr4:26,804,388:26,805,669
L2/3
Ndnf
L4
Vip
L5
Pvalb
(bin size=10,000)
Contact frequency
j
r = -0.82, p = 4e-10, n=38
r = -0.90, p = 0.002, n=8
r = 0.84, p = 3.1e-11, n=38
r = 0.88, p = 0.0039, n=8
Stim2 RNA
Enhancer mCG
Enhancer ATAC
r = -0.27, p = 0.1, n=38
r = -0.64, p = 0.09, n=8
r = 0.43, p = 0.0069, n=38
r = 0.64, p = 0.086, n=8
Src RNA
Major cell types (n=8)
Sub types (n=38)
k
r = -0.81, p = 6e-10, n=38
r = -0.74, p = 0.04, n=8
r = 0.82, p = 4.5e-10, n=38
r = 0.76, p = 0.028, n=8
Stim2 RNA
Enhancer mCG
Enhancer ATAC
r = -0.76, p = 4e-08, n=38
r = -0.98, p = 3e-05, n=8
r = 0.78, p = 7.2e-09, n=38
r = 0.95, p = 0.00026, n=8
Kctd1 RNA
Major cell types (n=8)
Sub types (n=38)
RNA - log10(CPM+1)
Enhancer mCG - (normalized by cell type mean)
Enhancer ATAC - log10(TPM+1)

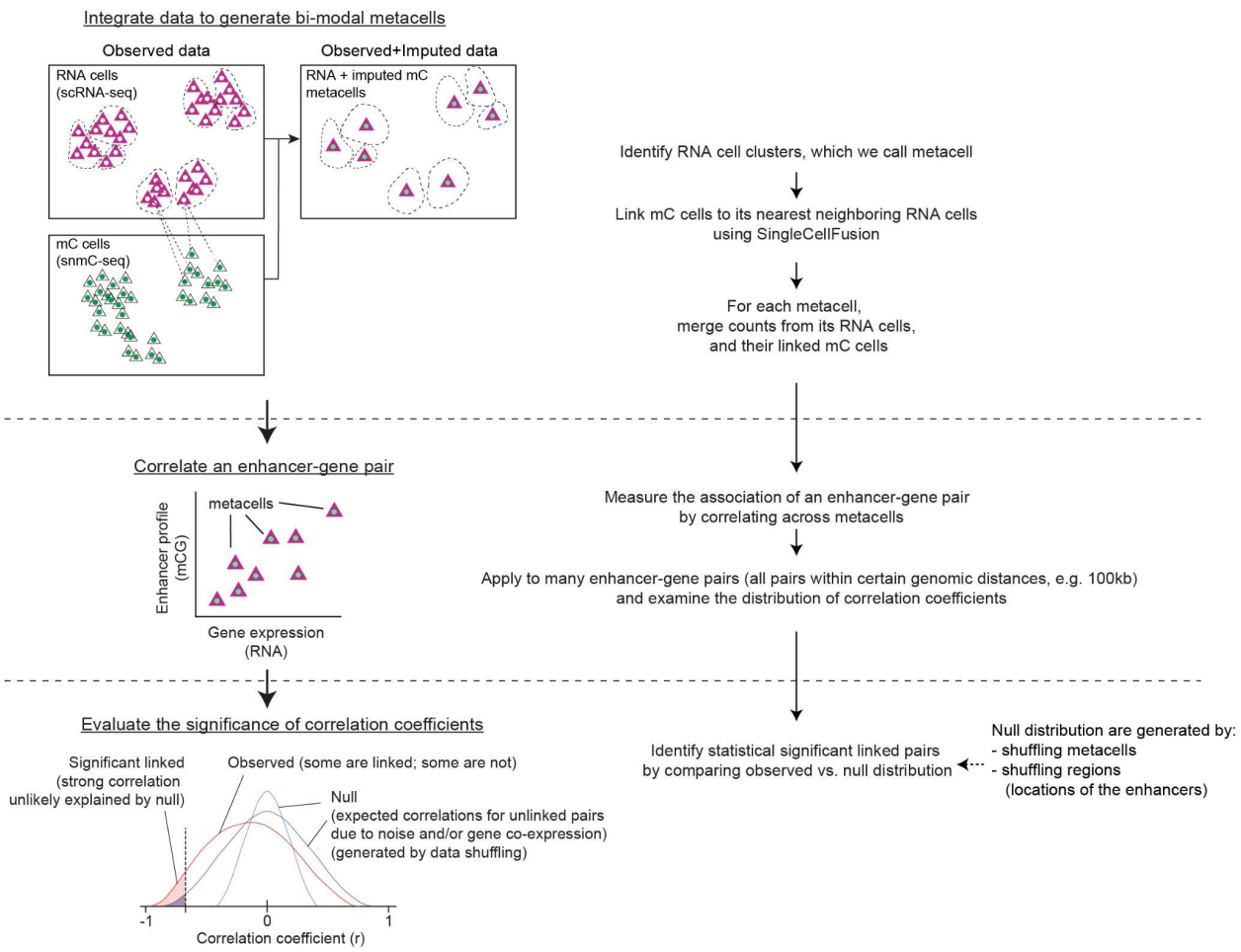


Figure S13: Method overview. The analysis involves three main steps. 1. Integrate transcriptomic and epigenomic data to generate metacells with bi-modal profiles. 2. Correlate enhancer-gene pairs to get correlation coefficients for individual enhancer-gene pairs. 3. Evaluate the statistical significance of correlations by comparing the observed correlations with null distributions generated by data shuffling.

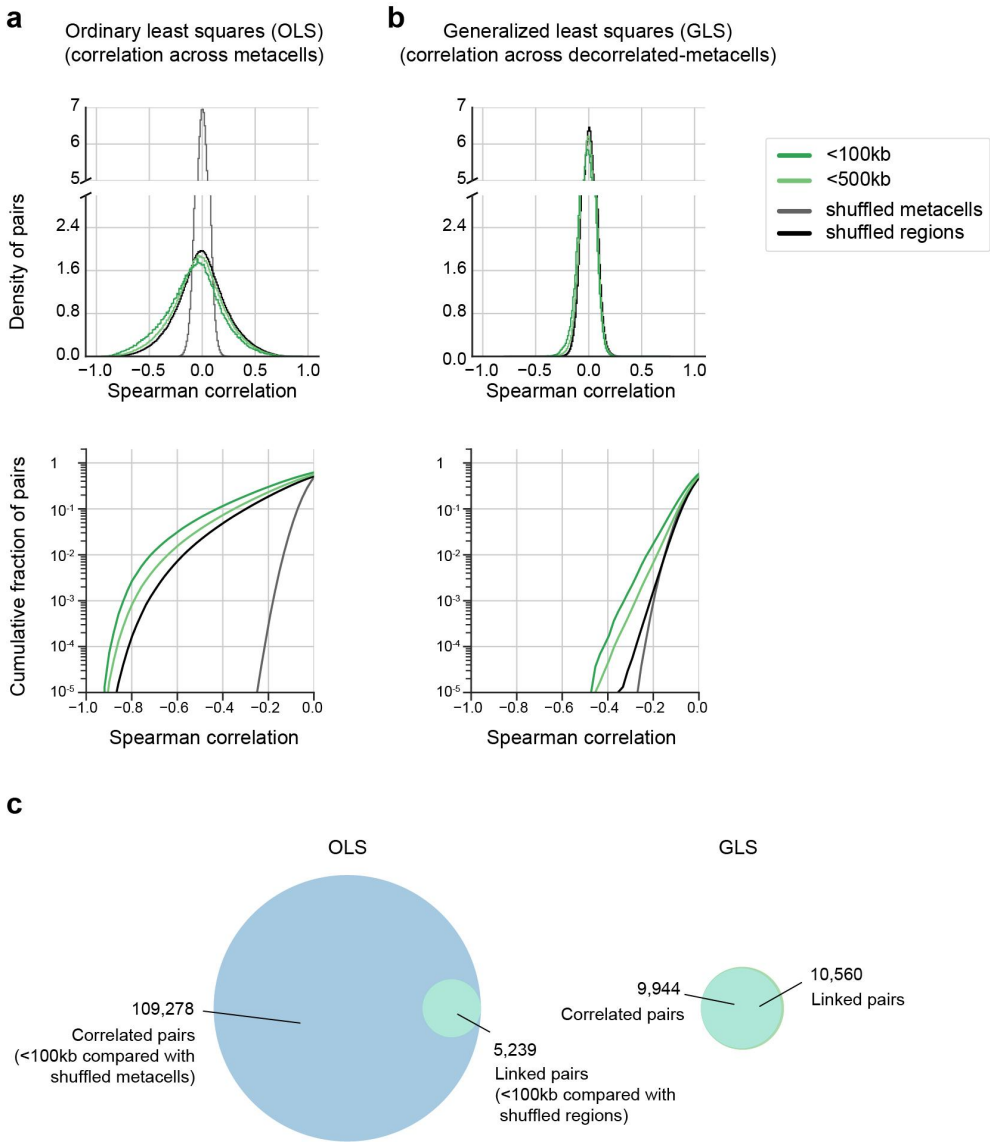
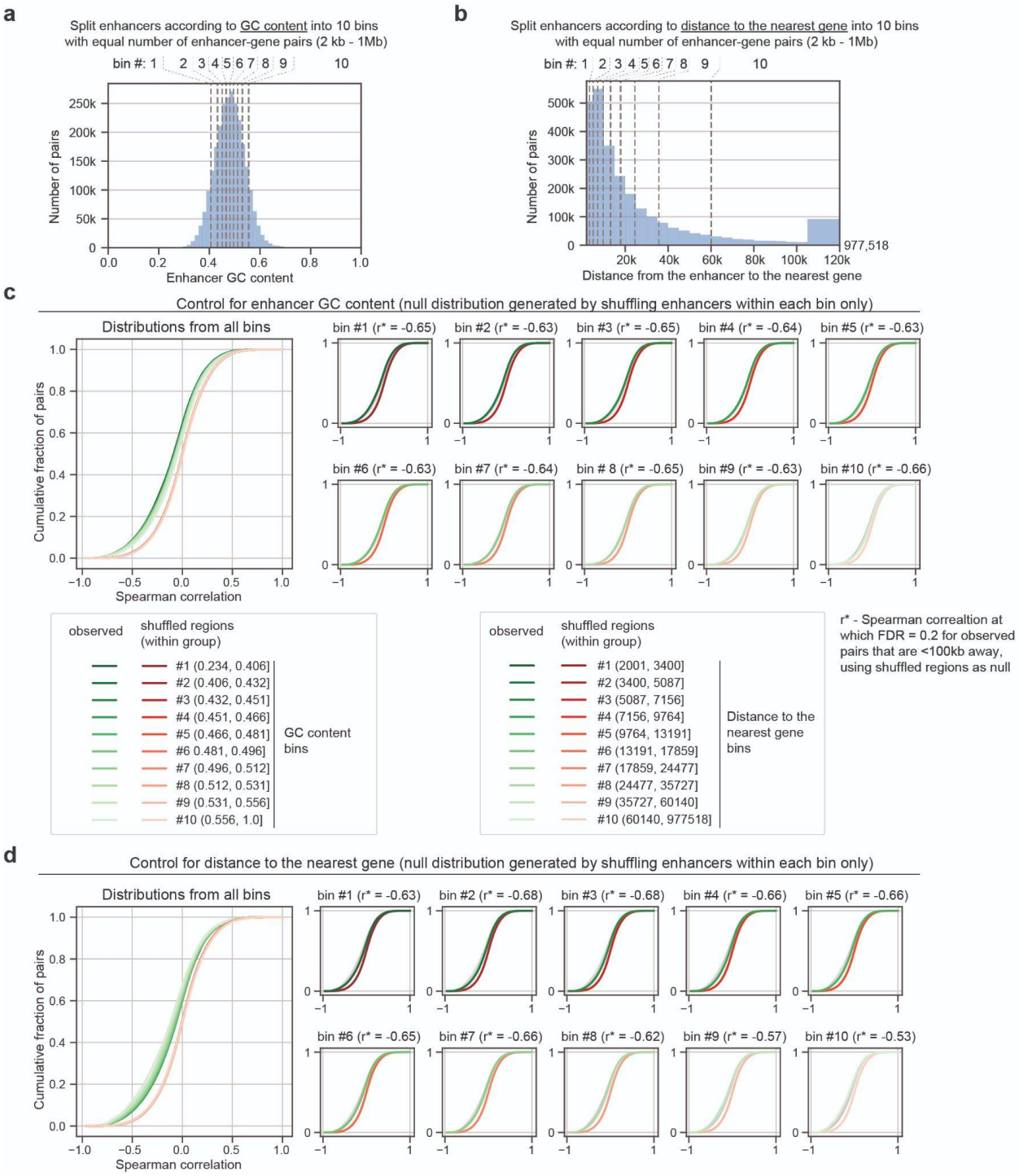


Figure S14: Generalized Least Squares (GLS) transformation abolishes the difference between shuffling metacells and shuffling regions. a-b. Density distribution (top) and cumulative distribution (bottom) of enhancer-gene correlations across metacells using OLS (**a**), and across decorrelated metacells using GLS (**b**). GLS transformation decouples the covariance across metacells, making the shuffling-regions distribution similar to the shuffling-metacell distribution (see Methods). **c**. Venn diagram showing the degree of overlap between correlated pairs and linked pairs, using OLS (left) and GLS (right) approaches. GLS abolishes the difference between correlated and linked pairs.

Figure S15: The shuffling-regions null distribution is robust with respect to enhancer GC content and distance to the nearest gene. a-b. Distribution of GC content (**a**) and distance to the nearest gene (**b**) for enhancers that are in all enhancer-gene pairs (2kb - 1Mb). In each case, they are grouped into 10 bins (deciles) with an equal number of enhancer-gene pairs. **c-d.** Cumulative distribution of enhancer-gene correlation (mCG-RNA; observed (<100kb) vs. null (shuffling regions)). The same analyses are applied to each of the 10 bins by enhancer GC content (**c**) and by distance to the nearest gene (**d**), respectively. Null distributions from different bins highly overlap.


a
Split enhancers according to GC content into 10 bins
with equal number of enhancer-gene pairs (2 kb - 1Mb)
bin #: 1 2 3 4 5 6 7 8 9 10
Number of pairs
Enhancer GC content
b
Split enhancers according to distance to the nearest gene into 10 bins
with equal number of enhancer-gene pairs (2 kb - 1Mb)
bin #: 1 2 3 4 5 6 7 8 9 10
Number of pairs
977,518
Distance from the enhancer to the nearest gene
c
Control for enhancer GC content (null distribution generated by shuffling enhancers within each bin only)
Distributions from all bins
Cumulative fraction of pairs
Spearman correlation
bin #1 (r* = -0.65)
bin #2 (r* = -0.63)
bin #3 (r* = -0.65)
bin #4 (r* = -0.64)
bin #5 (r* = -0.63)
bin #6 (r* = -0.63)
bin #7 (r* = -0.64)
bin # 8 (r* = -0.65)
bin #9 (r* = -0.63)
bin #10 (r* = -0.66)
observed
shuffled regions (within group)
#1 (0.234, 0.406]
#2 (0.406, 0.432]
#3 (0.432, 0.451]
#4 (0.451, 0.466]
#5 (0.466, 0.481]
#6 0.481, 0.496]
#7 (0.496, 0.512]
#8 (0.512, 0.531]
#9 (0.531, 0.556]
#10 (0.556, 1.0]
GC content bins
observed
shuffled regions (within group)
#1 (2001, 3400]
#2 (3400, 5087]
#3 (5087, 7156]
#4 (7156, 9764]
#5 (9764, 13191]
#6 (13191, 17859]
#7 (17859, 24477]
#8 (24477, 35727]
#9 (35727, 60140]
#10 (60140, 977518]
Distance to the nearest gene bins
r* - Spearman correaltion at which FDR = 0.2 for observed pairs that are <100kb away, using shuffled regions as null
d
Control for distance to the nearest gene (null distribution generated by shuffling enhancers within each bin only)
Distributions from all bins
Cumulative fraction of pairs
Spearman correlation
bin #1 (r* = -0.63)
bin #2 (r* = -0.68)
bin #3 (r* = -0.68)
bin #4 (r* = -0.66)
bin #5 (r* = -0.66)
bin #6 (r* = -0.65)
bin #7 (r* = -0.66)
bin #8 (r* = -0.62)
bin #9 (r* = -0.57)
bin #10 (r* = -0.53)

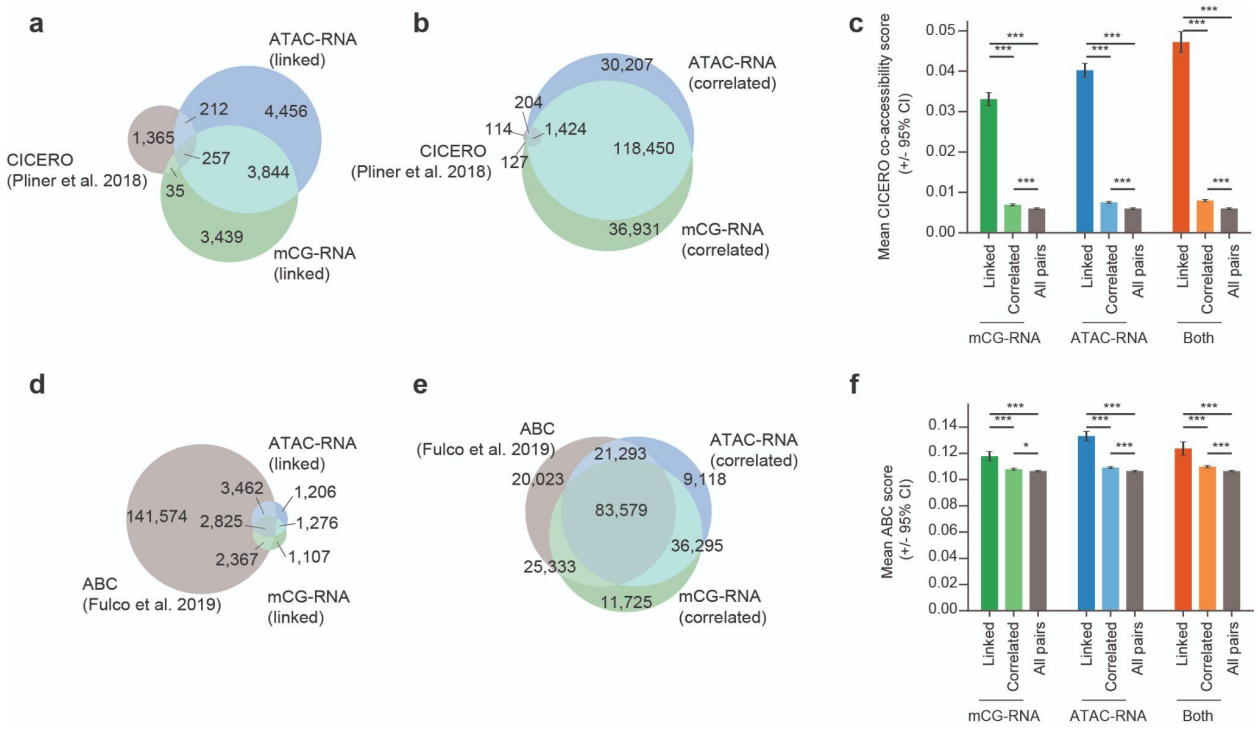


Figure S16: Comparison with CICERO[120] and the activity-by-contact (ABC) model[118]. a-b. Venn diagram comparing the enhancer-gene associations identified by applying CICERO[119] to the mouse MOp data[10] versus linked pairs (**a**) and correlated pairs (**b**) found in this study. **c.** Barplots comparing the mean CICERO scores across different groups of enhancer-gene pairs identified in this study. Error bars indicate 95% confidence intervals. Independent t-test are used to compare between groups (* p<0.05, *** p<0.001). **d-e.** Venn diagram comparing the enhancer-gene associations identified by applying ABC model[118] to the mouse MOp data[10] versus linked pairs (**d**) and correlated pairs (**e**) found in this study. **f.** Barplots comparing the mean ABC scores across different groups of enhancer-gene pairs identified by this study. ABC scores are generated for each enhancer-gene pair and cell type (n=38). We first took the maximum across cell types, followed by taking the mean of each group of enhancer-gene pairs. Error bars indicate 95% confidence intervals. Independent t-test are used to compare between groups (* p<0.05, *** p<0.001).

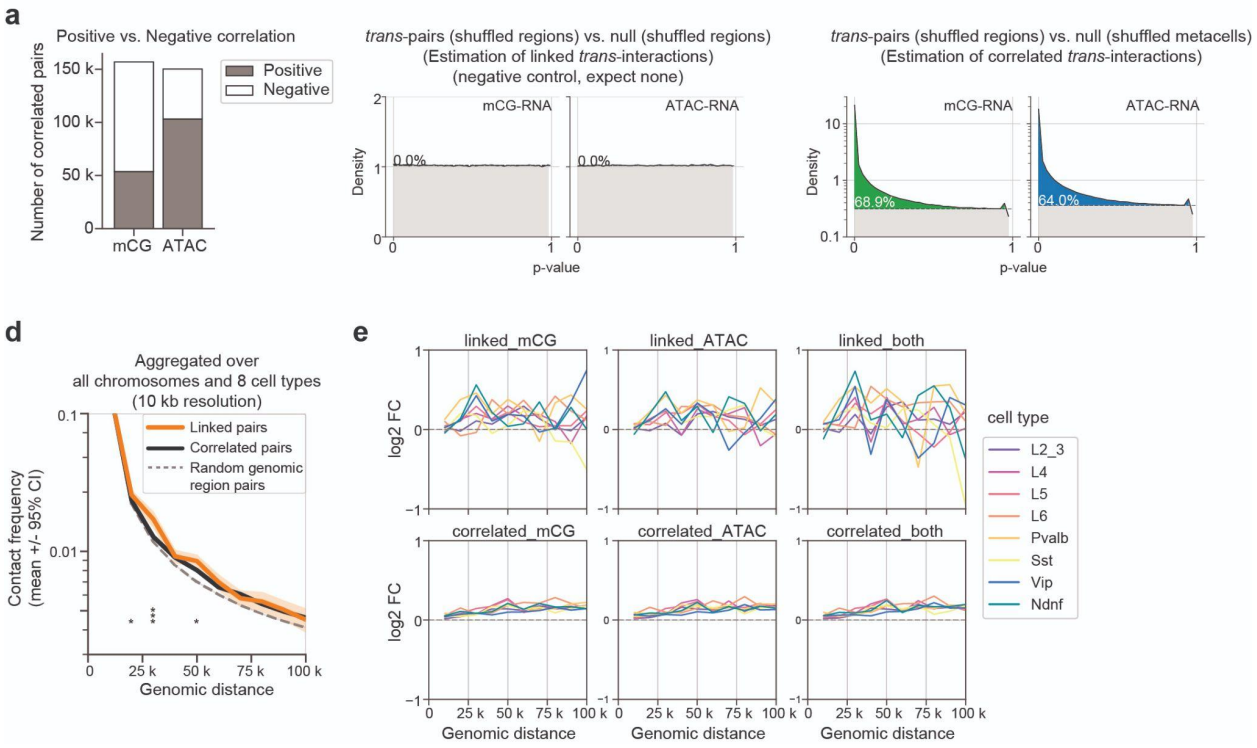


Figure S17: Linked vs. correlated enhancer-gene pairs have distinct characteristics. a. The number of positively or negatively correlated enhancer-gene pairs for mCG-RNA and ATAC-RNA, respectively. **b.** P-value histograms[129] of *trans*-enhancer-gene pairs using shuffling-regions as the null distribution. The histogram closely follows a uniform distribution, indicating trans-enhancer-gene pairs are linked. This serves as a negative control for Figure 9g. **c.** P-value histograms[129] of *trans*-enhancer-gene pairs, using shuffling-metacells as the null distribution. The numbers mark the fraction of p-values that deviate from the uniform distribution, which estimates the fraction of correlated *trans*-enhancer-gene pairs. **d.** Chromatin contact frequencies of pairs of genomic bins as a function of genomic distance. Linked and correlated pairs (lifted over from mm10 to hg38[138]) are compared with random genomic pairs. Results are aggregated over all chromosomes (autosomes + chrX) and 8 different human neuronal cell types (L2/3, L4, L5, L6, Pvalb, Sst, Vip, Ndnf) at 10kb resolution of chromatin contact maps[125]. **e.** Enrichment of contact frequency of linked and correlated enhancer-gene pairs compared with random genomic region pairs across 8 human neuronal cell types.

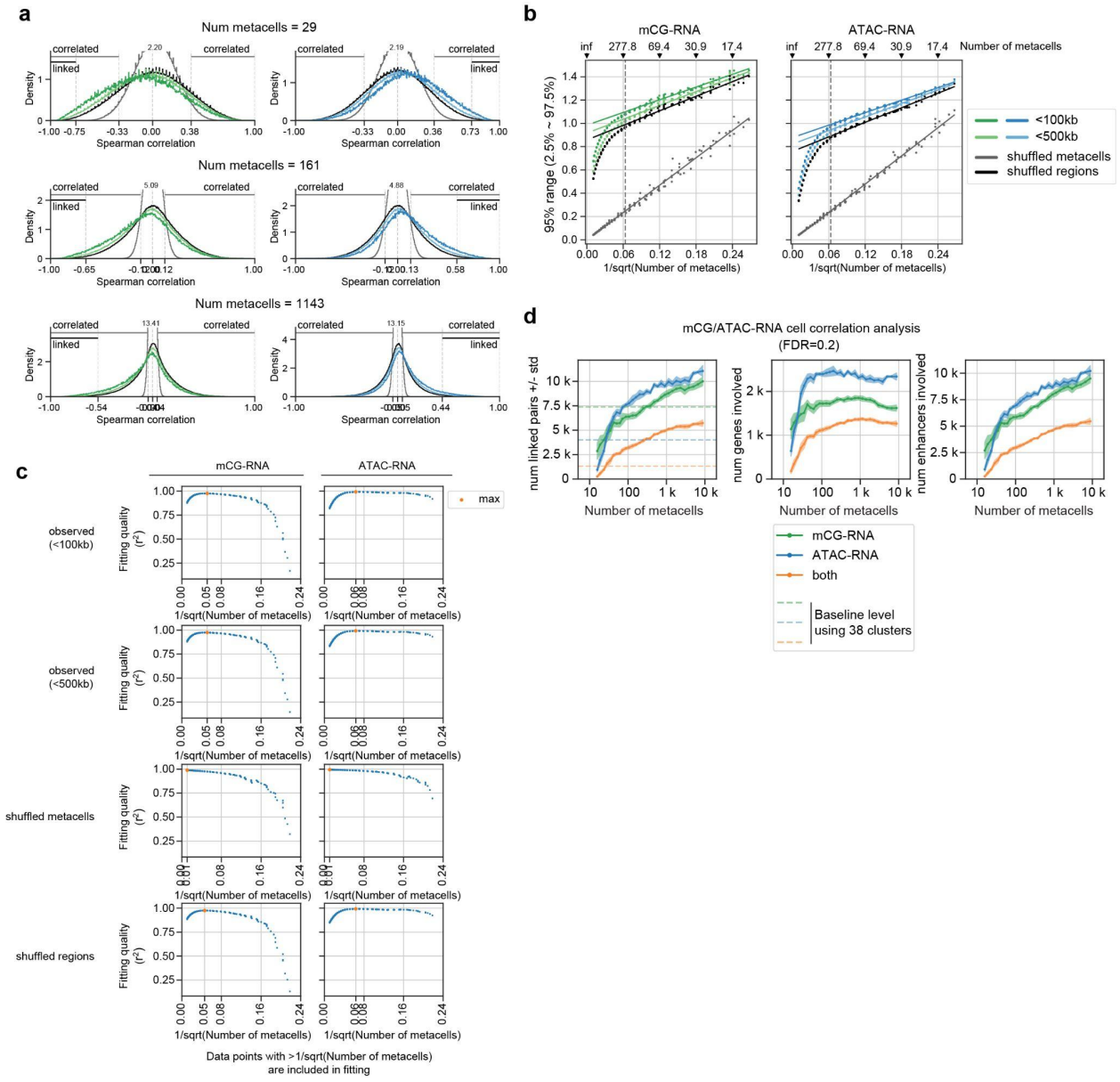


Figure S18: Effect of the granularity of metacells on enhancer-gene correlations. a. Distributions of correlation coefficients for different numbers of metacells. The distributions become narrower as the number of metacells increases. Here the number of metacells controls cell type granularity. **b.** Range of correlation coefficients (2.5%~97.5% range) as a function of $1/\sqrt{N}$, where N is the number of metacells. Data points are well fitted by a straight line for N < 277. **c.** Fitting quality, as measured by r^2, as a function of fitting cutoff--range of data points in (**b**) used for fitting. The fitting quality peaks at $1/\sqrt{N}$ = 0.05, i.e., N= 278. **d.** The number of linked pairs (left), number of genes involved (middle), and number of enhancers involved (right) as a function of the number of metacells.

REFERENCES

1. Ramon y Cajal, S. Histologie du système nerveux de l'homme et des vertébrés. *Maloine, Paris* **2,** 153–173 (1911).

2. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18,** 530–546 (2017).

3. Armand, E. J., Li, J., Xie, F., Luo, C. & Mukamel, E. A. Single-Cell Sequencing of Brain Cell Transcriptomes and Epigenomes. *Neuron* **109,** 11–26 (2021).

4. 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).

5. Saunders, A., Macosko, E. Z., Wysoker, A., Goldman, M., Krienen, F. M., de Rivera, H., Bien, E., Baum, M., Bortolin, L., Wang, S., Goeva, A., Nemesh, J., Kamitaki, N., Brumbaugh, S., Kulp, D. & McCarroll, S. A. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174,** 1015–1030.e16 (2018).

6. 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).

7. 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

8. 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).

9. 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).

10. 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).

11. Luo, C., Liu, H., Xie, F., Armand, E. J., Siletti, K., Bakken, T., Fang, R., Doyle, W. I., 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., Dixon, J. R., Linnarsson, S., Lein, E., Margarita Behrens, M., Ren, B., Mukamel, E. A. & Ecker, J. R. Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. *bioRxiv* 2019.12.11.873398 (2019). doi:10.1101/2019.12.11.873398

12. Xie, F., Armand, E. J., Yao, Z., Liu, H., Bartlett, A., Margarita Behrens, M., Li, Y. E., Lucero, J. D., Luo, C., Nery, J. R., Pinto-Duarte, A., Poirion, O., Preissl, S., Rivkin, A. C., Tasic, B., Zeng, H., Ren, B., Ecker, J. R. & Mukamel, E. A. Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes. *bioRxiv* 2021.10.25.465795 (2021). doi:10.1101/2021.10.25.465795

13. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. *Nature* **598,** 86–102 (2021).

14. Zeng, H. & Sanes, J. R. Neuronal cell-type classification: challenges, opportunities and the path forward. *Nat. Rev. Neurosci.* **18,** 530–546 (2017).

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. Scala, F., Kobak, D., Bernabucci, M., Bernaerts, Y., Cadwell, C. R., Castro, J. R., Hartmanis, L., Jiang, X., Laturnus, S., Miranda, E., Mulherkar, S., Tan, Z. H., Yao, Z., Zeng, H., Sandberg, R., Berens, P. & Tolias, A. S. Phenotypic variation of transcriptomic cell types in mouse motor cortex. *Nature* **598,** 144–150 (2021).

17. 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).

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

19. Efremova, M. & Teichmann, S. A. Computational methods for single-cell omics across modalities. *Nat. Methods* **17,** 14–17 (2020).

20. 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).

21. 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).

22. 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).

23. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0113-3

24. Abbott, L. F., Bock, D. D., Callaway, E. M., Denk, W., Dulac, C., Fairhall, A. L., Fiete, I., Harris, K. M., Helmstaedter, M., Jain, V., Kasthuri, N., LeCun, Y., Lichtman, J. W., Littlewood, P. B., Luo, L., Maunsell, J. H. R., Reid, R. C., Rosen, B. R., Rubin, G. M., Sejnowski, T. J., Seung, H. S., Svoboda, K., Tank, D. W., Tsao, D. & Van Essen, D. C. The mind of a mouse. *Cell* **182,** 1372–1376 (2020).

25. Yamawaki, N., Borges, K., Suter, B. A., Harris, K. D. & Shepherd, G. M. G. A genuine layer 4 in motor cortex with prototypical synaptic circuit connectivity. *Elife* **3,** e05422 (2014).

26. 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).

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

28. 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,** 1237905–1237905 (2013).

29. 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).

30. Mukamel, E. A. & Ngai, J. Perspectives on defining cell types in the brain. *Curr. Opin. Neurobiol.* **56,** 61–68 (2018).

31. Waddington, C. H. *The strategy of the genes*. (Routledge, 1957).

32. 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).

33. 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>

34. Vanderburg, C., Martin, C., Kozareva, V., Nadaf, N., Patel, N. & Macosko, E. Fresh Frozen Mouse Brain Preparation (for Single Nuclei Sequencing). (2020). doi:10.17504/protocols.io.bcbrism6

35. 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).

36. Miller, J. A., Gouwens, N. W., Tasic, B., Collman, F., van Velthoven, C. T., Bakken, T. E., Hawrylycz, M. J., Zeng, H., Lein, E. S. & Bernard, A. Common cell type nomenclature for the mammalian brain. *Elife* **9,** (2020).

37. Economo, M. N., Viswanathan, S., Tasic, B., Bas, E., Winnubst, J., Menon, V., Graybuck, L. T., Nguyen, T. N., Smith, K. A., Yao, Z., Wang, L., Gerfen, C. R., Chandrashekar, J., Zeng, H., Looger, L. L. & Svoboda, K. Distinct descending motor cortex pathways and their roles in movement. *Nature* **563,** 79–84 (2018).

38. Brodmann, K. *Brodmann's: Localisation in the Cerebral Cortex*. (Springer Science & Business Media, 2007).

39. Jabaudon, D., Shnider, S. J., Tischfield, D. J., Galazo, M. J. & Macklis, J. D. RORβ induces barrel-like neuronal clusters in the developing neocortex. *Cereb. Cortex* **22,** 996–1006 (2012).

40. Bakken, T. E., Hodge, R. D., Miller, J. A., Yao, Z., Nguyen, T. N., Aevermann, B., Barkan, E., Bertagnolli, D., Casper, T., Dee, N., Garren, E., Goldy, J., Graybuck, L. T., Kroll, M., Lasken, R. S., Lathia, K., Parry, S., Rimorin, C., Scheuermann, R. H., Schork, N. J., Shehata, S. I., Tieu, M., Phillips, J. W., Bernard, A., Smith, K. A., Zeng, H., Lein, E. S. & Tasic, B. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PLoS One* **13,** e0209648 (2018).

41. Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., Freier, S. M., Bennett, C. F., Sharma, A., Bubulya, P. A., Blencowe, B. J., Prasanth, S. G. & Prasanth, K. V. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **39,** 925–938 (2010).

42. 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).

43. Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A. & Trapnell, C. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14,** 309–315 (2017).

44. Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R. & Hemberg, M. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14,** 483–486 (2017).

45. Preissl, S., Wang, X. & Ren, B. Sequencing open chromatin of single cell nuclei: snATAC-seq protocol abstract. (2018). doi:10.17504/protocols.io.pjudknw

46. 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* **361,** 1380–1385 (2018).

47. Vruwink, M., Schmidt, H. H., Weinberg, R. J. & Burette, A. Substance P and nitric oxide signaling in cerebral cortex: anatomical evidence for reciprocal signaling between two classes of interneurons. *J. Comp. Neurol.* **441,** 288–301 (2001).

48. Peukert, D., Weber, S., Lumsden, A. & Scholpp, S. Lhx2 and Lhx9 determine neuronal differentiation and compartition in the caudal forebrain by regulating Wnt signaling. *PLoS Biol.* **9,** e1001218 (2011).

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

50. 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).

51. Harris, K. D., Hochgerner, H., Skene, N. G., Magno, L., Katona, L., Gonzales, C. B., Somogyi, P., Kessaris, N., Linnarsson, S. & Hjerling-Leffler, J. Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. *PLoS Biol.* **16,** e2006387 (2018).

52. Barkas, N., Petukhov, V., Nikolaeva, D., Lozinsky, Y., Demharter, S., Khodosevich, K. & Kharchenko, P. V. Wiring together large single-cell RNA-seq sample collections. *bioRxiv* 460246 (2018). doi:10.1101/460246

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

54. Hertler, B., Hosp, J. A., Blanco, M. B. & Luft, A. R. Substance P signalling in primary motor cortex facilitates motor learning in rats. *PLoS One* **12,** e0189812 (2017).

55. Harris, J. A., Hirokawa, K. E., Sorensen, S. A., Gu, H., Mills, M., Ng, L. L., Bohn, P., Mortrud, M., Ouellette, B., Kidney, J., Smith, K. A., Dang, C., Sunkin, S., Bernard, A., Oh, S. W., Madisen, L. & Zeng, H. Anatomical characterization of Cre driver mice for neural circuit mapping and manipulation. *Front. Neural Circuits* **8,** 76 (2014).

56. Madisen, L., Zwingman, T. A., Sunkin, S. M., Oh, S. W., Zariwala, H. A., Gu, H., Ng, L. L., Palmiter, R. D., Hawrylycz, M. J., Jones, A. R., Lein, E. S. & Zeng, H. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nature Neuroscience* **13,** 133–140 (2010).

57. 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).

58. 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).

59. Bortolin, L., Goldman, M. & McCarroll, S. Extraction of Nuclei from Brain Tissue v1 (protocols.io.2srged6). doi:10.17504/protocols.io.2srged6

60. Luo, C. & Ecker, J. R. Methyl-C sequencing of single cell nuclei: snmC-seq2 protocol abstract. (2018). doi:10.17504/protocols.io.pjvdkn6

61. 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).

62. 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).

63. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T. & Carey, V. J. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9,** e1003118 (2013).

64. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems* **8,** 329–337.e4 (2019).

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

66. Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., Olsen, B. N., Mumbach, M. R., Pierce, S. E., Corces, M. R., Shah, P., Bell, J. C., Jhutty, D., Nemec, C. M., Wang, J., Wang, L., Yin, Y., Giresi, P. G., Chang, A. L. S., Zheng, G. X. Y., Greenleaf, W. J. & Chang, H. Y. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37,** 925–936 (2019).

67. 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).

68. 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).

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

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

71. 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).

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

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

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

75. 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).

76. 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).

77. 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).

78. Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., Silva, T. C., Groeneveld, C., Wong, C. K., Cho, S. W., Satpathy, A. T., Mumbach, M. R., Hoadley, K. A., Robertson, A. G., Sheffield, N. C., Felau, I., Castro, M. A. A., Berman, B. P., Staudt, L. M., Zenklusen, J. C., Laird, P. W., Curtis, C., Cancer Genome Atlas Analysis Network, Greenleaf, W. J. & Chang, H. Y. The chromatin accessibility landscape of primary human cancers. *Science* **362,** (2018).

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

80. *Blacklist*. (Github). at <https://github.com/Boyle-Lab/Blacklist>

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

82. 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).

83. 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).

84. 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).

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

86. 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).

87. Wingender, E., Schoeps, T. & Dönitz, J. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.* **41,** D165–70 (2013).

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

89. 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).

90. 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).

91. Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., Krueger, F., Smallwood, S., 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).

92. 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).

93. 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).

94. 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).

95. 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).

96. 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).

97. 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).

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

99. 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).

100. 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).

101. 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* (2019). doi:10.1038/s41592-019-0619-0

102. 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

103. 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.* (2019). doi:10.1038/s41594-019-0323-x

104. 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).

105. 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).

106. 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).

107. 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).

108. 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).

109. Traag, V., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *arXiv [cs.SI]* (2018). at <http://arxiv.org/abs/1810.08473>

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

111. 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).

112. 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).

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

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

115. Gao, C., Liu, J., Kriebel, A. R., 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 single-cell

multi-omic integration using online learning. *Nat. Biotechnol.* (2021). doi:10.1038/s41587-021-00867-x

116. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21,** 292–310 (2020).

117. Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N. & Shendure, J. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176,** 1516 (2019).

118. 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).

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

120. 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).

121. Trevino, A. E., Sinnott-Armstrong, N., Andersen, J., Yoon, S.-J., Huber, N., Pritchard, J. K., Chang, H. Y., Greenleaf, W. J. & Pașca, S. P. Chromatin accessibility dynamics in a model of human forebrain development. *Science* **367,** (2020).

122. 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).

123. Sarropoulos, I., Sepp, M., Frömel, R., Leiss, K., Trost, N., Leushkin, E., Okonechnikov, K., Joshi, P., Giere, P., Kutscher, L. M., Cardoso-Moreira, M., Pfister, S. M. & Kaessmann, H. Developmental and evolutionary dynamics of cis-regulatory elements in mouse cerebellar cells. *Science* **373,** (2021).

124. Aitken, A. C. On Least Squares and Linear Combination of Observations. *Proceedings of the Royal Society of Edinburgh* **55,** 42–48 (1936).

125. 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).

126. Serwach, K. & Gruszczynska-Biegala, J. STIM Proteins and Glutamate Receptors in Neurons: Role in Neuronal Physiology and Neurodegenerative Diseases. *Int. J. Mol. Sci.* **20,** 2289 (2019).

127. Schoenfelder, S. & Fraser, P. Long-range enhancer-promoter contacts in gene expression control. *Nat. Rev. Genet.* **20,** 437–455 (2019).

128. Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A. & Tanay, A. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20,** 206 (2019).

129. Nettleton, D., Hwang, J. T. G., Caldo, R. A. & Wise, R. P. Estimating the number of true null hypotheses from a histogram of p values. *J. Agric. Biol. Environ. Stat.* **11,** 337 (2006).

130. 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).

131. Daigle, T. L., Madisen, L., Hage, T. A., Valley, M. T., Knoblich, U., Larsen, R. S., Takeno, M. M., Huang, L., Gu, H., Larsen, R., Mills, M., Bosma-Moody, A., Siverts, L. A., Walker, M., Graybuck, L. T., Yao, Z., Fong, O., Nguyen, T. N., Garren, E., Lenz, G. H., Chavarha, M., Pendergraft, J., Harrington, J., Hirokawa, K. E., Harris, J. A., Nicovich, P. R., McGraw, M. J., Ollerenshaw, D. R., Smith, K. A., Baker, C. A., Ting, J. T., Sunkin, S. M., Lecoq, J., Lin, M. Z., Boyden, E. S., Murphy, G. J., da Costa, N. M., Waters, J., Li, L., Tasic, B. & Zeng, H. A Suite of Transgenic Driver and Reporter Mouse Lines with Enhanced Brain-Cell-Type Targeting and Functionality. *Cell* **174,** 465–480.e22 (2018).

132. Graybuck, L. T., Daigle, T. L., Sedeño-Cortés, A. E., Walker, M., Kalmbach, B., Lenz, G. H., Morin, E., Nguyen, T. N., Garren, E., Bendrick, J. L., Kim, T. K., Zhou, T., Mortrud, M., Yao, S., Siverts, L. A., Larsen, R., Gore, B. B., Szelenyi, E. R., Trader, C., Balaram, P., van Velthoven, C. T. J., Chiang, M., Mich, J. K., Dee, N., Goldy, J., Cetin, A. H., Smith, K., Way, S. W., Esposito, L., Yao, Z., Gradinaru, V., Sunkin, S. M., Lein, E., Levi, B. P., Ting, J. T., Zeng, H. & Tasic, B. Enhancer viruses for combinatorial cell-subclass-specific labeling. *Neuron* **109,** 1449–1464.e13 (2021).

133. de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N. & Regev, A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.* **38,** 56–65 (2020).

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

135. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 289–300 (1995).

136. Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J. & Mirny, L. A. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9,** 999–1003 (2012).

137. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36,** 311–316 (2020).

138. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).