

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

A Single Cell Multiome Toolkit for Assessing Epigenome Diversity

Permalink

<https://escholarship.org/uc/item/8d17v7wk>

Author

XIE, YANG

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

A Single Cell Multiome Toolkit for Assessing Epigenome Diversity

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

in

Biomedical Sciences

by

Yang Xie

Committee in charge:

Professor Bing Ren, Chair
Professor Joseph Ecker
Professor Emma Farley
Professor Wei Wang
Professor Sheng Zhong

2024

Copyright

Yang Xie, 2024

All rights reserved.

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

University of California San Diego

2024

DEDICATION

To my families

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE.....	iii
DEDICATION	iv
TABLE OF CONTENTS	v
LIST OF FIGURES.....	vii
ACKNOWLEDGEMENTS	ix
VITA	xii
PUBLICATIONS	xii
ABSTRACT OF THE DISSERTATION	xiii
INTRODUCTION.....	1
Reference.....	10
Chapter 1 DROPLET-BASED SINGLE-CELL JOINT PROFILING OF HISTONE MODIFICATIONS AND TRANSCRIPTOMES	15
1.1 Abstract	15
1.2 Introduction	15
1.3 Results	17
1.4 Discussion	22
1.5 Acknowledgement.....	23
1.6 Figures	24
1.7 Supplementary Figures.....	28
1.8 Methods.....	43
1.9 References	55
Chapter 2 DROPLET HI-C FOR FAST AND SCALABLE PROFILING OF CHROMATIN ARCHITECTURE IN SINGLE CELLS.....	60
2.1 Abstract	60
2.2 Introduction	60

2.3 Results	63
2.4 Discussion	85
2.5 Acknowledgement.....	88
2.6 Figures.....	90
2.7 Supplementary Figures.....	99
2.8 Methods	116
2.9 References	145
DISCUSSION	156
Reference.....	159

LIST OF FIGURES

Figure 1.1: Joint profiling of transcriptomes and histone modifications in single cells using Droplet Paired-Tag.	24
Figure 1.2: Droplet Paired-Tag effectively resolves multiple cell types in the mouse frontal cortex (FC) and identifies the cCREs within each cell type.	26
Figure 2.1: Overview and performance of Droplet Hi-C.	90
Figure 2.2: Multi-scale chromatin structures are correlated with cell type-specific gene activity in adult mouse cortex.	91
Figure 2.3: Droplet Hi-C illuminates copy number, structural variations and ecDNA in cancer cells.	93
Figure 2.4: Droplet Hi-C reveals ecDNA dynamics under drug treatment.	94
Figure 2.5: Droplet Hi-C depicts ecDNA variations in patient samples.	96
Figure 2.6: Joint profiling of single-cell Hi-C and transcriptome.	98
Extended Data Figure 1.1: Comparison of the workflows of Droplet Paired-Tag and combinatorial-indexing-based Paired-Tag.	28
Extended Data Figure 1.2: Experimental design and quality control metrics of Droplet Paired-Tag.	29
Extended Data Figure 1.3: Quality control metrics of mESC Droplet Paired-Tag data.	31
Extended Data Figure 1.4: Annotation of cell clusters in the mouse frontal cortex Droplet Paired-Tag based on transcriptomic profiles.	33
Extended Data Figure 1.5: Integrative analysis of Droplet Paired-Tag transcriptomic profiles with public datasets.	35
Extended Data Figure 1.6: Quality control of mouse frontal cortex Droplet Paired-Tag histone modifications profiles.	37
Extended Data Figure 1.7: Landscape of histone modifications across cell types in mouse frontal cortex.	39
Extended Data Figure 1.8: Integrative analysis of histone modifications at candidates <i>cis</i>-regulatory elements across cell types in mouse frontal cortex.	41
Extended Data Figure 1.9: Nuclei gating strategy for mouse frontal cortex.	42
Extended Data Figure 2.1: Comparison between Droplet Hi-C and <i>in situ</i> Hi-C on cultured cells.	99
Extended Data Figure 2.2: Performance of Droplet Hi-C on adult mouse cortex.	101

Extended Data Figure 2.3: Validation of cell type identities in mouse cortex characterized by Droplet Hi-C.	103
Extended Data Figure 2.4: Droplet Hi-C illuminates CNV, SVs and ecDNA in cancer cells.	105
Extended Data Figure 2.5: Flowchart of ecDNA caller algorithms.	106
Extended Data Figure 2.6: Dynamics of ecDNA in GBM39 under erlotinib treatment.	107
Extended Data Figure 2.7: Cell state-specific chromatin structure and transcriptome differences in GBM patient samples.	108
Extended Data Figure 2.8: Chromatin structure differences between before and after treatment in BMMC patient samples.	110
Extended Data Figure 2.9: Workflow of Paired Hi-C.	111
Extended Data Figure 2.10: Joint chromatin organization and gene expression analysis using Paired Hi-C.	112
Extended Data Figure 2.11: Paired Hi-C illuminates alterations in the transcriptome associated with variations in chromatin structure.	113
Extended Data Figure 2.12: Downsampling analysis of chromatin interactions in ITL23GL subclass.	114
Extended Data Figure 2.13: Nuclei gating strategy for 7'AAD stained nuclei for HeLa S3 and mESC mixing experiment.	115

ACKNOWLEDGEMENTS

My journey in graduate studies began six years ago when I traveled across the vast Pacific Ocean. Reflecting on the history of biology, six years is long enough to witness significant transformations in the field; from the initial draft of the human genome to the introduction of the first Illumina Genome Analyzer; from rudimentary benchtop sequencer to the powerful machines capable of simultaneously generating hundreds of genomic datasets; and from bulk epigenomic assays to the widespread adoption of single-cell sequencing methods. The same amount of time also allows me to step into unfamiliar territories, acquire new skills, reshape ideas and perspectives, and question and embrace myself. Despite the new uncertainties I face today compared to when landed in San Diego, I am grateful to still possess a deep passion for my work. The rewards of this journey have far outweighed its challenges, leaving me profoundly satisfied.

I have been fortunate to enjoy abundant support and companionship throughout this journey. First and foremost, I am deeply grateful to my mom and dad for their unconditional love and encouragement in my pursuit of a career in science. They have consistently been there to share my happiness and shoulder my disappointments. I owe a special debt of gratitude to my dad; As a lifelong engineer, his motivation and dedication to continuous learning have profoundly influenced me, and I am certain that I will carry forward his spirit in my future endeavors.

My interest in biology traces back to high school, where my curiosity was ignited by the biology teacher Qingsong Chen. This early fascination turned biological sciences into my major in college, where I was fortunate to receive mentorships from investigators including Dr. Guodong Ren and Dr. Liang Cai, who nurtured my development as an experimental biologist. During my last year of undergraduate studies, thanks to the introduction by Dr. Fei Lan, I had the opportunity to work as a research assistant at UCSD and spent around half of the year in the Ren lab. It was

during this period that I became obsessed with epigenome and 3D genome, topics I continued to explore upon returning to sunny La Jolla.

I subsequently spent a year in Dr. Clodagh O'Shea's group alongside Dr. Horng Ou, who remains a close friend of mine. Eventually, I decided to reunite with the Ren lab to reignite my passion for 3D genome research. I am deeply grateful to Bing Ren, my thesis supervisor, whose guidance and exceptional mentorship have profoundly shaped my graduate school experience. Bing's foresight and unparalleled generosity have secured me resources as well as freedom to explore various new possibilities. During my time in the Ren lab, I had the privilege of working with Dr. Chenxu Zhu, a brilliant scientist who laid me a foundation for single-cell methodologies, and Dr. Lei Chang, my perseverant and dedicated collaborator. My committee members, Dr. Emma Farley, Dr. Joseph Ecker, Dr. Sheng Zhong, and Dr. Wei Wang have provided me with invaluable suggestions and assistance in shaping my thesis. I am immensely grateful to all my mentors who have guided me along this journey. As I move forward, I hope to have chances to pay forward the kindness and mentorship I have, helping others who may need help.

Friendship has been an important part of my journey. My experience in the Ren lab could not be more amusing without a group of fantastic colleagues and friends. Special thanks to Dr. Yang Li, Ethan Armand, and my great roommate Dr. Zhaoning Wang for organizing activities and sharing tons of memorable moments with me. I am also grateful to Bojing Jia, whose passion for science sparked numerous late-night discussions during the worst period of the pandemic. I found invaluable emotional support within the Chinese cohort of the Biological Science program, including Xiaohui Lyh, Jingwen Liao, Jiahan Peng, Xiaoxue Lin, and Yuhui Luo, who shared in both my challenges and joys. While I regret not returning home during my six years away, I deeply cherished the unwavering support from all my old friends afar in different chatting groups or on

WeChat, especially from my undergrad roommate Ziyue Wang. I'm delighted to see both of us continue to pursue our original goals and beliefs. To all my friends, I sincerely hope you continue to pursue your passions and dreams.

Chapter 1 is, in full, a reformatted reprint of the material as it appears as “Droplet-based single-cell joint profiling of histone modifications and transcriptomes” in *Nature Structural & Molecular Biology*, 2023. Yang Xie, Chenxu Zhu, Zhaoning Wang, Melodi Tastemel, Lei Chang, Yang Eric Li and Bing Ren. The dissertation author was the primary investigator and author of this paper.

Chapter 2 is, in full, a reformatted reprint of the material as it appears as “Droplet Hi-C for Fast and Scalable Profiling of Chromatin Architecture in Single Cells” in *bioRxiv*, 2024. Lei Chang, Yang Xie, Brett Taylor, Zhaoning Wang, Jiachen Sun, Tuyet R. Tan, Rafael Bejar, Clark C. Chen, Frank B. Furnari, Ming Hu, Bing Ren. The dissertation author was the primary investigator and co-first author of this paper.

VITA

Bachelor of Science in Biological Sciences, Fudan University	2018
Doctor of Philosophy in Biomedical Sciences, University of California San Diego	2024

PUBLICATIONS

Chang, L.*, **Xie, Y.***, Taylor, B., Wang, Z., Sun, J., Tan, T. R., et al. (2024). Droplet Hi-C for Fast and Scalable Profiling of Chromatin Architecture in Single Cells. *bioRxiv*.

Xie, Y.*, Zhu, C.*, Wang, Z., Tastemel, M., Chang, L., Li, Y. E., & Ren, B.# (2023). Droplet-based single-cell joint profiling of histone modifications and transcriptomes. *Nature Structural & Molecular Biology*, 30(10), 1428–1433.

Li, Y. E., Preissl, S., Miller, M., Johnson, N. D., Wang, Z., Jiao, H., Zhu, C., Wang, Z., **Xie, Y.**, Poirion, O., et al. (2023). A comparative atlas of single-cell chromatin accessibility in the human brain. *Science*, 382(6667), eadf7044.

Zu, S.*, Li, Y. E.*, Wang, K.*, Armand, E. J., Mamde, S., Amaral, M. L., Wang, Y., Chu, A., **Xie, Y.**, Miller, M., et al. (2023). Single-cell analysis of chromatin accessibility in the adult mouse brain. *Nature*, 624(7991), 378–389.

Zemke, N. R.*, Armand, E. J.*, Wang, W., Lee, S., Zhou, J., Li, Y. E., Liu, H., Tian, W., Nery, J. R., Castanon, R. G., Anna, B., Julia, K. O., Li, D., Zhou, X., Xu, V., Chang, L., Dong, K., Indralingam, H. S., Rink, J. A., **Xie, Y.**, Miller, M., Krienen, F. M. et al. (2023). Conserved and divergent gene regulatory programs of the mammalian neocortex. *Nature*, 624(7991), 390–402.

* co-first authors

ABSTRACT OF THE DISSERTATION

A Single Cell Multiome Toolkit for Assessing Epigenome Diversity

by

Yang Xie

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2024

Professor Bing Ren, Chair

The epigenome controls the spatial and temporal specificity of gene expression in mammalian cells, guiding different cell fates and developmental trajectories. Our current

understanding of how the epigenome regulates transcription is primarily based on bulk epigenetic assays. While these assays provide valuable information, the data represents population averages, which obscure detailed cell types or cell state-specific information in heterogeneous systems. Moreover, bulk data is dominated by the most abundant cells in a sample, leading to the underrepresentation of rare cell types.

Recent developments in single-cell epigenetic profiling methods have mitigated these issues, although large parts of the epigenome, such as histone modifications and chromatin interactions, remain relatively under-explored at the single-cell level. In my thesis, I developed a set of sequencing tools to link transcriptional outcomes with epigenetic information, including histone modifications and chromatin interactions, in single cells. This enables an understanding of the cell type-specific roles of under-explored repressive elements and chromatin structures in both healthy and diseased mammalian brains.

In Chapter 1, I describe Droplet Paired-Tag, a droplet-based single-cell joint profiling method for histone modifications and transcriptomes. This method addresses the challenges of current single-cell histone modification mapping techniques by offering shorter handling times, a simplified protocol, and higher data quality. Through this approach, I successfully annotated the chromatin states of candidate cis-regulatory elements (cCREs) and predicted both active and repressive cCREs in various cell types in the adult mouse cerebral cortex.

In Chapter 2, I detail the development of Droplet Hi-C, a high-throughput method for profiling chromatin conformation in single cells. This technique not only enables large-scale analysis of cell type-specific chromatin structures in heterogeneous tissues, such as the mouse brain, but also has been pivotal in detecting copy number and structural variations, including

extrachromosomal DNA (ecDNA), in brain cancer cells such as glioblastoma. Furthermore, I demonstrate that Droplet Hi-C can be extended to a multiomics approach called Paired Hi-C, which facilitates the simultaneous analysis of gene expression and chromatin conformation in the same cell. This advancement provides an opportunity to connect ecDNA copy numbers with gene expression variations in brain cancer cells under drug treatment.

INTRODUCTION

It is estimated that only about 2% of the human genome consists of protein-coding sequences, while the remaining 98% was long considered 'junk DNA' due to the lack of methods to characterize its function over a long time^{1,2}. The notion of 'junk DNA' began to change as evidence emerged showing that cis-regulatory elements (CREs) within these non-coding regions play critical roles in regulating transcription in various tissue samples^{3,4,5}. These CREs are not only prevalent and conserved across species, but they also harbor the majority of disease-associated alleles identified by genome-wide association studies⁶. This underscores their essential role in regulating gene function and maintaining genomic integrity.

Even after the release of the human genome sequence, accurately identifying CREs based solely on sequence information remains challenging, particularly in determining their spatial and temporal activities⁷. Epigenetic features on chromatin, such as chemical modifications and physical conformation, have long been known to influence the accessibility and activity of underlying DNA sequences^{8,9}. However, the genomic distribution of these epigenetic features and their association with CREs remained elusive until the development and widespread use of microarray and next-generation sequencing (NGS) technologies.

The advancements in methodologies have enabled several large-scale studies, including the Roadmap Epigenome Project and the Encyclopedia of DNA Elements (ENCODE) consortium, to profile epigenetic features in numerous cell lines and tissue samples under various conditions^{10,11,12}. These efforts have led to the successful annotation of millions of candidate CREs (cCREs), including but not limited to promoters, enhancers, insulators, and silencers¹³. Each cCRE is characterized by a specific combination of DNA and histone modifications that indicate its identity and potential function¹⁴. The same cCRE can acquire additional modifications that reflect

its functional state across different tissues. It can also switch roles in some cases, such as changing from an enhancer to a silencer during cell differentiation¹⁵. Furthermore, multiple cCREs can cooperate to make the same target genes either robust to perturbations or susceptible to extrinsic signals¹⁶. To achieve this coordination, cCREs often form chromatin loops, bringing them spatially proximal to their target genes¹⁷. This complex and dynamic regulatory network is crucial for controlling proper gene expression patterns during cell fate determination and other cellular behaviors. Malfunctions in any part of this regulatory network can lead to disease or severe developmental defects¹⁸.

Our understanding of cCREs and their associated epigenetic features has significantly advanced thanks to the development of new techniques. Chromatin immunoprecipitation (ChIP) was one of the earliest methods piloted to identify transcription factor (TF) binding sites and histone modification landscapes. When coupled with microarray (ChIP-on-chip) or next-generation sequencing (ChIP-seq, ChIP-exo)^{19,20,21}. ChIP has provided invaluable insights into these regulatory regions. Most TFs bind to accessible regions of the genome, which can be measured using assays such as FAIRE-seq, DNase-seq, and ATAC-seq^{21,22,23}. In contrast, relatively inactive, inaccessible genomic regions are often marked by repressive histone marks, or DNA methylation, which can be measured by whole-genome bisulfite sequencing (WGBS)²⁴. Interactions between cCREs and genes can be studied by chromatin conformation capture methods like Hi-C, and this method has greatly helped us reveal hierarchical chromatin structures, from chromosome territories and compartments to topologically associating domains (TADs) and cCREs-gene loops²⁵. Insulators, frequently bound by CTCF, demarcate TAD boundaries and play crucial roles in restricting the spread of repressive heterochromatin marks or preventing unwanted interactions between genes and non-target enhancers²⁶. Promoters and their regulated enhancers

typically reside within the same TAD, and their interactions can be captured more efficiently by methods with higher specificity or resolution, such as HiChIP, PLAC-seq, Micro-C, and various imaging-based techniques^{27,28,29}. The continuous development of new sequencing tools allows for more sensitive, efficient, and convenient measurement of epigenetic features, facilitating broader applications to study different biological questions. These advancements are moving us closer to a comprehensive understanding of cCRE functions.

Despite the tremendous progress, the current characterization of cCREs faces several obstacles. While we already own a rich catalog of cCREs, establishing functional links between these elements and their target genes remains challenging due to the inefficiency of manipulating genomic loci in situ. High-throughput screening methods like STARR-seq and MPRA have partially addressed this issue, but presenting cCREs in an artificial genomic environment or altered cellular states may not accurately reflect their original functions^{30,31}.

Moreover, the current catalog of cCREs is primarily defined using datasets generated at the bulk level. This means that in complex tissues, the data only represents a population average across various cell types, lacking the resolution needed to capture cell-type-specific or single-cell information. The functional states of cCREs are often specific to a few cell types or particular cell states, which cannot be distinguished using bulk data from heterogeneous primary tissues. This averaging effect can obscure the contributions of low-abundance or poorly characterized cell types, whose epigenetic profiles may still play crucial roles or carry significant information for understanding diseases. For example, cancer stem cells, though low in number, can have important functions and provide critical insights into tumor biology and potential therapeutic targets³².

Fortunately, the limitations posed by cell type heterogeneity in cCRE characterization can be

addressed with the development of single-cell epigenomic techniques. These methods typically involve dissociating tissue samples and partitioning single cells or nuclei into isolated chambers, such as micro-wells (well-based) or oil droplets (droplet-based). Each isolated chamber allows individual cells to be tagged with a unique molecular identifier, enabling distinct reactions to occur within each cell³³.

Facilitated by mature microfluidic platforms, droplet-based methods are faster and can achieve higher throughput compared to microwell-based methods. Microwell-based methods, on the other hand, are useful for studying systems with limited cell numbers, such as zygotes and early-stage embryos, while droplet-based methods can accommodate a broad range of sample types, processing between 1,000 and 10,000 cells^{34,35,36}. The throughput of single-cell experiments can be further enhanced using combinatorial indexing strategy (split-and-pool-based), where cells are repeatedly distributed into multiple barcoding plates, each time ligating with unique barcodes. Each well in a plate can contain multiple cells during each round of barcode ligation³⁷. Although cells may share the same initial barcode, repeated redistribution ensures they can acquire unique barcode combinations, exponentially scaling the combinations to millions. This split-and-pool-based method opens up the possibility of profiling highly complex and heterogeneous systems, such as late-stage embryos or the entire mammalian brain, albeit with a more complex workflow compared to droplet-based methods³⁸.

These single-cell strategies were initially adopted for transcriptome profiling, creating extensive resources for depicting cell type or cellular state heterogeneity, analyzing RNA isoforms, and inferring co-expression modules in the complex tissues^{39,40}. But in recent years, these strategies have been coupled with bulk sequencing methods to profile distinct epigenetic features including but not limited to chromatin accessibility, DNA and histone modifications, and

chromatin interaction at the cell type level.

To measure chromatin accessibility, single-cell (sc-) and single-nucleus (sn-) ATAC-seq have become popular due to their simple yet robust protocols as well as high performance⁴¹. For histone modifications, ChIP-seq was adapted to single-cell strategies with improvements to accommodate fewer input cells⁴². However, due to ChIP's low signal-to-noise ratio, alternative methods with higher sensitivity are preferred for single-cell analysis. Methods like CUT&RUN, which uses antibody-targeted Protein A or Protein G conjugated with micrococcal nuclease (MNase) to cleave antibody-labeled chromatin, and CUT&Tag, which has the same design but employs transposase Tn5 instead of MNase, have been developed and quickly adopted for their sensitivity, simplicity, and ability to perform reactions in intact nuclei^{43,44}. Built on these new bulk assays, methods such as scChIC-seq and scCUT&Tag allow efficient profiling of chromatin modifications and even transcription factor binding in individual cells^{45,46,47}. To measure DNA methylation, bisulfite sequencing combined with well-based methods has led to techniques like scWGBS and snmC-seq, enabling genome-wide DNA methylation analysis in single cells^{48,49}. Methods with higher throughput, like sciMET, have also become available⁵⁰. For chromatin interactions, most methods involve adaptations of the conventional Hi-C protocol, including scHi-C, Dip-C, and sci-Hi-C^{51,52,53}. Other methods like scSPRITE use several rounds of split-and-pool barcoding to create identical barcodes for DNA fragments in close proximity⁵⁴.

Together, these single-cell methods offer opportunities for detailed characterization of cCREs with cell type resolution in the brain, tumors, and other tissues or diseases. For example, through snATAC-seq profiling of the adult mouse cerebellum, over 98% of known cCREs were found to exhibit variable activities across cell types⁵⁵. The cCREs catalog also expands with the identification of new elements specific to only one or a few cell types. Within each cell type,

cCREs activity can be further classified as repressed, primed, active, or weak by integrating information about DNA and histone modifications. Combining this information with disease-related risk variants helps us better interpret which cCREs and cell types are more likely to be affected in neurological disorders such as schizophrenia.

Single-cell epigenetic tools profiling one modality at a time have already provided in-depth insights into gene regulation across diverse cell types and states. However, they offer only a partial view of the gene regulation process. To accurately identify gene-targeting enhancers, infer the hierarchy of gene regulatory networks, and create a codebook of key transcription factors, it is necessary to link cCREs with their target genes. Assays that measure more than one modality have gradually become possible, employing strategies such as physically separating the cytoplasm and nucleus or adding orthogonal molecular barcodes to different modalities⁵⁶.

The inclusion of multiple modalities allows for joint analysis of transcriptome and epigenome, serving as a reference to bridge the analysis of single-modality datasets. For example, by profiling transcriptome and histone modification data in the same cells using Paired-Tag, a split-and-pool-based multiomic method, we can obtain landscapes of different histone modifications within single cells⁵⁷. Using the transcriptome modality as a common anchor, integrating multiple histone modifications belonging to the same cell types becomes feasible. This integration allows for detailed characterization of the cCREs chromatin states within different cell types, revealing information that was previously inaccessible.

Currently, the majority of single-cell epigenetic studies on cCREs focus on chromatin accessibility and active enhancers due to the efficiency and broad adaptability of snATAC-seq. However, the cCRE catalog includes other types of elements, and at least 10% of accessible cCREs

do not carry active enhancer marks. Achieving a comprehensive understanding of the variations in epigenetic features of cCREs remains a challenge.

One major limitation is the applicability and efficiency of new epigenetic profiling methods. Microwell-based methods, while useful, require expensive automated machines or customized consumables such as microwell plates or chips, therefore they cannot be efficiently scaled up. The split-and-pool-based method is powerful yet involves tedious workflows and requires extensive labor. In contrast, droplet-based methods are convenient, and are among the first to be commercialized³⁴. Regardless of their widespread use, many droplet-based platforms are not open source, making it challenging to repurpose them for profiling other understudied epigenetic features, including histone modifications or chromatin interaction profiles.

In response to the efficiency problem, in chapter 1, my colleagues and I developed Droplet Paired-Tag, a robust and sensitive single-cell multiomic method for the joint analysis of chromatin modifications and transcriptome within the same single cell. As a successor to the split-and-pool-based Paired-Tag method, Droplet Paired-Tag offers greater convenience and higher data quality. By applying this method to the adult mouse cortex, we systematically characterized both active and repressive cCREs across different mouse cortex cell types. Additionally, we identified a set of active and repressive cCREs associated with genes closely related to cell type-specific functions. Our dataset serves as one of the earliest resources for a comprehensive understanding of gene regulatory networks in the mammalian brain.

cCREs are known to exert their functions by looping with their target genes over long genomic distances. To bridge the information between cell-type-specific cCREs and gene expression, obtaining cell type-specific chromatin interaction profiles is crucial. Additionally,

chromatin interaction data generated by Hi-C are frequently used to infer copy number and structural variations, both of which are common and highly heterogeneous but hard to measure among cancer cells. However, even with existing single-cell Hi-C methods, chromatin conformation profiling has not been widely adopted for primary tissues. We also have limited knowledge about cell type-specific chromatin contacts that can confidently be used for annotating cell identities.

To overcome this hurdle, in chapter 2, my colleagues and I introduced Droplet Hi-C, a technique based on a commercial microfluidic device for high-throughput, single-cell chromatin conformation profiling. Droplet Hi-C surpasses previous methods in terms of throughput, convenience, and cost-effectiveness. After developing an analysis framework by leveraging the relationship between transcription and gene body chromatin interactions, we applied Droplet Hi-C to the mouse cortex and accurately annotated cell type identities using chromatin interaction profiles. With this framework, we successfully mapped the multi-layer chromatin architecture at single-cell resolution, enabling an analysis of gene regulatory programs in major cortical cell types.

Furthermore, by harnessing the unique power of Hi-C in revealing structural alterations in the genome, we applied Droplet Hi-C to detect complex structural variations, including extrachromosomal DNA in cancer cells and patient tumor samples. We also developed a deep-learning-based ecDNA caller, which, in combination with Droplet Hi-C, helps us uncover ecDNA dynamics during drug treatment processes. Finally, we extended Droplet Hi-C to achieve joint profiling of chromatin architecture and the transcriptome in single cells. This advancement provides a more exhaustive exploration of the connections between chromatin architecture and gene expression in both normal tissues and tumors.

Adoption is crucial for generalizing a method and fostering insightful discoveries. By introducing a comprehensive toolkit for exploring relatively understudied epigenetic features, I believe our work will advance our understanding of regulatory elements in gene regulation and continually contribute to biomedical research and precision medicine.

Reference

1. Palazzo, A. F., & Gregory, T. R. (2014). The case for junk DNA. *PLoS genetics*, 10(5), e1004351. <https://doi.org/10.1371/journal.pgen.1004351>
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001). <https://doi.org/10.1038/35057062>
3. Creighton, H. B., & McClintock, B. A correlation of cytological and genetical crossing-over in *Zea mays*. *Proceedings of the National Academy of Sciences* 17, 492–497 (1931)
4. Chung, J.H., Whiteley, M. & Felsenfeld, G. A. element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. *Cell* 74, 505–514 (1993).
5. Banerji, J., Olson, L., & Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, 33(3), 729–740. [https://doi.org/10.1016/0092-8674\(83\)90015-6](https://doi.org/10.1016/0092-8674(83)90015-6)
6. Maurano, Matthew T., et al. "Systematic localization of common disease-associated variation in regulatory DNA." *Science* 337.6099 (2012): 1190-1195.
7. Alexander, R. P., Fang, G., Rozowsky, J., Snyder, M., & Gerstein, M. B. (2010). Annotating non-coding regions of the genome. *Nature Reviews Genetics*, 11(8), 559-571.
8. Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33(3), 245-254.
9. Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: a landscape takes shape. *Cell*, 128(4), 635-638.
10. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57.
11. Roadmap Epigenomics Consortium. *Nature* 518, 317–330 (2015).
12. The ENCODE Project Consortium., Snyder, M.P., Gingeras, T.R. et al. Perspectives on ENCODE. *Nature* 583, 693–698 (2020). <https://doi.org/10.1038/s41586-020-2449-8>
13. Ngan, C. Y., Wong, C. H., Tjong, H., Wang, W., Goldfeder, R. L., Choi, C., He, H., Gong, L., Lin, J., Urban, B., Chow, J., Li, M., Lim, J., Philip, V., Murray, S. A., Wang, H., & Wei, C. L. (2020). Chromatin interaction analyses elucidate the roles of PRC2-bound silencers in mouse development. *Nature genetics*, 52(3), 264–272. <https://doi.org/10.1038/s41588-020-0581-x>
14. Ernst, Jason, and Manolis Kellis. "Discovery and characterization of chromatin states for systematic annotation of the human genome." *Nature biotechnology* 28.8 (2010): 817-825.

15. Huang, D., & Ovcharenko, I. (2022). Enhancer–silencer transitions in the human genome. *Genome Research*, 32(3), 437-448.
16. Hong, Joung-Woo, David A. Hendrix, and Michael S. Levine. "Shadow enhancers as a source of evolutionary novelty." *Science* 321.5894 (2008): 1314-1314.
17. Jerkovic, I., & Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology*, 22(8), 511-528.
18. Cavalli, G., & Heard, E. (2019). Advances in epigenetics link genetics to the environment and disease. *Nature*, 571(7766), 489-499.
19. Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., & Young, R.A. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290: 2306-2309.
20. Johnson, D. S. et al. Genome-wide mapping of in vivo protein–DNA interactions. *Science* 316, 1497–1502 (2007)
21. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147, 1408–1419 (2011).
22. Nagy P.L., Cleary M.L., Brown P.O., Lieb J.D., Cleary M.L., Brown P.O., Lieb J.D., Brown P.O., Lieb J.D., Lieb J.D. Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. *Proc. Natl. Acad. Sci.* 2003;100:6364–6369.
23. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. High -resolution mapping and characterization of open chromatin across the genome. *Cell*. 2008;132:311–22.
24. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, 10(12), 1213-1218.
25. Lister, R., Pelizzola, M., Downen, 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. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), 315–322. <https://doi.org/10.1038/nature08514>
26. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., Sandstrom, R., Bernstein, B., Bender, M.A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L.A., Lander, E.S., Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289-293.
27. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., ... & Ren, B. (2012). Topological

- domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376-380.
28. Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods*, 13(11), 919-922.
 29. Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A. D., & Ren, B. (2016). Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell research*, 26(12), 1345-1348.
 30. Hsieh, T. H. S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., & Rando, O. J. (2015). Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell*, 162(1), 108-119.
 31. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123), 1074-1077.
 32. Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.-I., Cooper, G.M., Ahituv, N., Pennacchio, L.A., Shendure, J., (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature biotechnology*, 30(3), 265-270.
 33. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., Louis, D.N., Rozenblatt-Rosen, O., Suvà, M.L., Regev, A., Bernstein, B.E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190), 1396-1401.
 34. Preissl, S., Gaulton, K. J., & Ren, B. (2023). Characterizing cis-regulatory elements using single-cell epigenomics. *Nature Reviews Genetics*, 24(1), 21-43.
 35. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., Lao, K., Surani, M.A., (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5), 377-382.
 36. Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., Kirschner, M.W., (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5), 1187-1201.
 37. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A., McCarroll, S.A., (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5), 1202-1214.
 38. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745), 496–502.

39. Boeshaghi, A. S., Yao, Z., van Velthoven, C., Smith, K., Tasic, B., Zeng, H., & Pachter, L. (2021). Isoform cell-type specificity in the mouse primary motor cortex. *Nature*, 598(7879), 195-199.
40. Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., ... & Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature methods*, 14(11), 1083-1086.
41. Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., ... & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561), 486-490.
42. Rotem, A., Ram, O., Shores, N., Sperling, R. A., Goren, A., Weitz, D. A., & Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology*, 33(11), 1165-1172.
43. Skene, P. J., & Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *elife*, 6, e21856.
44. Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., ... & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1), 1930.
45. Ku, W. L., Nakamura, K., Gao, W., Cui, K., Hu, G., Tang, Q., ... & Zhao, K. (2019). Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nature methods*, 16(4), 323-325.
46. Wu, S. J., Furlan, S. N., Mihalas, A. B., Kaya-Okur, H. S., Feroze, A. H., Emerson, S. N., ... & Patel, A. P. (2021). Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nature biotechnology*, 39(7), 819-824.
47. Bartosovic, M., Kabbe, M., & Castelo-Branco, G. (2021). Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nature biotechnology*, 39(7), 825-835.
48. Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., ... & Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8), 817-820.
49. Luo, C., Keown, C. L., Kurihara, L., Zhou, J., He, Y., Li, J., ... & Ecker, J. R. (2017). Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science*, 357(6351), 600-604.
50. Mulqueen, R. M., Pokholok, D., Norberg, S. J., Torkenczy, K. A., Fields, A. J., Sun, D., ... & Adey, A. C. (2018). Highly scalable generation of DNA methylation profiles in single cells. *Nature biotechnology*, 36(5), 428-431.

51. Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., ... & Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469), 59-64.
52. Tan, L., Xing, D., Chang, C. H., Li, H., & Xie, X. S. (2018). Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405), 924-928.
53. Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., ... & Shendure, J. (2017). Massively multiplex single-cell Hi-C. *Nature methods*, 14(3), 263-266.
54. Arrastia, M. V., Jachowicz, J. W., Ollikainen, N., Curtis, M. S., Lai, C., Quinodoz, S. A., ... & Guttman, M. (2022). Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nature biotechnology*, 40(1), 64-73.
55. Li, Y. E., Preissl, S., Hou, X., Zhang, Z., Zhang, K., Qiu, Y., ... & Ren, B. (2021). An atlas of gene regulatory elements in adult mouse cerebrum. *Nature*, 598(7879), 129-136.
56. Vandereyken, K., Sifrim, A., Thienpont, B., & Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. *Nature Reviews Genetics*, 24(8), 494-515.
57. Zhu, C., Zhang, Y., Li, Y. E., Lucero, J., Behrens, M. M., & Ren, B. (2021). Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature methods*, 18(3), 283-292.

Chapter 1 DROPLET-BASED SINGLE-CELL JOINT PROFILING OF HISTONE MODIFICATIONS AND TRANSCRIPTOMES

1.1 Abstract

We previously reported Paired-Tag, a combinatorial indexing-based method that can simultaneously map histone modifications and gene expression at single-cell resolution at scale. However, the lengthy procedure of Paired-Tag has hindered its general adoption in the community. To address this bottleneck, we developed a droplet-based Paired-Tag protocol that is faster and more accessible than the previous method. Using cultured mammalian cells and primary brain tissues, we demonstrate its superior performance at identifying candidate *cis*-regulatory elements and associating their dynamic chromatin state to target gene expression in each constituent cell type in a complex tissue.

1.2 Introduction

The chemical modifications to the histone proteins and nucleic acids along the chromosomes, collectively referred to as the cell's epigenome, regulate the spatiotemporal gene expression patterns in multicellular eukaryotic organisms¹. Epigenome analysis has been successfully used to reveal mechanisms of gene regulation and dysregulation in development, disease pathogenesis, and aging, however, considerable technical barriers exist in profiling the epigenome of complex tissues due to the heterogeneity of cell types and states within the biospecimens. Single-cell epigenomic techniques can circumvent this barrier by revealing the landscape of DNA methylation², chromosome conformation³, chromatin accessibility⁴, histone modifications⁵⁻⁷ and transcription factor binding⁸ at single-cell resolution. Single-cell multiomic

assays that jointly survey multiple molecular modalities including gene expression⁹⁻¹³ or protein abundance^{14,15} further help to decipher the complex interplay between the epigenome and transcriptional machinery. In particular, single-cell CUT&Tag has enabled the characterization of active or silenced chromatin states at candidate *cis*-regulatory elements within different cell types of primary tissues^{16,17}. However, current single-cell epigenomics assays have been slow for general adoption, owing to factors such as lengthy procedures and lack of general accessibility.

Here, we report Droplet Paired-Tag, a fast and broadly accessible technique for producing high-quality single-cell joint profiles of histone modification and transcriptome in parallel, which has the potential to be quickly adopted by the research community for interrogating the dynamic and cell-type-specific epigenome landscapes in complex tissues. The key modifications compared to the initial combinatorial indexing-based Paired-Tag protocol are the adaptation of a commercially available microfluidics platform (i.e. 10x Chromium Single Cell Multiome) to introduce cellular barcodes, and a much-simplified protocol to prepare sequencing libraries. The new procedure, therefore, offers three key advantages. First, changing to a droplet-based platform greatly shortens the hands-on time in both the molecular barcoding and library preparation steps (Extended Data Fig. 1.1). As a result, Droplet Paired-Tag can be performed in less than 1.5 days from nuclei preparation to sequencing library construction, considerably shorter than the conventional 3-day-long Paired-Tag procedure. Second, this design can also be more easily adapted owing to the wide availability of the commercial 10X Chromium platform and reagent kits. Third, the simplified procedure also brings improved performance in the identification of *cis*-regulatory elements and correlating chromatin states of distal elements to expression levels of putative target genes.

1.3 Results

In Droplet Paired-Tag, nuclei are first permeabilized, followed by targeted tagmentation using primary antibodies against a histone modification that are pre-coupled with protein A-Tn5 transposase fusion proteins (pA-Tn5), in a procedure modified from the reported CUT&Tag¹⁸ method (See Methods). The resulting nuclei and barcoded beads are then co-encapsulated into droplets with a microfluidic device (10x Genomics Chromium X controller). Two types of oligos with the same set of barcodes are embedded in the beads: a barcoded poly-dT oligo to label cDNA and a capturing oligo to label DNA fragments derived from tagmentation. Reverse transcription is carried out in the droplets with the barcoded poly-dT oligo. A ligation reaction is simultaneously performed to attach barcoded capturing oligo to the tagmented chromatin fragments. To facilitate this ligation process with our in-house protein A-Tn5 fusion (pA-Tn5), we designed the pA-Tn5 adaptor with 3'-extended bridge-pMENTS sequence reverse complementary with the capturing oligo sequence (Fig. 1.1a, Extended Data Fig. 1.2a). Upon completion of the reactions, the reverse transcription products as well as tagmented DNA fragments from the same cells are both labeled with the same unique cellular barcodes. The droplets are then dissolved, and the cDNA and chromatin fragments are purified, amplified, and then split for sequencing library construction by following manufacturer-recommended protocols (10x Chromium Single Cell Multiome) (Extended Data Fig. 1.2b-d).

As a proof-of-concept, we first used Droplet Paired-Tag to analyze a histone modification (either H3K27ac or H3K27me3) jointly with transcriptome in individual mouse embryonic stem cells (mESCs). The complexity of the DNA-dedicated libraries corresponding to histone modifications (median number of 1,448 and 3,224 fragments per nucleus for H3K27ac and H3K27me3, respectively) was comparable to those of the combinatorial indexing-based Paired-

Tag. The library complexity was also similar to or higher than those from other published methods for analyzing histone modification in single cells such as the scCUT&Tag⁷, CoTECH and scCUT&Tag-pro (Fig. 1.1b). Compared to H3K27ac, a higher fraction of reads from H3K27me3 Droplet Paired-Tag experiments correspond to mono-, di- and tri-nucleosomes fragments (Extended Data Fig. 1.3b), consistent with the more compact chromatin structures within the Polycomb complex repressed regions¹⁹. For transcriptome analysis, Droplet Paired-Tag yielded gene expression dedicating libraries with comparable complexity as the other commonly used scRNA-seq platforms (Fig. 1.1c-d). The gene expression profile detected by Droplet Paired-Tag was in excellent agreement with bulk RNA-seq data from mouse ES cells (Extended Data Fig. 1.3c). As expected, the detected gene expression levels are in general positively correlated with the H3K27ac signal over the transcription start site (TSS), and inversely correlated with H3K27me3 deposition across gene bodies (Extended Data Fig. 1.3d). These results indicate that Droplet Paired-Tag can reliably capture histone modification and gene expression simultaneously from the same cell in a high throughput fashion.

To evaluate the sensitivity and specificity of Droplet Paired-Tag in histone modification profiling, we compared the mESC single-cell data with bulk CUT&Tag and public ChIP-seq datasets generated from the mouse ES cells^{20,21}. For both H3K27ac and H3K27me3, the aggregated single-cell signals faithfully resembled those from bulk CUT&Tag and ChIP-seq experiments (Fig. 1.1e), and showed high enrichment over peaks identified from ChIP-seq datasets (Extended Data Fig. 1.3e). As an example, single-cell H3K27ac reads from Droplet Paired-Tag (Fig. 1.1f) marked the promoter region of the pluripotent gene *Sox2* and its downstream super-enhancer, while H3K27me3 reads are deposited on genes involved in cellular differentiation and expressed in specific cell lineages (e.g., *Gata4*). 72% of the peaks identified from aggregated single-cell signals

from Droplet Paired-Tag overlap with those from ChIP-seq experiments (Fig. 1.1g). By down-sampling the number of nuclei profiled in Droplet Paired-Tag experiments, we found that the number of H3K27ac peaks detected reaches saturation after about 2,500 nuclei, or around 6 million total reads. By comparison, Paired-Tag dataset requires 60% more reads to reach saturation (Fig. 1.1h, Extended Data Fig. 1.3f).

To demonstrate the utility of Droplet Paired-Tag for single-cell epigenomic analysis of primary tissues, we used it to analyze histone modifications and gene expression at single-cell resolution in the adult mouse frontal cortex. We performed Droplet Paired-Tag experiments to profile either H3K27ac or H3K27me3, together with RNA expression, each in three biological replicates. After filtering out low-quality cells and potential doublets, we recovered 22,054 nuclei in total, of which 11,874 nuclei were profiled for H3K27ac and 10,180 for H3K27me3. Clustering of these nuclei based on their transcriptome profiles identified 20 major cell clusters corresponding to 9 glutamatergic neuron types (*Snap25+/*Slc17a7*+*), 6 GABAergic neuron types (*Snap25+/*Gad1*+*) and 5 non-neuron cell types (*Snap25-*) (Fig. 1.2a, Extended Data Fig. 1.4a-d). Most cell types were known to exist in the frontal cortex regions, except for three cell types found in one sample, namely D12MSN, OBGA, OBGL, that likely originate from anatomical regions proximal to the frontal cortex due to variations in surgical sectioning during the sample preparation (Extended Data Fig. 1.4b). Interestingly, out of the 20 major cell clusters, 17 and 18 can also be recovered by independent clustering using histone modification H3K27ac or H3K27me3 signals (Fig. 1.2a), respectively, suggesting that cell type annotations are highly concordant between transcriptome- and epigenome-based clustering (Fig. 1.2b, Extended Data Fig. 1.5a-d). Additionally, the cell clusters reported in this study were also consistent with those from the

previous Paired-Tag dataset and the BICCN reference 10X single nucleus RNA-seq dataset generated from the mouse primary motor cortex (Extended Data Fig. 1.5e-h)²².

To jointly analyze Droplet Paired-Tag data corresponding to different histone modifications, we used transcriptome-based clustering and cell type annotation in all subsequent analyses. For histone modality, pseudo-bulk level signals showed high concordance with both bulk CUT&Tag and ChIP-seq experiments (Extended Data Fig. 1.6a-b). Pseudo-bulk single-cell histone signals from cells within each cluster showed that H3K27ac signal is abundant at transcription start sites of cell-type specific genes, whereas these regions are generally silenced by H3K27me3 in other cell types (Fig. 1.2c, Extended Data Fig. 1.6c, Extended Data Fig. 1.7a-d). Compared with scCUT&Tag⁶, Droplet Paired-Tag yielded a comparable fraction of reads in peaks (FRiP) but higher numbers of unique fragments per nuclei for both histone modifications. Compared with the combinatorial-indexing-based Paired-Tag¹², Droplet Paired-Tag recovered fewer unique reads per nuclei, but showed higher FRiP and thus captured a higher number of peak-associated reads. The improvements in both signal sensitivity and specificity likely contributed to the higher resolution in separating cell types (Fig. 1.2e, f, Extended Data Fig. 1.6f-i). Compared to the list of open chromatin regions identified in the same brain cell types from a recent snATAC-seq atlas (the Center for Epigenomic of Mouse Brain Atlas, CEMBA)²³, Droplet Paired-Tag yielded the lowest level of H3K27me3 signals at the open chromatin regions, indicating minimal off-target Tn5 transposase activities in our procedure (Fig. 1.2g). To evaluate the sensitivity of Droplet Paired-Tag, we identified the peaks of H3K27ac signals in each cell cluster and retained those that appeared in two or more replicates. The resulting union set of 63,799 peaks was two times more than those detected in the previous combinatorial-indexing-based Paired-Tag (27,522) from a similar number of nuclei (11,874 versus 11,749) (Extended Data Fig. 1.6d). A higher

fraction of H3K27ac peaks detected in Droplet Paired-Tag overlapped with the open chromatin regions from the same brain cell types than the previous Paired-Tag dataset (90.8% vs. 80.5%) (Fig. 1.2h). Taken together, these results suggest that Droplet Paired-Tag can generate high quality transcriptomic and epigenomic profiles at single-cell resolution from complex tissues.

To further demonstrate the utility of Droplet Paired-Tag in characterizing candidate *cis*-regulatory elements (cCREs) activity states, we examined variations of chromatin states (H3K27ac or H3K27me3) at identified cCREs in different brain cell types²³. We classified cCREs as distal or proximal based on their distance to promoter regions (see Methods), then performed non-negative matrix factorization to group all the cCREs with sufficient levels of H3K27ac (RPKM > 1) and H3K27me3 (RPKM > 1) signals in at least one brain cell type into 20 cCRE modules, each showing a distinct pattern of cell-type specificity (Fig. 1.2i-j, Extended Data Fig. 1.8a). For proximal cCREs, promoter region H3K27ac signal again showed a strong positive correlation with transcription, while H3K27me3 signal showed an overall inverse correlation (Fig. 1.2i). Transcriptional factor (TF) binding motif analysis of each CRE module revealed known transcription factors involved. For example, cCRE module corresponding to the ITL23GL cluster (excitatory neurons from cortex layer 2/3) was enriched for NEUROD1 and MEF2C motifs, and the OGC cCRE module was enriched for motifs of oligodendrocytes-critical transcriptional factors like SOX10 (Fig. 1.2k, Extended Data Fig. 1.8b).

Droplet Paired-Tag data enables the prediction of putative target genes of distal cCREs due to the joint profiling of gene expression level and chromatin state from the same cells. We calculated the pairwise Spearman's correlation coefficients (SCC) between the histone modification signals at distal cCREs and promoter regions of potential target genes within a 500kb window (Fig. 1.2l). This analysis identified 20,241 significant cCREs-gene pairs with a positive

correlation between H3K27ac signals and gene expression, and 4,738 pairs with a negative correlation between H3K27me3 and gene expression (FDR < 0.05, Fig. 1.2m). Interestingly, Droplet Paired-Tag data captured stronger linkages between cCREs and genes over background than the combinatorial indexing Paired-Tag (Extended Data Fig. 1.8c). Gene ontology (GO) analysis showed that H3K27ac-associated genes in OGC population were enriched for terms related to myelination process, consistent with the likely enhancer function of the distal cCREs. On the other hand, the H3K27me3-associated genes in the same cell type are enriched for terms related to neuronal functions (Fig. 1.2n). Through integrative analysis with a recently published mouse brain single-cell chromosome contacts dataset²⁴, we found that cell-type specific cCREs-gene pairs exhibiting long-range chromatin contacts were overall more positively (for H3K27ac) correlated than cCRE-gene pairs showing no detectable chromatin contacts (Extended Data Fig. 1.8d-e).

1.4 Discussion

In summary, Droplet Paired-Tag is a fast and robust method for joint profiling of histone modification and gene expression in single cells. We demonstrated the utility and superior performance of this method for analyzing cell-type-specific gene regulatory programs in complex tissue. The shortened, more easily adaptable procedure due to the use of a widely available microfluidic device (i.e., 10x Genomics Chromium) would likely facilitate the quick adaption of this method in the field of epigenetics. Droplet Paired-Tag adds a new toolkit for investigation of the gene regulatory mechanisms in disease and lifespan.

1.5 Acknowledgement

Chapter 1 is, in full, a reformatted reprint of the material as it appears as “Droplet-based single-cell joint profiling of histone modifications and transcriptomes” in *Nature Structural & Molecular Biology*, 2023. Yang Xie, Chenxu Zhu, Zhaoning Wang, Melodi Tastemel, Lei Chang, Yang Eric Li and Bing Ren. The dissertation author was the primary investigator and author of this paper.

We thank K. Zhang and E. Armand for discussion and suggestions for data analysis. We thank QB3 MacroLab for the Protein A-Tn5 fusion protein production. This study was funded by grant nos. R41MH128993 and the Ludwig Institute for Cancer Research (to B.R.); grant nos. 4R00HG011483 and 5RM1HG011014 (to C.Z.). Z.W. is a DDBrown Awardee of the Life Science Research Foundation.

1.6 Figures

Figure 1.1: Joint profiling of transcriptomes and histone modifications in single cells using Droplet Paired-Tag.

a, Schematic overview of Droplet Paired-Tag; TSO, template switch oligo. **b**, Distribution of the unique number of fragments per cell across different single-cell epigenomic assays with the indicated samples and histone targets; $n = 13,664$ (Droplet Paired-Tag, H3K27ac), 4,501 (Droplet Paired-Tag, H3K27me3), 2,066 (Paired-Tag, H3K27ac), 418 (Paired-Tag, H3K27me3), 3,031 (CoTECH, H3K27me3), 5,111 (scCUT&Tag, H3K27ac), 9,039 (scCUT&Tag-pro, H3K27ac) and 12,585 (scCUT&Tag-pro, H3K27me3); PBMCs, peripheral blood mononuclear cells. **c,d**, Distribution of unique transcript (**c**) or gene (**d**) numbers per cell across different single-cell multiomic assays. For all box plots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median and whiskers denoting a maximum $2\times$ the interquartile range; $n = 13,664$ (Droplet Paired-Tag, H3K27ac), 4,501 (Droplet Paired-Tag, H3K27me3), 2,066 (Paired-Tag, H3K27ac), 418 (Paired-Tag, H3K27me3), 3,031 (CoTECH, H3K27me3) and 2,711 (10x Multiome, ATAC). **e**, Genome-wide SCCs between mESC histone modification datasets from Droplet Paired-Tag (indicated with 'sc'), bulk CUT&Tag and ChIP-seq (indicated with 'bulk'); rep, replicate. **f**, Genome browser view showing examples of pseudo-bulk and single-cell histone modification signals of Droplet Paired-Tag in mESCs, along with bulk ChIP-seq and CUT&Tag. The pluripotent gene (*Sox2*), along with its superenhancer (SCR), and the repressed gene (*Gata4*) are highlighted in pink; chr, chromosome. **g**, Number and overlap of peaks called using H3K27ac Droplet Paired-Tag compared to those from ChIP-seq datasets. **h**, Scatter plot showing the relationships between the total number of H3K27ac peaks called and the number of nuclei after downsampling. The dashed line indicates the cutoff of the number of nuclei that could recover 85% of peaks.

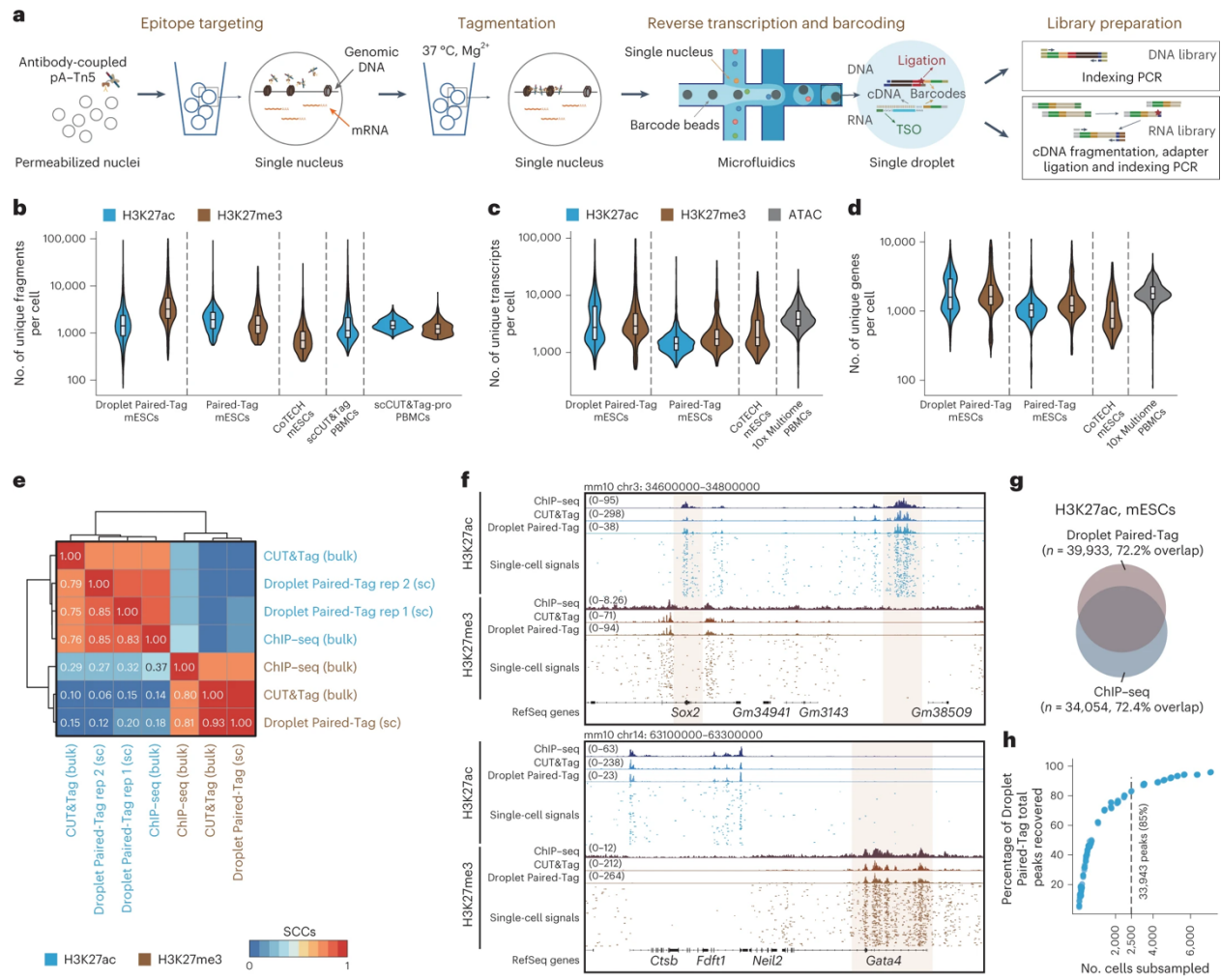
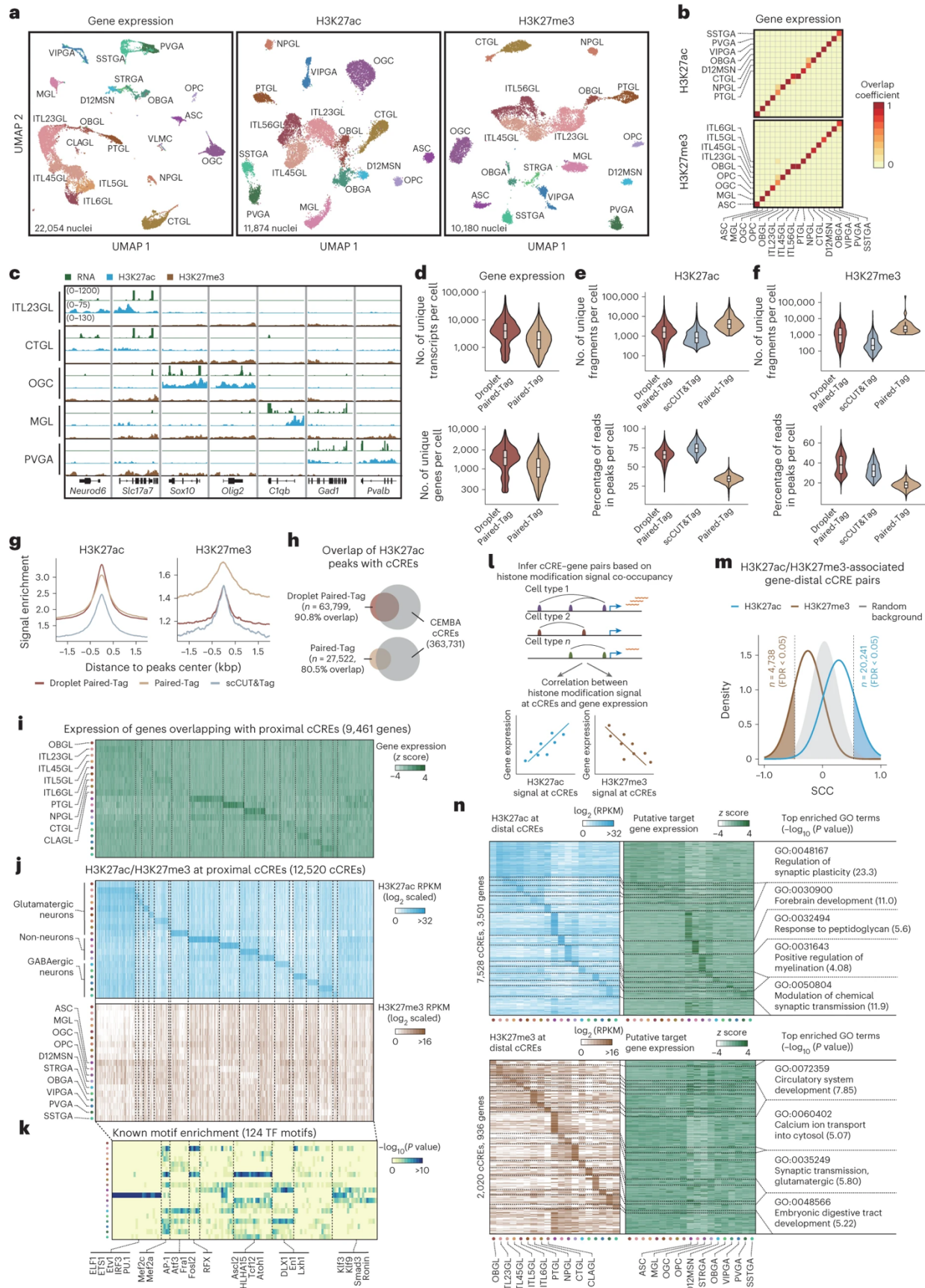
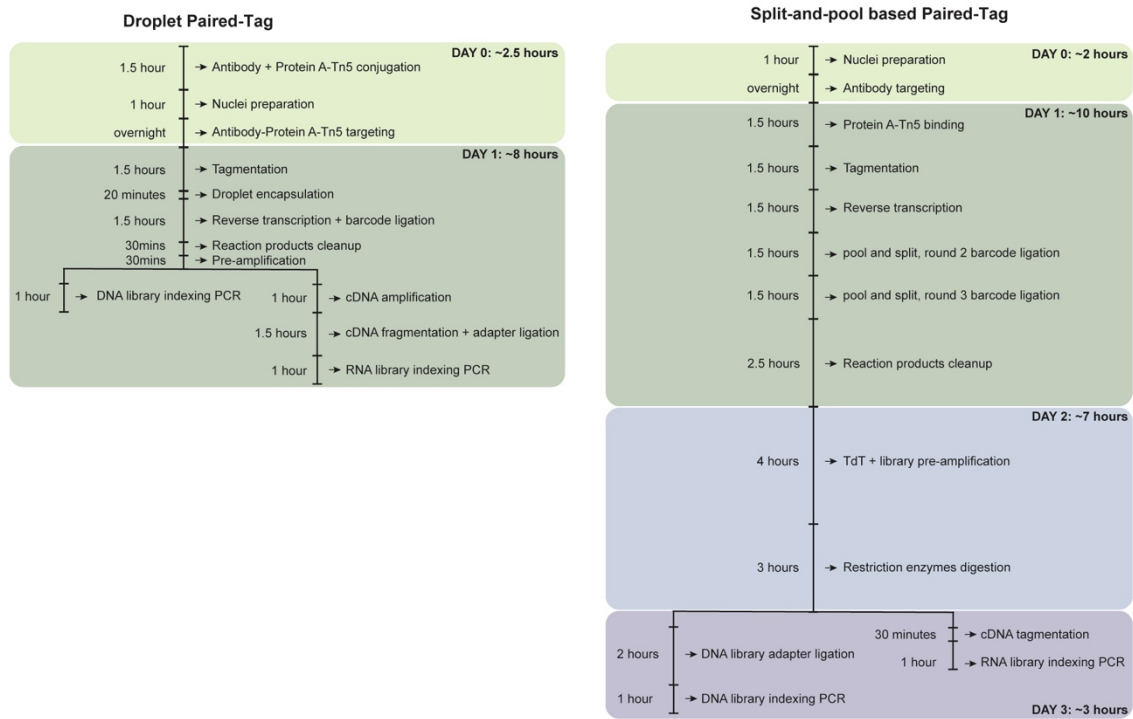


Figure 1.2: Droplet Paired-Tag effectively resolves multiple cell types in the mouse frontal cortex (FC) and identifies the cCREs within each cell type.

a, Uniform manifold approximation and projection (UMAP) embedding visualization of frontal cortex Droplet Paired-Tag data clustered and annotated based on the transcriptome (gene expression) and histone modifications (H3K27ac and H3K27me3). **b**, Overlap of shared annotations between transcriptome and epigenome clustering. **c**, Representative genome browser view of gene expression and H3K27ac and H3K27me3 distribution over cell-type-specific marker genes. **d**, Comparison of the number of unique transcripts and genes detected in each cell between Droplet Paired-Tag and Paired-Tag; $n = 22,054$ (Droplet Paired-Tag, RNA) and $11,026$ (Paired-Tag, RNA). **e,f**, Comparison of the unique fragments and FRiPs in each cell between Droplet Paired-Tag and other methods measuring the histone modifications H3K27ac (**e**) and H3K27me3 (**f**) in single cells. All box plot hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median and whiskers denoting $2 \times$ the interquartile range; $n = 11,874$ (Droplet Paired-Tag, H3K27ac), $5,049$ (scCUT&Tag, H3K27ac), 886 (Paired-Tag, H3K27ac), $10,180$ (Droplet Paired-Tag, H3K27me3), $4,019$ (scCUT&Tag, H3K27me3) and 60 (Paired-Tag, H3K27me3). **g**, Signal enrichment over CEMBA cCREs in Droplet Paired-Tag and other methods measuring single-cell histone modifications; kbp, kilobase pairs. **h**, Comparison of the number of H3K27ac peaks from the Droplet Paired-Tag dataset with the original Paired-Tag dataset, intersected with CEMBA-identified cCREs. **i**, Heat map showing gene expression values from promoter-proximal cCREs. **j**, Signals of both histone modifications over promoter-proximal cCREs across different cell types. **k**, Heat map of known motif enrichment for each cCRE module of promoter-proximal cCREs. Examples of known motifs are shown along with the heat map. P values were calculated by one-sided binomial test. FDRs were then calculated to select enriched motifs; TF, transcription factor. **l**, Schematics for identifying potential targets for cCREs. **m**, Frequency density plots showing the distribution of SCCs between distal cCRE histone modification signals and their putative target gene expression level. Cutoffs (FDR = 0.05) used to identify cell-type-specific cCRE–gene pairs are also indicated. **n**, Heat map showing histone modification signals at distal cCREs with potential active or repressive roles and their putative target gene expression levels. Example overrepresented GO terms for genes in selected cell types are shown. P values were calculated by Fisher’s exact test. Benjamini–Hochberg FDRs were then calculated to select overrepresented GO terms.



1.7 Supplementary Figures

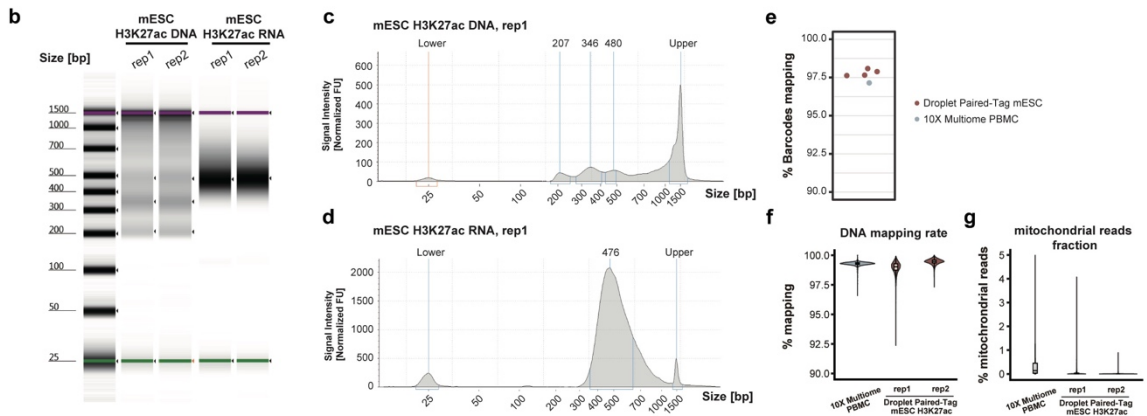


Extended Data Figure 1.1: Comparison of the workflows of Droplet Paired-Tag and combinatorial-indexing-based Paired-Tag.

Detailed workflow describing the end-to-end procedure of Droplet Paired-Tag (left) is compared to the conventional combinatorial-indexing-based version of Paired-Tag (right).

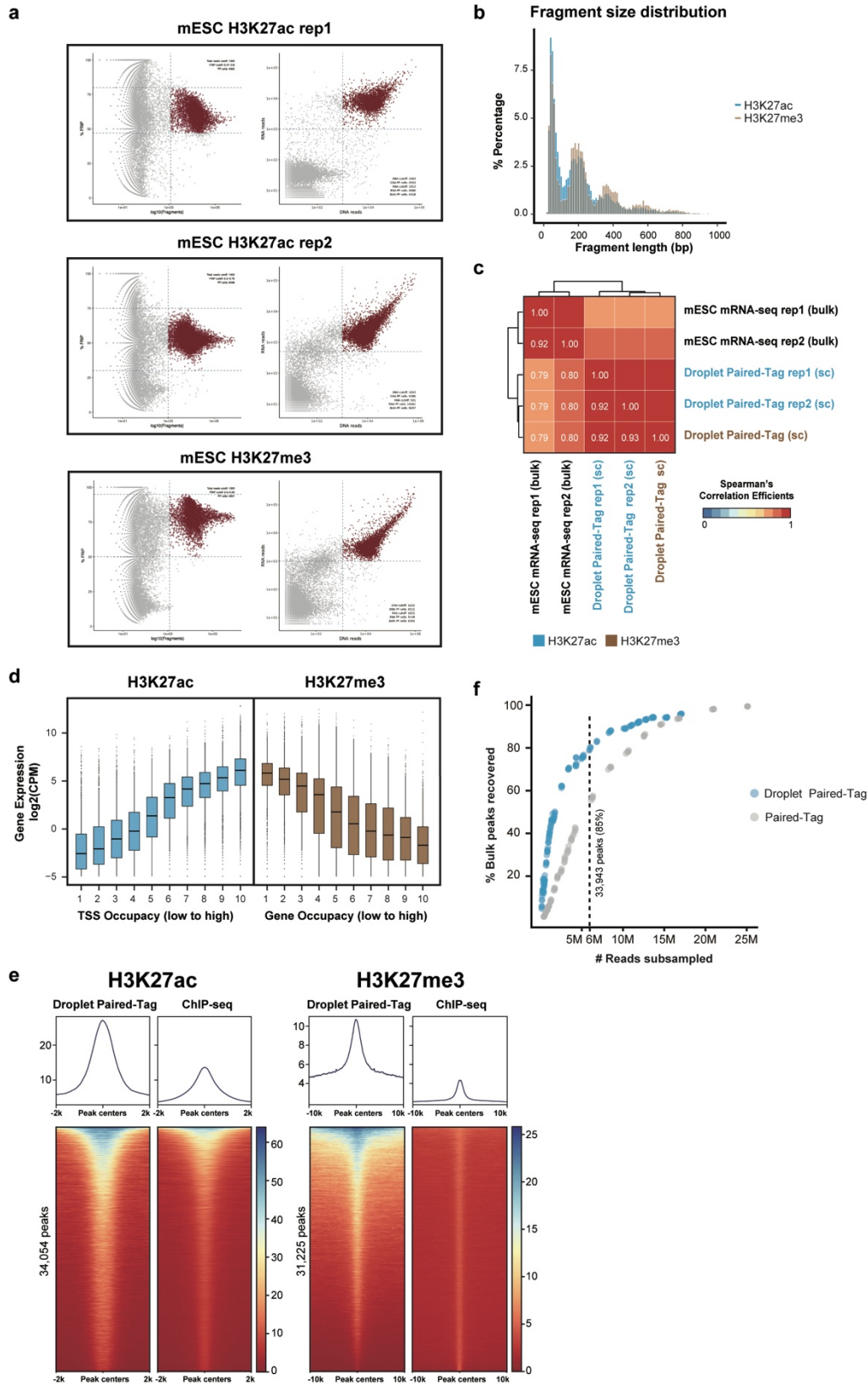
Extended Data Figure 1.2: Experimental design and quality control metrics of Droplet Paired-Tag.

a, Schematics for Droplet Paired-Tag barcoding and library preparation process. Modification on oligos is not shown for simplicity. **b–d**, Example capillary electrophoretic gel images from a TapeStation showing final libraries size distribution for Droplet Paired-Tag DNA (**c**) and RNA (**d**). All replicates are shown in **b**. **e**, Barcode mapping rate for Droplet Paired-Tag and standard 10X Multiome experiment. **f,g**, Sequence mapping rate (**f**) and fraction of mitochondrial reads (**g**) for Droplet Paired-Tag DNA libraries. For all boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 2 interquartile range (IQR). n = 4,302 (Droplet Paired-Tag, H3K27ac, replicate#1), 9,346 (Droplet Paired-Tag, H3K27ac, replicate#2), 2,711 (10X Multiome, ATAC).

[illegible]

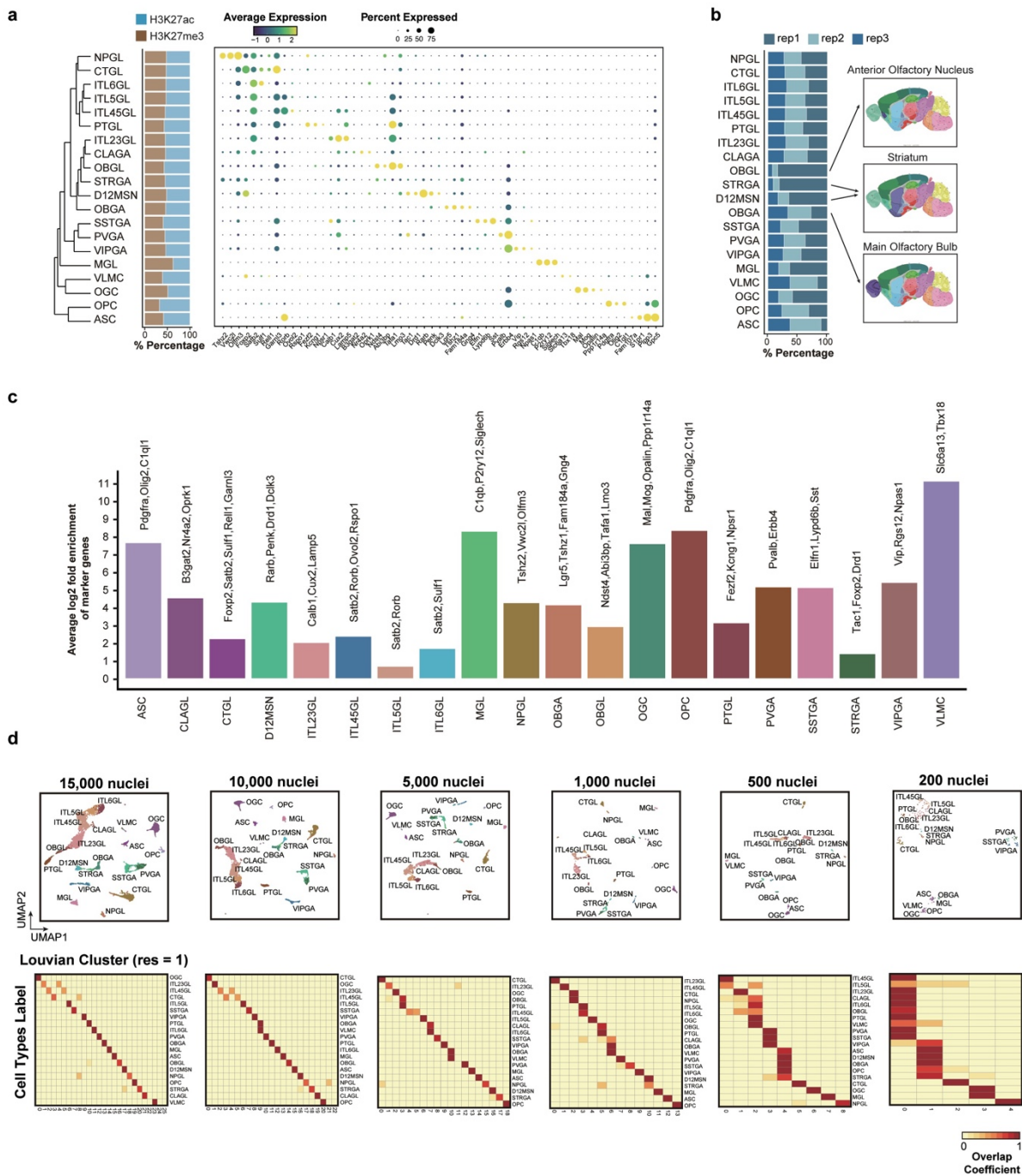
Extended Data Figure 1.3: Quality control metrics of mESC Droplet Paired-Tag data.

a, Strategies of valid nuclei selection. DNA barcodes were filtered based on total reads per nuclei and fraction of reads in peak regions. Valid nuclei were further selected based on the pairing of valid DNA and RNA nuclei in the scatterplot of total reads per nuclei (right). Cutoffs were set by manual inspection and depicted as dash lines and are also annotated inside the scatterplot. **b**, Distribution for fragment lengths of the sequenced fragments from the H3K27ac and H3K27me3 Droplet Paired-Tag experiments. **c**, Heatmap showing pairwise Spearman's correlation coefficients of expression profiles from single-cell Droplet Paired-Tag experiment and with the bulk mRNA-seq datasets. **d**, Boxplots showing the expression level of genes grouped by histone marks occupancy at their TSS regions (H3K27ac) or gene bodies (H3K27me3). For all boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 1 interquartile range (IQR), outlier indicated with dots. For H3K27ac group 1-10, gene number $n = 22,456$ (2,246 each quantile group except group 10); for H3K27me3 group 1-10, $n = 22,969$ (2,297 each quantile group except group 10). **e**, Enrichment of histone modification signals over ChIP-seq peaks, compared between Droplet Paired-Tag and ChIP-seq data. **f**, Scatter plot showing the relationships between total number of H3K27ac peaks called and the number of reads from Droplet Paired-Tag or Paired-Tag dataset in a down-sampling test, similar to Fig. 1.1g. Dashed line indicates 85% peaks recovered for Droplet Paired-Tag dataset.



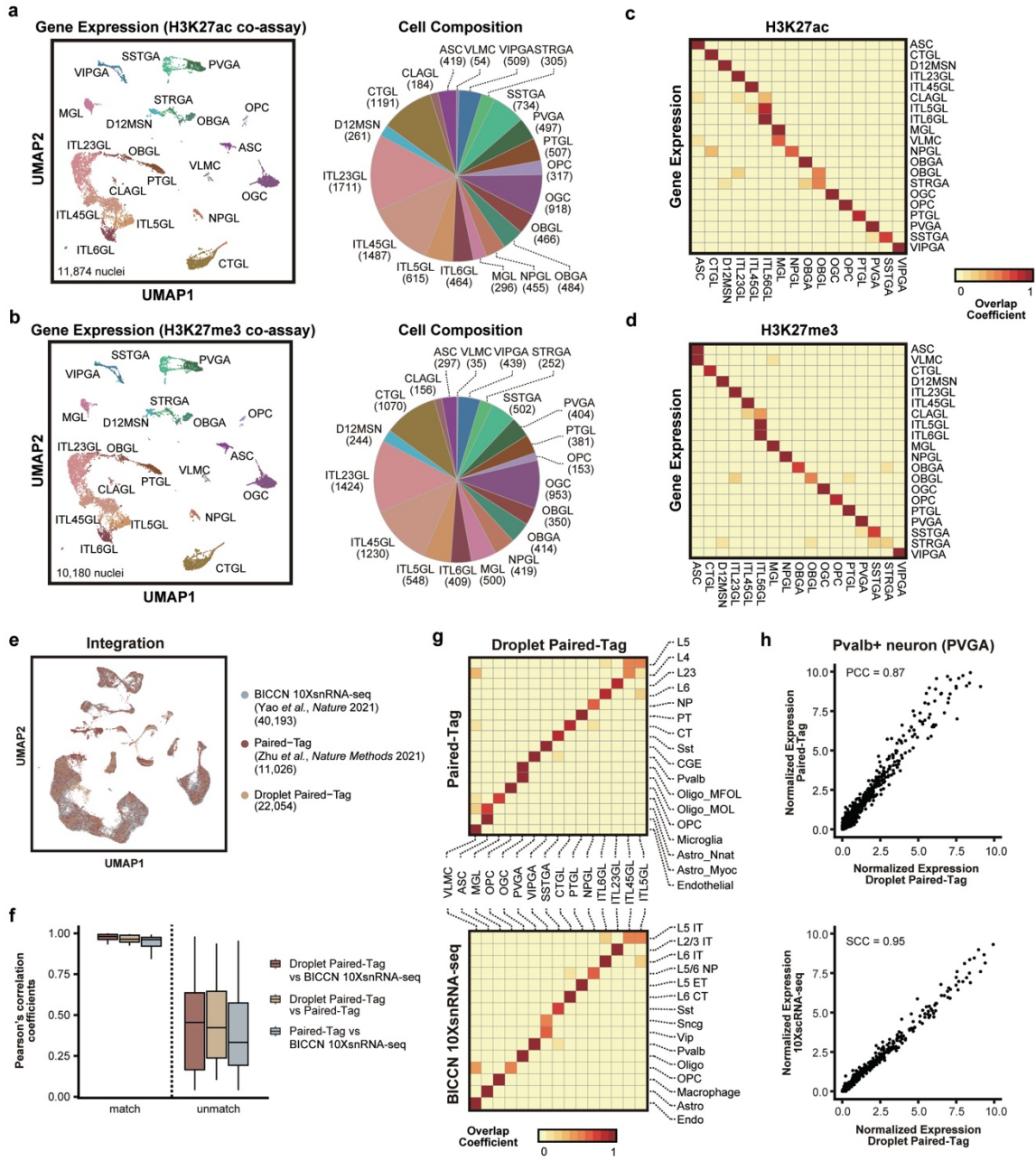
Extended Data Figure 1.4: Annotation of cell clusters in the mouse frontal cortex Droplet Paired-Tag based on transcriptomic profiles.

a, Expression of marker genes in 20 mouse brain cell types, and the fraction of nuclei by set of experiments. **b**, Fraction of nuclei in each cell type by replicates. Cell types with biased distribution in any of the replicates were from anatomical regions (right) proximal to frontal cortex. **c**, Averaged Log₂ fold enrichment ratio of the normalized selected marker genes expression (RPKM) in cluster of interest versus the rest of the dataset. **d**, UMAP embeddings and overlap scores based on Droplet Paired-Tag transcriptome profiles down-sampled to different nuclei numbers.



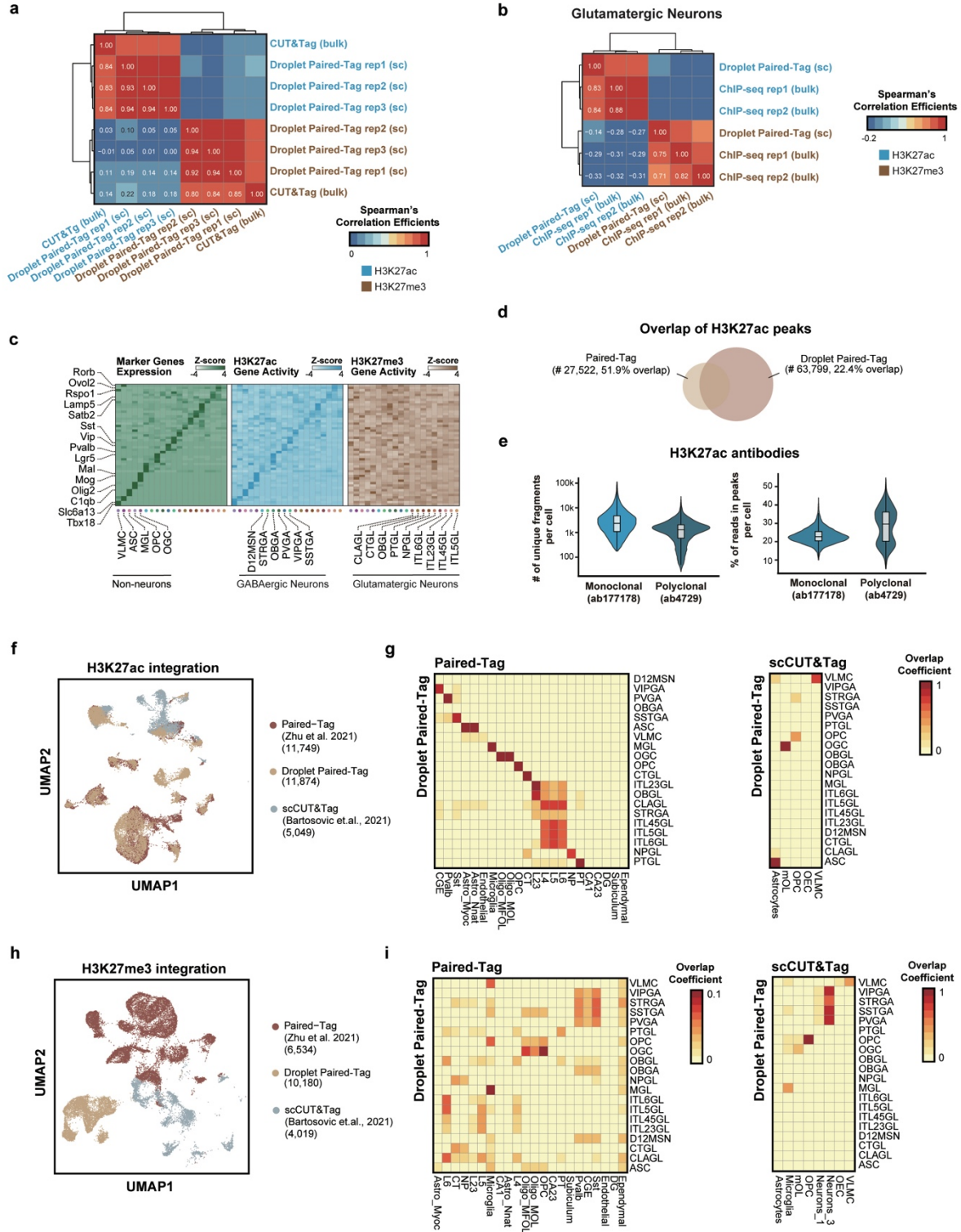
Extended Data Figure 1.5: Integrative analysis of Droplet Paired-Tag transcriptomic profiles with public datasets.

a-b, UMAP embedding, cell type compositions of Droplet Paired-Tag transcriptome profiles based on histone modification co-profiled. **c-d**, Overlap of all annotations between transcriptomic and epigenomic clustering. **e**, UMAP co-embedding of single nuclei transcriptomic profile from Droplet Paired-Tag, Paired-Tag and reference BICCN 10X snRNA-seq datasets on mouse motor cortex regions. **f**, Boxplot showing Pearson correlation coefficients of variable genes for matched and unmatched cell types between Droplet Paired-Tag, Paired-Tag and reference 10X snRNA-seq datasets. For all boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 1.5 interquartile range (IQR). Number of matched cell types: $n = 16$ (Droplet Paired-Tag vs snRNA-seq), 18 (Droplet Paired-Tag vs Paired-Tag), 16 (Paired-Tag vs snRNA-seq). Number of unmatched cell types: $n = 304$ (Droplet Paired-Tag vs snRNA-seq), 322 (Droplet Paired-Tag vs Paired-Tag), 256 (Paired-Tag vs snRNA-seq) **g**, Overlap of shared annotations between Droplet Paired-Tag and the previously published Paired-Tag transcriptomic clustering results. Cell types not from frontal cortex are excluded in comparison. **h**, Scatterplots showing expression levels of variable genes in *Pvalb*⁺ neurons (PVGA) in Droplet Paired-Tag, Paired-Tag and reference 10X snRNA-seq datasets.



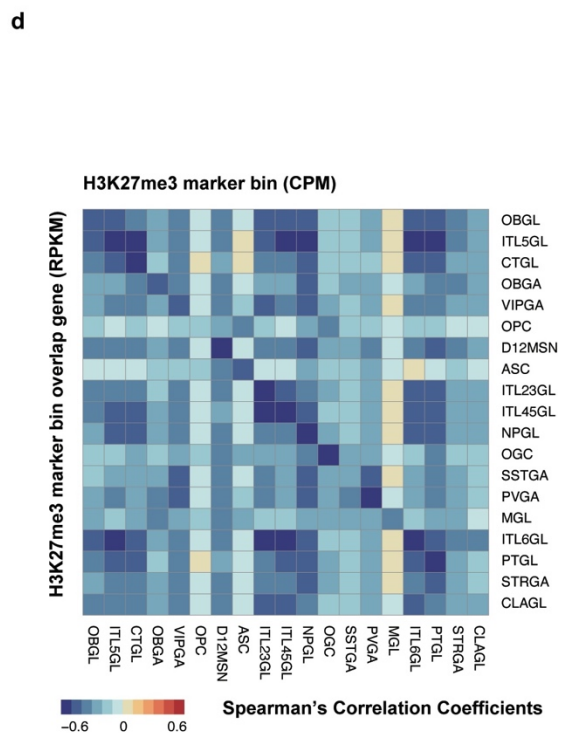
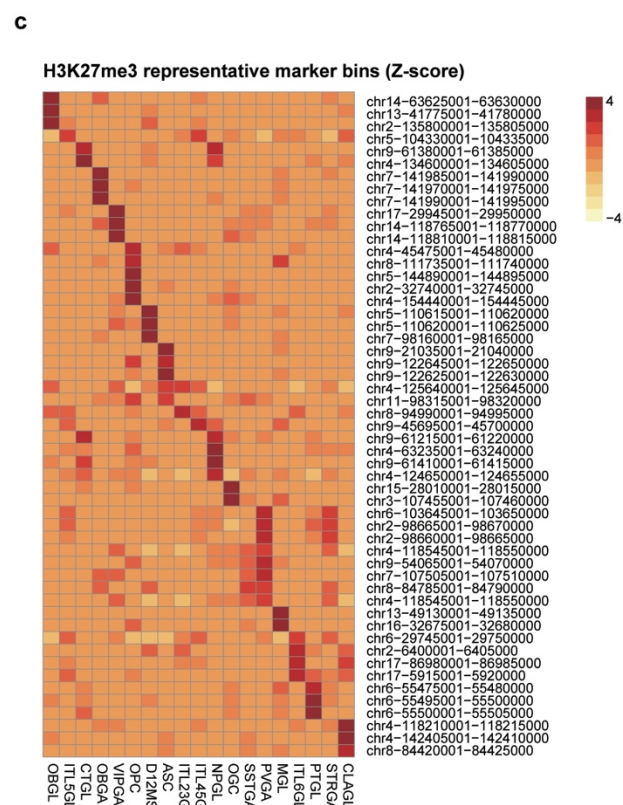
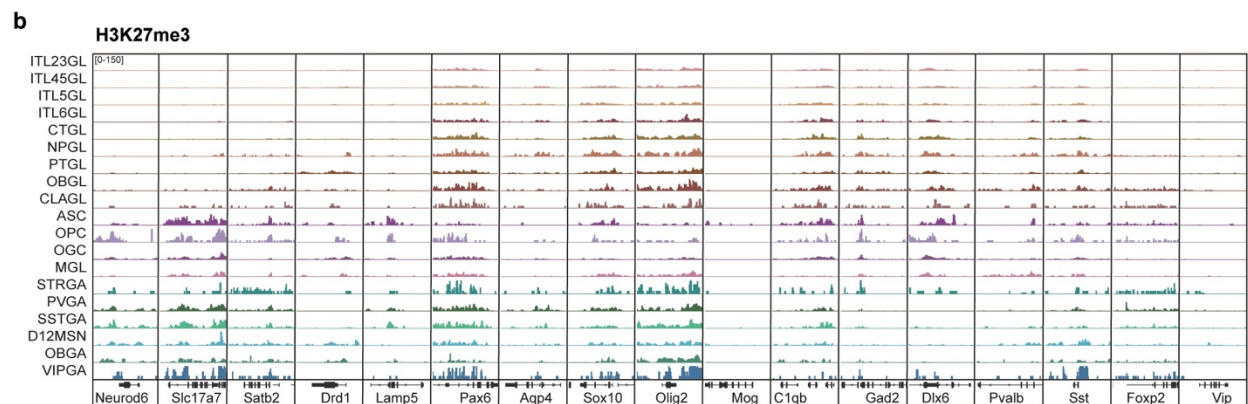
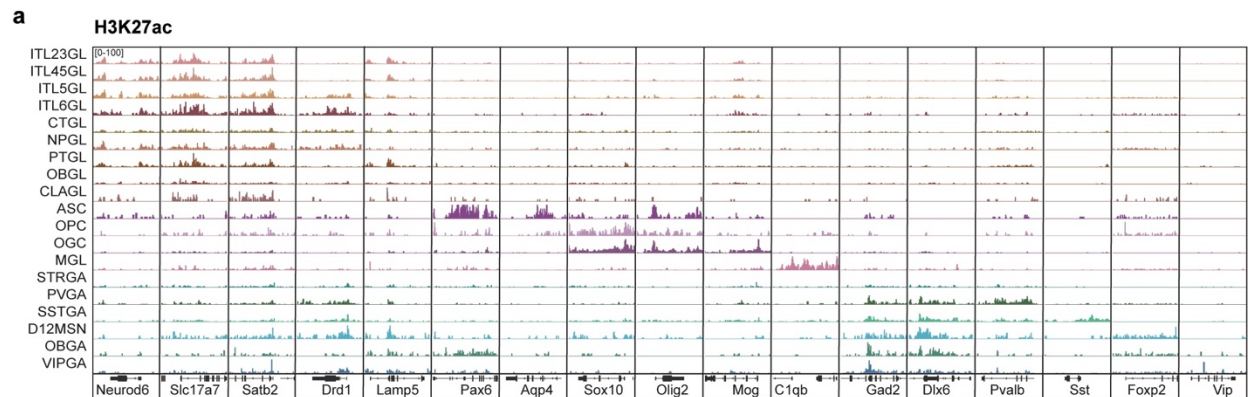
Extended Data Figure 1.6: Quality control of mouse frontal cortex Droplet Paired-Tag histone modifications profiles.

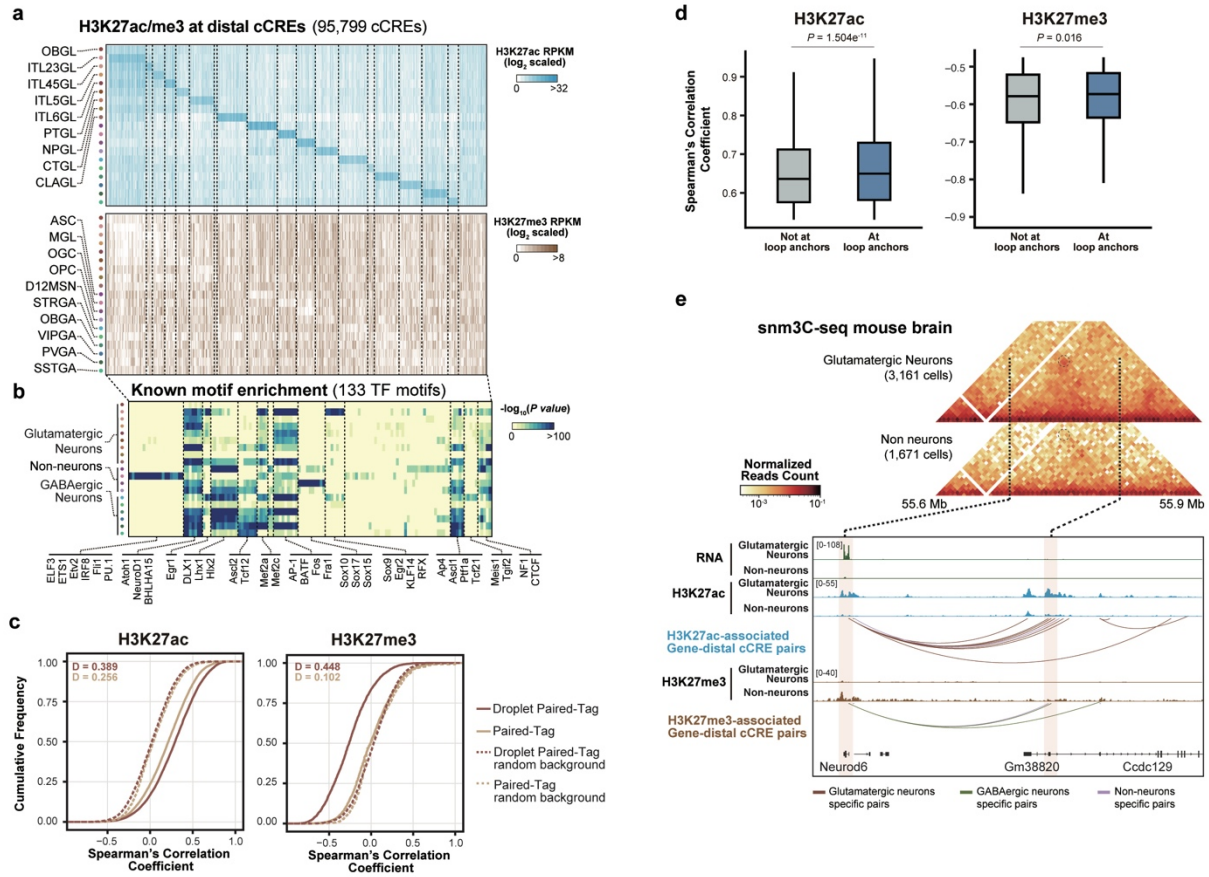
a, Genome-wise Spearman's correlation coefficients between mouse frontal cortex histone modification datasets from single-cell Droplet Paired-Tag and bulk CUT&Tag. **b**, Genome-wise Spearman's correlation coefficients between mouse frontal cortex glutamatergic neurons histone modification datasets from single-cell Droplet Paired-Tag and bulk ChIP-seq. **c**, Heatmap showing marker genes expression level and promoters / gene bodies histone modifications signal in each cell type. Examples of well-known marker genes are shown. **d**, Overlap of the number of H3K27ac peaks called from Droplet Paired-Tag and Paired-Tag datasets. Both datasets are down-sampled to similar number of cells. **e**, Comparison of the number of unique fragments and FRiP in single cell between monoclonal and polyclonal H3K27ac antibodies used. For all boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 2 interquartile range (IQR). $n = 8,807$ (monoclonal antibody), $10,069$ (polyclonal antibody). **f**, UMAP co-embedding of single nuclei H3K27ac profile from Droplet Paired-Tag, Paired-Tag and scCUT&Tag. **g**, Overlap of shared annotations between Droplet Paired-Tag and public single-cell H3K27ac datasets. **h**, UMAP co-embedding of single nuclei H3K27me3 profile from Droplet Paired-Tag, Paired-Tag and scCUT&Tag. **i**, Overlap of shared annotations between Droplet Paired-Tag and public single-cell H3K27ac datasets.



Extended Data Figure 1.7: Landscape of histone modifications across cell types in mouse frontal cortex.

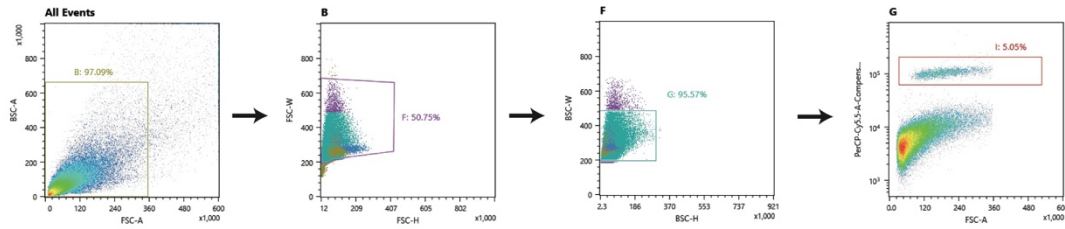
a-b, Representative Genome browser view showing H3K27ac (**a**) or H3K27me3 (**b**) signal over marker genes in each mouse brain cell type. **c**, Top 10 H3K27me3 marker bins from each cell type with >100 nuclei profiled. **d**, Spearman's correlation coefficients between cell-type specific marker bins H3K27me3 signal (CPM) and bin-overlapped genes expression level (RPKM) from different cell types.





Extended Data Figure 1.8: Integrative analysis of histone modifications at candidate *cis*-regulatory elements across cell types in mouse frontal cortex.

a, Heatmap showing H3K27ac and H3K27me3 signals at the distal cCREs across different cell types. **b**, Heatmap of known motifs enrichment for each cCRE module of distal cCREs. Examples of known motifs are shown along with the heatmap. P value: one-sided binomial test. False Discovery Rate is then calculated to select enriched motifs. **c**, Cumulative distribution function plot of Spearman's correlation coefficients between distal cCREs histone modification signal and putative target genes expression level. Distributions from both Droplet Paired-Tag and the Paired-Tag are shown, and the greatest separation between real and random background (D) for each dataset is annotated inside the plot. **d**, boxplot showing Spearman's correlation coefficients of H3K27ac- or H3K27me3- associated distal cCREs-gene pairs intersected at or not at loop anchors identified in snm3C-seq dataset⁵². P value, two-tailed Wilcoxon rank-sum test. For all boxplots, hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median, whiskers with maximum 1.5 interquartile range (IQR). $n = 11,231$ (H3K27ac, not at loop anchors), 9,010 (H3K27ac, at loop anchors), 2,553 (H3K27me3, not at loop anchors), 2,185 (H3K27me3, at loop anchors). **e**, Representative snm3C-seq contact heatmap and genome browser view showing class-specific (glutamatergic versus non-neurons), active / repressive putative cCREs regulating gene *Neurod6*. Distal cCREs and proximal regions are highlighted in pink.



Extended Data Figure 1.9: Nuclei gating strategy for mouse frontal cortex.

After nuclei extraction, nuclei were stained with DRAQ7 and proceeded to FANS. First, potential nuclei were identified using forward scatter (FSC) area and backscatter (BSC) area (left dot plot). Next, potential doublets were removed based on BSC as well as FSC signal width (two middle dot plots). Finally, 200 - 500k diploid nuclei (2n) were collected for antibody incubation (right dot plot).

1.8 Methods

Cell culture

Mouse embryonic stem cells (mESC) used in this study have been described in a previous study²⁰. mESC were maintained in feeder-free and serum-free 2i media at 37 °C with 5% CO₂. To isolate nuclei, mESC were dissociated with Accutase (Innovative Cell Technologies, AT104), collected by centrifugation, washed twice with PBS (Gibco, 10010023) then resuspended in cold nuclei permeabilization buffer (10 mM Tris-HCl pH 7.4 (Sigma, T4661), 10 mM NaCl (Sigma, S7653), 3 mM MgCl₂ (Sigma, 63069), 1×Protease Inhibitor (Roche, 05056489001), 0.5 U μl⁻¹ RNaseOUT (Invitrogen, 10777-019), 0.5 U μl⁻¹ SUPERase Inhibitor (Invitrogen, AM2694), 0.1% IGEPAL CA630 (Sigma, I8896) and 0.02% Digitonin (Sigma, Cat#D141) for 1 minute. Nuclei were counted using a BioRad TC20 cell counter with 0.4% Trypan Blue (Gibco, 15250061) staining. For each Droplet Paired-Tag experiment, 0.5 million nuclei were used.

Mouse brain dissection and nuclei extraction

All animal work described in this manuscript has been approved and conducted under the oversight of the UC San Diego Institutional Animal Care and Use Committee. Male C57BL/6J mice were purchased from the Jackson Laboratory (#000664) at 12 weeks of age and housed in the barrier facility at UC San Diego in a 12-hour light/dark cycle in a temperature-controlled room, with *ad libitum* access to water and food, until euthanasia and tissue collection at 16 weeks of age. The temperature in the animal facility is maintained within the range of 20°C (68°F) to 22.2°C (72°F), while the humidity levels can vary between 35% and 60%. The frontal cortex was dissected from 16-week male mice, snap-frozen in liquid nitrogen, and stored at -80°C before proceeding to nuclei extraction.

Single cell suspensions were prepared from frozen tissues by douncing in the douncing buffer (0.25 M sucrose (Sigma, S7903), 25 mM KCl (Sigma, P9333), 5 mM MgCl₂, 10 mM Tris-HCl pH 7.4, 1 mM DTT (Sigma, D9779), 1× Protease Inhibitor, 0.5 U μl⁻¹ RNaseOUT, 0.5 U μl⁻¹ SUPERase Inhibitor). The cell suspension was then filtered through a 30-μm Cell-Tric (Sysmex) for debris removal, and spun down for 10 mins at 300g, 4 °C. Cell pellets was washed once with douncing buffer, spun down again, and resuspended in cold nuclei permeabilization buffer for 10mins. Permeabilized nuclei were pelleted by centrifuge for 10 mins at 1,000g, 4 °C, and washed with sort buffer (1×PBS (Gibco, 10010023), 1×Protease Inhibitor (Roche, 05056489001), 0.5 U μl⁻¹ RNaseOUT (Invitrogen, 10777-019), 0.5 U μl⁻¹ SUPERase Inhibitor (Invitrogen, AM2694), 1mM EDTA (Invitrogen, 15575020), 1% BSA (Sigma, A1595) once. After resuspension in sort buffer, nuclei are stained with 2 μM 7-AAD (Invitrogen, A1310) for 10mins on ice, and proceed to Fluorescence-activated nuclei sorting (FANS) with an SH800 cell sorter (Sony) for isolation of single nucleus (Extended Data Fig. 1.9). Nuclei are collected in collection buffer (1×PBS (Gibco, 10010023), 5×Protease Inhibitor (Roche, 05056489001), 2.5 U μl⁻¹ RNaseOUT (Invitrogen, 10777-019), 2.5 U μl⁻¹ SUPERase Inhibitor (Invitrogen, AM2694), 1mM EDTA (Invitrogen, 15575020), 5% BSA (Sigma, A1595)) at 5 °C, then immediately proceed to centrifuge for 10 mins at 750g, 4 °C. Nuclei were washed twice with sort buffer and counted on an RWD C100-pro fluorescence cell counter with DAPI staining to better estimate the nuclei number. For each histone modification, around 0.5 million sorted nuclei were aliquoted and proceeded to Paired-Tag experiments.

Assembly of active transposon complex

Sequences for all custom oligos used in this study are provided in Supplementary Table 1. To assemble Protein A-Tn5 (pA-Tn5) complex suitable for 10X Single Cell Multiome ATAC +

Gene Expression platform, we mixed transposome oligos (100 μ M) in two separated PCR tubes (USA Scientific, 1402-2300): 25 μ L AdapterA + 25 μ L bridge-pMENTS, or 25 μ L AdapterB + 25 μ L pMENTS. Oligos were annealed in a thermal cycler with the following program: 95 $^{\circ}$ C for 5 mins, then slowly cool down to 12 $^{\circ}$ C at the speed of 0.1 $^{\circ}$ C/s. Each 1 μ L of the annealed transposome DNAs was mixed separately with 6 μ L of unloaded pA-Tn5 (0.5 mg/mL, MacroLab), mixed, and quickly spun down. The mixtures were incubated at room temperature for 30mins then at 4 $^{\circ}$ C for an additional 10 min, and an equal volume of assembled pA-Tn5-AdapterA and pA-Tn5-AdapterB were mixed to form functional pA-Tn5 complex. Assembled transposon complexes can be stored at -20 $^{\circ}$ C for up to six months, and transposon activity is validated with bulk CUT&Tag assay before use.

Antibody

Antibodies used in this study include: H3K27ac (Abcam, ab177178, Lot GR3202987-20 (recombinant); Abcam, ab4729, Lot GR3442886-1 (polyclonal)) and H3K27me3 (Abcam, ab192985, Lot GR3399022-3 (recombinant)). We found that antibody specificity is critical for high-quality signals of single-cell histone data. For H3K27ac, although recombinant antibody yielded a higher fragment number per cell than polyclonal antibodies, its enrichment at transcription starting sites or ChIP-seq peaks were lower (Extended Data Fig. 1.6e). Therefore, except for replicate 1 (rep1) of the mouse frontal cortex datasets, all other experiments targeting H3K27ac were carried out with the polyclonal antibody. Per Droplet Paired-Tag reaction, 1 μ g antibody was used.

Experimental Protocol for Droplet Paired-Tag

A brief description of the Droplet Paired-Tag experimental procedure is provided below. A more detailed, step-by-step protocol is provided in Protocol Exchange²⁵.

Antibody-guided tagmentation

pA-Tn5 and primary antibody were pre-conjugated during nuclei extraction. 1 μg primary antibody and 1 μL assembled pA-Tn5 are premixed in 35 μL MED1 buffer (20 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1 $\times$ Protease Inhibitor cocktail, 0.5 U μL^{-1} SUPERase IN, 0.5 U μL^{-1} RNaseOUT, 0.01% IGEPAL CA630, 0.01% Digitonin, 2 mM EDTA and 1% BSA) and rotated at room temperature for 1 hour, as a previous study showed that high salt condition is critical to reducing undesired open chromatin background⁹. Nuclei extracted with the above-described protocol were also resuspended in MED1 buffer, and 0.15 - 0.50 million nuclei are distributed into the pre-mixed antibody/pA-Tn5 to a final volume 75 μL . The mixtures were rotated at 4 °C overnight for epitope targeting along with pA-Tn5 tethering.

After overnight incubation, nuclei were isolated by centrifugation for 10 mins at 300g, 4 °C and washed with MED2 buffer (20 mM HEPES pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1 $\times$ Protease Inhibitor cocktail, 0.5 U μL^{-1} SUPERase IN, 0.5 U μL^{-1} RNaseOUT, 0.01% IGEPAL CA630, 0.01% Digitonin and 1% BSA) three times to remove excessive antibody and pA-Tn5. Tagmentation is next carried out in 50 μL MED2 buffer supplemented with 10mM MgCl_2 (Sigma, M1028) at 550 r.p.m., 37 °C for 60 mins in a ThermoMixer (Eppendorf). The tagmentation reaction was terminated by adding an equal volume of stop solution (10mM Tris-HCl pH 8.0, 20mM EDTA, 2% BSA, 2 $\times$ Protease Inhibitor cocktail, 1 U μL^{-1} SUPERase IN and 1 U μL^{-1} RNaseOUT). Nuclei were spun down for 10 mins at 500g, 4 °C, and washed once with 1 $\times$ Nuclei Buffer (10X Genomics PN-2000207, supplemented with 1mM DTT, 0.5 U μL^{-1} SUPERase IN and

0.5 U μL^{-1} RNaseOUT). Finally, nuclei were resuspended in 10 μL 1 $\times$ Nuclei Buffer, and 10,000 – 16,000 nuclei were aliquoted into PCR tubes with a total volume of 8 μL . Normally, the nuclei recovery rate is around 30% - 60% depending on the starting input materials. To assess the CUT&Tag performance, around 10,000 - 50,000 nuclei were used for bulk analysis. the fragmented DNA in these nuclei was purified with MiniElute (QIAGEN, 28004) column, and amplified for quality control assessment. 7 μL ATAC Buffer B (10X Genomics PN-2000193) is added to 8 μL nuclei mixture to reach a final reaction volume of 15 μL , same as specified in the user manual of Chromium Next GEM Single Cell Multiome kit (CG000338, RevF), except that we substitute 3 μL ATAC enzyme B (10X Genomics PN-2000265/ 2000272) with 1 $\times$ Nuclei Buffer.

Reverse transcription, cell barcoding, library preparation

Barcoding reaction mixes are prepared as described in the manual of Chromium Next GEM Single Cell Multiome kit. 60 μL prepared master mix is added to 15 μL nuclei mixture before being loaded onto Chromium Next GEM Chip J and proceeding to droplet generation with Chromium X microfluidics system (10X Genomics). Reverse transcription and cell barcoding were carried out inside 10X GEM. Final DNA and RNA libraries amplification was performed according to the Chromium Single Cell ATAC Library kit manual, except that we used an increased number of amplification cycles (in total 12-13) for histone modality library.

Preprocessing of Droplet Paired-Tag data

All sequencing was carried out with an Illumina Nextseq2000 sequencer. Droplet Paired-Tag fastq files are demultiplexed using cellranger-arc (v2.0.0) with command `cellranger-arc mkfastq`; However, DNA and RNA data were preprocessed using cellranger-atac (v2.0.0) and

cellranger (v6.1.2), respectively, and barcodes were manually paired with custom script using the matching relationship provided in cellranger-arc. To select high-quality nuclei, we first aggregated histone modification data from the same sample and performed peaks calling to find narrow peaks (for H3K27ac) or broad peaks (for H3K27me3) using MACS2 (v2.1.2)²⁶. Then, we filtered the histone modality based on the per-cell fragment number and fraction of reads in peaks (FRiP). Next, we selected nuclei by pairing pass-filtered nuclei from both modalities (histone modification and transcriptome), as shown in Extended Data Fig. 1.3a. Before clustering, nuclei with a high fraction of mitochondrial and ribosomal RNA reads were filtered out. Nuclei with an extremely high number of reads were also filtered out since most of them are doublet. Potential doublets were identified and removed using Scrublet²⁷ for individual RNA datasets, corresponding DNA barcodes were also removed.

For genome browser track generation, sample or cell-type specific bigwig files were generated from bam files with deepTools (v3.5.1)²⁸ and visualized in IGV (v2.15.4)²⁹.

For FRiP calculation, duplicated reads were removed using Samtools (v1.14)³⁰ and Picard MarkDuplicates (v2.25.0)³¹, taking barcode information into account. Peaks calling was then performed using macs2 with default parameters, except that for H3K27me3 we called broad peaks with `--broad` due to its broad domain enrichment. Pre-processing pipelines and scripts are shared at <https://github.com/Xieeeee/Droplet-Paired-Tag>.

Analysis of Droplet Paired-Tag data

Signal enrichment calculation

Density plot and heatmap of signal enrichment over ChIP-seq peaks or CEMBA cCREs were carried out using deepTools. Peaks overlapped with ENCODE blacklist (v2) or CUT&RUN blacklist regions were removed during the calculation of enrichment^{32,33}.

Clustering and annotation

Clustering of single-cell transcriptome data was carried out in R using Seurat (v4.1.0)³⁴ and Signac (v1.6.0)³⁵. In short, gene counts were normalized, and the top 2,500 variable genes were selected for dimension reduction by principal component analysis (PCA). For all datasets, the first 35 principal components were used to correct batch effects with harmony³⁶, followed by UMAP visualization and Louvain clustering. Marker genes for each cluster are identified using Seurat, with the Log₂ fold change threshold set to 0.25. Annotation of cluster identities was done with the marker genes characterized in previous studies. For epigenome data, 10X fragment files were converted to cell-by-bin matrices using 5kb non-overlapping genomic bins, and clustering was carried out using Signac. Briefly, sequencing depth was normalized using the two-step term frequency-inverse document frequency (TF-IDF). The top 85% of genomic bins were selected for linear dimension reduction by singular value decomposition (SVD), and batch effects were corrected with harmony, again followed by UMAP visualization and Louvain clustering. Gene activity score was computed by signal density in promoter and gene body regions.

To compare clustering results from different modalities, we first annotated the epigenome clusters by nominating the most abundant cell type identified with transcriptome clustering. Overlap coefficients were calculated according to the proportion of cells sharing the same labels from both the transcriptome (*A*) and epigenome clusters (*B*), in the transcriptome clusters:

$$O_i = \max \left(\frac{A_x \cap B_i}{A_i} \right)$$

Integration with public snRNA-seq datasets

Integration of Droplet Paired-Tag RNA modality with the original Paired-Tag dataset and 10X snRNA-seq dataset is carried out using Seurat. Briefly, gene counts for all datasets were normalized, and the top 2,000 shared variable genes across datasets were identified as integration features. Dimensional reduction (CCA) was carried out to project all nuclei into the same embedding, and “anchors” (pairs of cells from different datasets) were identified by mutual nearest neighbors searching. Low confidence anchors were filtered out, and finally shared neighbor overlap between anchor and query cells in an overall neighbor graph was computed. Louvain clustering on the overall neighbor graph is used for co-embedded cluster identification. To compare clustering results from different datasets, overlap coefficients were calculated based on the number of cells sharing the same labels from both the query (A) and reference clusters (B), in the co-embedding clusters (C) (i = query cell type, j = reference cell type, k = co-embedding cluster):

$$O_{i,j} = \min \left(\left(\frac{A_i \cap C_k}{A_i} \right), \max \left(\frac{B_j \cap C_k}{B_j} \right) \right)$$

Identification of peak set

To identify peaks using H3K27ac data, we adopted a previously described method to perform peak calling and merge peaks across replicates²³. Properly paired reads from all pass-filtered nuclei in the same replicate were merged to generate a pseudo-bulk dataset for all biological replicates. After shift correction for pA-Tn5 cleavage site, peak calling was performed

on the using MACS2 with these parameters: ``-q 0.01 --nomodel --shift -75 --extsize 150 --keep-dup all -B --SPMR``. Since we used two different antibodies, to reduce variation caused by antibody affinity and specificity, we retained peaks identified in at least two replicates as conserved peaks during merging. Finally, we extended peak summits to a fixed width of 500bp for merging and downstream analysis.

Identification of cCRE modules

To ensure a fair comparison between different techniques, we filtered the CEMBA cCREs list for elements with an arbitrary cutoff (histone modification signal RPKM > 1) in at least one transcriptome-based cluster from both Droplet Paired-Tag and Paired-Tag datasets, and retained cCRE with H3K27ac (289,437, 87.9% of cCREs) or H3K27me3 (127,005, 34.9% of cCREs) signal RPKM > 1 in at least one cluster for downstream analysis. For visualization, shared cCREs (108,319) from H3K27ac and H3K27me3 groups are selected for plotting. cCREs were classified as distal or proximal based on their distance to ± 2 kb of the TSS in GENCODE mm10 (vm25). Proximal or distal cCREs were grouped into 20 different modules by non-negative matrix factorization (NMF)³⁷ based on their histone modification signal intensity. Downstream motif enrichment and gene ontology analysis are based on this classification of cCRE modules. For visualization, 95,799 distal cCREs and 12,520 proximal cCREs pass both H3K27ac and H3K27me3 signal cutoffs are plotted. Cell types with <100 nuclei are excluded for further analysis (VLMC, Vascular and leptomeningeal cells).

Linking cCREs with putative target genes

We used a previously described method to link cCREs with their putative regulatory genes for both histone modifications³⁸. First, cCREs with co-occupancy of H3K27ac or H3K27me3

within 500kb genomic distance were identified using Cicero³⁹. cCREs with co-occupancy (Cicero score) >0.1 were retained for further analysis. Next, cCREs were classified as distal or proximal based on their distance to ± 2 kb of the TSS in GENCODE mm10 (vm25). In our analysis, only distal-to-proximal pairs were selected for comparison. We calculated Spearman's correlation coefficients (SCC) between gene expression and histone modification signal over cCRE across clusters to examine the relationship between co-accessibility pairs. To estimate random background levels, we shuffled the cell identities for each read and calculated the corresponding SCCs. Finally, we fit a normal distribution model and set a cutoff SCCs score with an empirically defined significance threshold of $FDR < 0.05$, to select significant positively (H3K27ac) or negatively (H3K27me3) correlated cCRE-gene pairs. To compare with the original Paired-Tag dataset on the strength of putative cCRE-gene pairs, we used Kolmogorov–Smirnov test to calculate the difference between putative cCREs-gene linkages over random backgrounds. For comparison with the previously published Paired-Tag dataset, The greatest distance (D) between real and background distributions is calculated for each dataset, respectively.

Motif enrichment and gene ontology analysis

Motif enrichment for each cCRE module was performed using HOMER (v4.11) 'findMotifsGenome.pl'⁴⁰. The displayed motif heatmap in Fig. 1.2k and Extended Data Fig. 1.8b were from the results of known motif discovery. Gene ontology analysis for each enriched gene set is carried out using PANTHER with default parameters, and biological process terms were used for annotation⁴¹. To exclude ambiguous terms, we only selected the top enriched terms ranked by fold enrichment $\times -\text{Log}_{10}(\text{adjusted } p\text{-value})$.

Integrative analysis of Droplet Paired-Tag and snm3C-seq data

Mouse brain snm3C-seq dataset was downloaded from GEO with accessing number GSE156683²⁴. Contact pairs from individual cells are merged and visualized using pairtools (v1.0.2) at 5kb resolution⁴². To summarize putative cCREs-genes pairs at loop anchors, we first performed loop calling using HiCCUPS (Juicer tools v1.22.01) at resolution of 5kb, 10kb and 25kb⁴³. Merged loop sets are then intersected with cCRE-gene pairs using bedtools pairtobed (v2.27.1)⁴⁴.

Data availability

Raw data obtained in this study have been deposited at the NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) with accession number GSE224560. The processed data can also be accessed as supplementary files in GEO. Datasets for mESC H3K27ac ChIP-seq were downloaded from 4DN data portal with the accession numbers 4DNESTVGLCD9. Other external datasets were downloaded from NCBI Gene Expression Omnibus (GEO) with the following accession numbers: mESC H3K27me3 ChIP-seq ([GSE156589](#)), Paired-Tag ([GSE152020](#)), CoTECH ([GSE158435](#)), scCUT&Tag on PMBC ([GSE157910](#)), scCUT&Tag on brain ([GSE163532](#)), scCUT&Tag-pro ([GSE195725](#)), snm3C-seq on brain ([GSE156683](#)) and ChIP-seq on mouse cortex excitatory neurons ([GSE141587](#)). The CEMBA snATAC-seq datasets and BICCN 10XsnRNA-seq MOp data are downloaded via the NeMO archive (RRID: SCR_016152) (<https://assets.nemoarchive.org/dat-ch1nqb7>). 10x PBMC scRNA-seq and E18 embryonic mouse brain Multiome datasets were downloaded from the 10x Genomics website (<https://www.10xgenomics.com/resources/datasets>). Source data for statistical analysis are provided along in this paper.

Code availability

Scripts and code are available at <https://github.com/Xieeeee/Droplet-Paired-Tag>.

1.9 References

1. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
2. Smallwood, S. A., Lee, H. J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S. R., Stegle, O., Reik, W., & Kelsey, G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
3. Chi, Y., Shi, J., Xing, D. & Tan, L. Every gene everywhere all at once: High-precision measurement of 3D chromosome architecture with single-cell Hi-C. *Front. Mol. Biosci.* **9**, 959688 (2022).
4. Preissl, S., Gaulton, K. J. & Ren, B. Characterizing cis-regulatory elements using single-cell epigenomics. *Nat. Rev. Genet.* **24**, 21–43 (2022).
5. Rotem, A., Ram, O., Shores, N., Sperling, R. A., Goren, A., Weitz, D. A., & Bernstein, B. E. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
6. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. *Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues.* *Nature Biotechnology* **39**, 825–835 (2021).
7. Wu, S. J., Furlan, S. N., Mihalas, A. B., Kaya-Okur, H. S., Feroze, A. H., Emerson, S. N., Zheng, Y., Carson, K., Cimino, P. J., Keene, C. D., Sarthy, J. F., Gottardo, R., Ahmad, K., Henikoff, S., & Patel, A. P. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nat. Biotechnol.* **39**, 819–824 (2021).
8. Patty, B. J. & Hainer, S. J. Transcription factor chromatin profiling genome-wide using uliCUT&RUN in single cells and individual blastocysts. *Nat. Protoc.* **16**, 2633–2666 (2021).
9. 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).
10. Zhu, C., Yu, M., Huang, H., Juric, I., Abnoui, A., Hu, R., Lucero, J., Behrens, M.M., Hu, M. & Ren, B. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat. Struct. Mol. Biol.* **26**, 1063–1070 (2019).
11. 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).

12. Zhu, C., Zhang, Y., Li, Y.E., Lucero, J., Behrens, M.M. & Ren, B. Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nat. Methods* **18**, 283–292 (2021).
13. Xiong, H., Luo, Y., Wang, Q., Yu, X. & He, A. Single-cell joint detection of chromatin occupancy and transcriptome enables higher-dimensional epigenomic reconstructions. *Nat. Methods* **18**, 652–660 (2021).
14. Mimitou, E. P., Lareau, C. A., Chen, K. Y., Zorzetto-Fernandes, A. L., Hao, Y., Takeshima, Y., Luo, W., Huang, T. S., Yeung, B. Z., Papalexi, E., Thakore, P. I., Kibayashi, T., Wing, J. B., Hata, M., Satija, R., Nator, K. L., Sakaguchi, S., Ludwig, L. S., Sankaran, V. G., Regev, A., & Smibert, P.. *Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat. Biotechnol.* **39**, 1246-1258 (2021).
15. Chen, A.F., Parks, B., Kathiria, A.S., Ober-Reynolds, B., Goronzy, J.J. & Greenleaf, W.J. NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat. Methods* **19**, 547–553 (2022).
16. Zhang, B., Srivastava, A., Mimitou, E., Stuart, T., Raimondi, I., Hao, Y., Smibert, P. & Satija, R.. Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat. Biotechnol.* **40**, 1220–1230 (2022).
17. Stuart, T., Hao, S., Zhang, B., Mekerishvili, L., Landau, D.A., Maniatis, S., Satija, R. & Raimondi, I. Nanobody-tethered transposition enables multifactorial chromatin profiling at single-cell resolution. *Nat. Biotechnol.* **41**, 806–812 (2023).
18. Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K. & Henikoff, S. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).
19. Aranda, S., Mas, G. & Di Croce, L. Regulation of gene transcription by Polycomb proteins. *Sci. Adv.* 1(11):e1500737 (2015).
20. Huang, H., Zhu, Q., Jussila, A., Han, Y., Bintu, B., Kern, C., Conte, M., Zhang, Y., Bianco, S., Chiariello, A. M., Yu, M., Hu, R., Tastemel, M., Juric, I., Hu, M., Nicodemi, M., Zhuang, X., & Ren, B. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nat. Genet.* **53**, 1064–1074 (2021).
21. Kinoshita, M., Barber, M., Mansfield, W., Cui, Y., Spindlow, D., Stirparo, G. G., Dietmann, S., Nichols, J., & Smith, A. Capture of Mouse and Human Stem Cells with Features of Formative Pluripotency. *Cell Stem Cell* **28**, 453-471.e8 (2021).
22. 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., Boeshaghi, 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).
23. 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).
 24. 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).
 25. Xie, Y., Zhu, C., Wang, Z., Tastemel, M., Chang, L., Li, Y. E., & Ren, B. Droplet-based Single-cell Joint Profiling of Histone Modification and Transcriptome. Preprint at *Protocol Exchange* (2023).
 26. Zhang, Y., Liu, T., Meyer, C. A., Eeckhoutte, 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**, (2008).
 27. 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).
 28. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, 187–191 (2014).
 29. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
 30. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
 31. Broad Institute. Picard Toolkit. <http://broadinstitute.github.io/picard/> (2018).
 32. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

33. Nordin, A., Zambanini, G., Pagella, P. & Cantù, C. The CUT&RUN Blacklist of Problematic Regions of the Genome. Preprint at *bioRxiv* (<https://doi.org/10.1101/2022.11.11.516118>) (2022).
34. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P. & Satija, R. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
35. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
36. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P. R., & Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
37. Li, Y. E., Xiao, M., Shi, B., Yang, Y. T., Wang, D., Wang, F., Marcia, M., & Lu, Z. J. Identification of high-confidence RNA regulatory elements by combinatorial classification of RNA-protein binding sites. *Genome Biol.* **18**, 169 (2017).
38. Xie, F., Armand, E. J., Yao, Z., Liu, H., Bartlett, A., Behrens, M. M., Li, Y. E., Lucero, J. D., Luo, C., Nery, J. R., Pinto-Duarte, A., Poirion, O. B., 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, *Cell Genomics* **3**, 100342 (2023).
39. 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).
40. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., & Glass, C. K. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).
41. Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L. P., & Mi, H. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022).
42. Open2C, Abdennur, N., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., Imakaev, M., & Venev, S. V. Pairtools: from sequencing data to chromosome contacts. Preprint at *bioRxiv* (<https://doi.org/10.1101/2023.02.13.528389>) (2023).
43. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).

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

Chapter 2 DROPLET HI-C FOR FAST AND SCALABLE PROFILING OF CHROMATIN ARCHITECTURE IN SINGLE CELLS

2.1 Abstract

Comprehensive analysis of chromatin architecture is crucial for understanding the gene regulatory programs during development and in disease pathogenesis, yet current methods often inadequately address the unique challenges presented by analysis of heterogeneous tissue samples. Here, we introduce Droplet Hi-C, which employs a commercial microfluidic device for high-throughput, single-cell chromatin conformation profiling in droplets. Using Droplet Hi-C, we mapped the chromatin architecture at single-cell resolution from the mouse cortex and analyzed gene regulatory programs in major cortical cell types. Additionally, we used this technique to detect copy number variation (CNV), structural variations (SVs) and extrachromosomal DNA (ecDNA) in cancer cells, revealing clonal dynamics and other oncogenic events during treatment. We further refined this technique to allow for joint profiling of chromatin architecture and transcriptome in single cells, facilitating a more comprehensive exploration of the links between chromatin architecture and gene expression in both normal tissues and tumors. Thus, Droplet Hi-C not only addresses critical gaps in chromatin analysis of heterogeneous tissues but also emerges as a versatile tool enhancing our understanding of gene regulation in health and disease.

2.2 Introduction

In eukaryotic cells, chromosomes are organized within the nucleus in a non-random manner¹⁻⁴. Each chromosome is folded into a complex and dynamic structure, forming chromatin

loops⁵, topologically associating domains^{6, 7}, and A/B compartments⁸⁻¹⁰. Chromatin organization plays a crucial role in both embryonic development^{11, 12} and the progression of various diseases^{13, 14}. During development, the chromatin structure influences which genes are turned on or off by the distal regulatory elements¹⁵⁻¹⁸. Abnormal chromatin organization can lead to aberrant gene expression or silencing, contributing to cancer development and progression¹⁴. Indeed, components of the nuclear machinery that regulates chromatin organization are among the most frequently mutated genes in human cancers¹⁹⁻²². Thus, comprehensive analysis of the chromatin organization is crucial for understanding the gene regulatory programs involved in both normal development and disease pathogenesis.

Analyzing chromatin organization in primary tissues and tumor biopsies presents unique challenges due to the complexity of chromatin dynamics and the technical limitations of current methodologies. First, the biospecimens available for assays are often heterogeneous, containing a mixture of different cell types. Existing techniques such as *in situ* Hi-C for analysis of 3D chromatin organization typically require bulk tissues as input and could not resolve specific chromatin organization patterns relevant to particular cell types, especially in tumors where cancer cells coexist with stromal and immune cells. Second, the resulting data obtained from bulk chromatin organization assays are complex, and relating changes in chromatin organization to specific functional outcomes can be difficult. Addressing these challenges requires the development of more sensitive single-cell chromatin analysis, better methods for handling tissue samples, and advanced computational tools for data analysis.

Great strides have been made in recent years in single-cell Hi-C technologies¹⁻⁴. In general, single-cell Hi-C methods can be categorized into low-throughput microwell-based approaches and high-throughput combinatorial indexing-based approaches. In the case of microwell-based single-

cell Hi-C methods, cells/nuclei are individually dispensed into microwells, and library construction is carried out in each microwell in parallel, frequently with the help of automated liquid handlers. Notably, several single-cell Hi-C methods have been developed, such as those by Nagano *et al.* (2013)²³, Nagano *et al.* (2017)²⁴, Stevens *et al.* (2017)²⁵, Flyamer *et al.* (2017)²⁶, Tan *et al.* (2018)²⁷ and third-generation sequencing (TGS) based scNanoHi-C²⁸. Recently, additional microwell-based single-cell Hi-C methods have emerged that combine Hi-C analysis with assays for one or more additional molecular modalities in a single cell. For example, Methyl-HiC²⁹ and single-nucleus methyl-3C sequencing (sn-m3C-seq)³⁰ capture chromatin interactions and DNA methylation patterns from the same cell. HiRES³¹ jointly performs Hi-C and RNA sequencing to explore the functional relationship between 3D genome organization and transcriptome dynamics. In general, microwell-based techniques are limited in scale, and difficult to adopt broadly due to lengthy procedures and high cost. On the other hand, combinatorial indexing-based single-cell Hi-C methods utilize combinatorial barcoding strategies to achieve high throughput and scalability. Previously, a single-cell combinatorial indexed Hi-C (sci-Hi-C)³² method allowed the generation of single-cell Hi-C libraries from a few thousand cells, albeit with limited genomic coverage in each cell. Building upon the same strategy, GAGE-seq (genome architecture and gene expression by sequencing)³³ was developed to simultaneously profile chromatin interactions and gene expression in single cells, to achieve high throughput, multimodality, and high coverage per cell. However, the lengthy and largely manual combinatorial indexing procedure still poses a challenge for its general adoption.

Here, we introduce a highly scalable, generally accessible, droplet-based single-cell Hi-C method, Droplet Hi-C, which combines an *in situ* chromosomal conformation capture (3C) assay with commercially available droplet microfluidics, to simultaneously capture the 3D genome

structure from tens of thousands of individual cells in a single experiment. We demonstrate the utility of Droplet Hi-C data in resolving cell type-specific chromatin architecture in complex tissues such as the mouse brain. We further show that Droplet Hi-C can be used to identify aberrant chromatin structure in cancer cells. Most excitingly, we used Droplet Hi-C to identify extra-chromosomal DNA (ecDNA) and map their chromatin interactome in tumor cells at single-cell resolution. Finally, we extended this method to enable the simultaneous capture of transcriptome and chromatin architecture in single cells.

2.3 Results

Development of Droplet Hi-C

Droplet Hi-C builds upon the *in situ* Hi-C procedure⁵. It captures spatial proximity genome-wide between chromatin fibers in formaldehyde-crosslinked cells/nuclei through restriction digestion and ligation *in situ* (Fig. 2.1a). After SDS treatment to remove histone proteins, DNA fragmentation and capture is then carried out in a commercially available microfluidic platform (i.e. 10x Genomics single cell ATAC kit), where cell-specific DNA barcodes are added to the DNA fragments. This is followed by sequencing library construction and next-generation sequencing (NGS) (Fig. 2.1a). The whole procedure lasts about 10 hours from fixed cells/nuclei to final sequencing libraries, and 8 samples can be processed in parallel, enabling profiling of 40,000 or more cells simultaneously (Fig. 2.1a-c). The throughput of Droplet Hi-C surpasses plate-based single-cell Hi-C methods by an order of magnitude, offering the shortest experimental duration and hands-on time (Fig. 2.1b). Given the widespread use of commercial microfluidic

systems, ease of the experimental procedure, and relatively low cost (Fig. 2.1c), Droplet Hi-C has the potential to be rapidly adopted by the research community.

To test the performance of Droplet Hi-C, we first applied it to a mixture of human HeLa S3 cell line and mouse embryonic stem cell (mESC) line. Equal number of cells from two cell lines were combined after cross-linking and subjected to Droplet Hi-C. The results showed that 1,773 human and 3,489 mouse high-quality cells (with total read pairs number > 1,000) were recovered after shallow sequencing, along with 284 potential doublets (Extended Data Fig. 2.1a). We further carried out Droplet Hi-C with a mixture of three human cell lines (K562, GM12878 and HeLa S3). After filtering and clustering using Higashi³⁴, we obtained 3,709 high-quality cells with a median of 108,439 unique read pairs in each cell (duplication rate 19.0%), including 1,604 HeLa S3 cells (median 101,460 read pairs), 1,359 GM12878 cells (median 109,518 read pairs), 606 K562 cells (median 111,255 read pairs) and 140 cells with chromatin interaction pattern similar to those in cell cycle (median 119,573 read pairs, Extended Data Fig. 2.1b). By performing genome-wide correlation analysis with bulk *in situ* Hi-C references, each except the cell cycle population can be confidentially assigned to one cell type (Extended Data Fig. 2.1c-f). The identity of cell cycle population is assigned based on the similar contact frequency decay by distance pattern compared to HeLa S3 cells arrested at prometaphase (PM)³⁵, and the signature contact maps enriched with frequent interactions along the diagonal (Extended Data Fig. 2.1d-e). On the chromosome level, cell type-specific chromatin structures can be observed, such as the distinct Hi-C pattern of chromosomal rearrangements on chromosome 11 in HeLa S3 cells (Extended Data Fig. 2.1e). When focusing on a region encompassing the K562-specific upregulated gene *LGR4*³⁶, we observed an active A compartment at the *LGR4* genomic locus that is specifically shown in the K562 cells (Extended Data Fig. 2.1g). The differences in compartment scores as well as insulation

scores among different cell lines can also be faithfully reproduced as in the published datasets (Extended Data Fig. 2.1g-h). We also performed Droplet Hi-C on a normal-like cell mixture (GM12878 and WTC-11) to prove that single cell analysis of chromatin conformation enables to separate different cell lines without genomic rearrangement information. Clustering of the 3,669 high quality cells revealed three populations (Extended Data Fig. 2.1i), including 2,236 GM12878 and 1,202 WTC-11 cells which showed high correlation with *in situ* Hi-C references (Extended Data Fig. 2.1j), along with 231 cells identified as in cell cycle (Extended Data Fig. 2.1k-l). In summary, these results demonstrate that Droplet Hi-C can accurately distinguish cell type-specific chromatin organization.

Droplet Hi-C reveals cell type-specific chromatin structure in mouse cortex

To demonstrate the feasibility of applying Droplet Hi-C to primary heterogeneous tissues, we performed this assay using the cortex dissected from 8-week-old mice. In total, we generated 6,235 high-quality single-cell profiles of chromatin architecture from two replicates, with a median of 175,021 unique read pairs per cell (with 22,456 *cis*-long contacts (> 1 kb) and 9,954 trans contacts per cell, at 58% duplication rate) (Extended Data Fig. 2.2a-d). On the single-cell level, despite the sparse contact signals, we can still discern an enrichment of intra-chromosomal interactions (Extended Data Fig. 2.2e). To benchmark the performance of Droplet Hi-C, we compared our data with previous single-cell Hi-C methods, Dip-C³⁷ and sn-m3C-seq³⁸, data from the mouse cortex. The decay curve of contact probability by genomic distance showed a similar trend among these assays, with sn-m3C-seq profiles showing a slightly lower fraction of extreme long-range interactions than Droplet Hi-C and Dip-C (Extended Data Fig. 2.2f). Droplet Hi-C

yields a higher fraction of *cis*-long (> 1 kb) interactions than Dip-C but lower than sn-m3C-seq (Extended Data Fig. 2.2g-h). A similar *cis* long-*trans* ratio in chromatin contacts was detected between Droplet Hi-C and snm3C-seq (Extended Data Fig. 2.2g).

To verify the fidelity of Droplet Hi-C profiles, we examined chromatin structures on contact heatmaps spanning multiple resolutions, and found them to be nearly identical between Droplet Hi-C and sn-m3C-seq (Extended Data Fig. 2.2i). At the chromatin compartment level, Interaction frequency within the active A or inactive B compartments are similar with sn-m3C-seq (Extended Data Fig. 2.2j). Although Droplet Hi-C used Tn5 transposase for chromatin fragmentation in nucleus, it shows minimal open chromatin bias (Extended Data Fig. 2.2k). The compartment scores and insulation scores from Droplet Hi-C data are highly correlated with other two single-cell Hi-C methods (Extended Data Fig. 2.2l-o). Overall, these results indicated that Droplet Hi-C can reliably and robustly detect chromatin features in complex tissue.

Chromatin interactions within the gene body have been shown to be positively associated with gene expression levels, and therefore were used to represent features of single-cell Hi-C data at the gene level³⁹. To annotate cell types, we first summarized the chromatin contacts on the gene body region of each gene as gene associating domain (GAD) scores⁴⁰, and then generated a cell-by-gene GAD score feature matrix. We then leveraged single-nuclei RNA-seq data from previous Droplet Paired-Tag⁴¹ on mouse cortex to co-embed and annotate our Droplet Hi-C data (Extended Data Fig. 2.3a). With this strategy, we successfully resolved 20 subclasses, including 5 non-neuronal types, 9 glutamatergic neuronal types and 6 GABAergic neuronal types (Fig. 2.1d, Extended Data Fig. 2.3b). The annotated subclasses show overall high prediction score (Extended Data Fig. 2.3c-d). When compared between different or same neuronal types, the annotated subclasses always show substantial variations in both GAD scores and chromatin interaction

pattern (Extended Data Fig. 2.3e-h), indicating the accuracy of our annotations on Droplet Hi-C data. The annotations accuracy were further validated based on the consistent changes in aggregated GAD scores on known cell type markers with gene expression level from reference snRNA-seq (Extended Data Fig. 2.3i). Through integration with the published sn-m3C-seq data, we also showed that our subclass annotation is highly correlated with the matched annotations from sn-m3C-seq (Extended Data Fig. 2.3j-k). Among these subclasses, neuronal cells exhibit greater similarity to each other than to non-neuronal cells (Fig. 2.1e). The hierarchical chromatin organization features, such as compartments, domains, and chromatin loops, are distinctly discernible in the aggregated single-cell profiles across various cell types (Fig. 2.1f-h). For example, the glutamatergic neurons marker gene *Satb2* displayed long-range interactions with the region nearby *Gm32949* gene that are specific in glutamatergic neurons (ITL6GL, Cortex L6 excitatory neurons; PTGL, Cortex NP excitatory neurons), but not in GABAergic (VIPGA, MGE-derived neurogliaform cells *Vip*⁺; PVGA, MGE-derived neurogliaform cells *Pvalb*⁺) or non-neuronal cells (OGC, oligodendrocytes; OPC, oligodendrocytes precursor cells) (Fig. 2.1h). Accompanying with this glutamatergic neurons-specific contact, the *Satb2* gene also displayed an elevated expression level as well as association with an active histone modification, H3K27ac, in ITL6GL and PTGL cells when analyzed in conjunction with our Droplet Paired-Tag data (Fig. 2.1h). By contrast, in other cell types, this gene remained inactive and was even silenced with the presence of repressive H3K27me3 signals.

To systematically probe the differences in 3D genome structure across different cell types and elucidate their relationships with chromatin state and gene expression, we first examined differences in A/B compartments at 100-kb resolution among cell types. These compartments were found to correspond with transcriptionally active and inactive chromatin states. In total, we

identified 895 genomic regions displaying variable compartment scores across the brain cell types (Fig. 2.2a). When combined with single-cell histone modification profiles previously reported for the mouse cortical region, the compartment scores displayed a positive correlation with H3K27ac signals and a negative correlation with H3K27me3 signals across different cell types, suggesting that variable compartments are associated with dynamic chromatin states (Fig. 2.2b).

Chromatin domains orchestrate enhancer-promoter interactions and regulate cell type-specific gene activities^{18, 37}. We defined domain boundaries in each of the 20 cortical cell types, and subsequently summarized the boundaries across all the cells to obtain the boundary probability for each cell type. Notably, genes with highest expression variation among cell types were often found to reside within or near cell type-specific domain boundaries³⁴. For example, for the *Pdgfra* gene, which plays a crucial role in controlling oligodendrocyte differentiation⁴², we found an OPC specific-chromatin domain boundary overlapping the gene body of *Pdgfra* (Fig. 2.2c). We identified and fetched genes within a 100-kb window centering on variable domain boundaries, and then checked corresponding single-nuclei RNA-seq datasets to classify such genes as constant genes, housekeeping genes and variable genes based on their expression levels across different cell types (Fig. 2.2d). The results showed that the genes with variable expression levels had higher positive correlation with boundary probability than the constant genes and housekeeping genes (Fig. 2.2e).

Finally, we examined chromatin loops within each cell type at 10-kb resolution. To check whether loops are associated with functional chromatin states, we compared both H3K27ac and H3K27me3 histone modification signals at loop anchors across all the cell types. Notably, the active chromatin mark H3K27ac showed a higher enrichment at loop anchors identified in the corresponding cell types than other cell types, while repressive chromatin mark H3K27me3

showed the opposite trend, and was more likely to be depleted at the corresponding cell type-specific loop anchors (Fig. 2.2f). To systematically assess the functions of genes located near cell type-specific loop anchors, we performed GREAT analysis⁴³ for VIPGA- and OGC- specific loop anchors. The VIPGA-specific loop anchors are near genes related to neurotransmitters and ion transport, consistent with the signal transduction functions in neurons, while OGC-specific loop anchors are near genes related to myelination and axon ensheathment (Fig. 2.2g). Taken together, the above analyses showed that Droplet Hi-C can accurately reveal cell type-specific chromatin organization in heterogeneous tissues.

Multi-way interactions is known to happen between enhancers and promoters in regulating lineage-specific genes. To identify chromatin hubs which are enriched multi-way interactions from our single cell data, we first identify 10-kb bins involving multi-way interactions in each cell from our adult mouse cortex Droplet Hi-C data, and then calculate the distribution of interaction frequency among all the cells in each subclass annotation. A 10-kb bin is defined to be in chromatin hub if it has significantly higher interaction frequency in the distribution. After identifying chromatin hubs, we calculate the enrichment of chromatin hubs at cell type specific super enhancers or marker genes, to evaluate their roles in mediating gene regulation (Fig. 2.2h). Using the published datasets to identify super-enhancers or marker genes, we showed that subclass-specific chromatin hubs tend to enriched in super-enhancers and marker genes from the same subclass compared to other subclasses (Fig. 2.2i-j), suggesting that the single cell contacts information is indicative of the cell type specific gene regulation patterns.

Droplet Hi-C detects structural alterations and ecDNA in cancer cells

Cancer cells often exhibit large-scale genomic rearrangements, including copy number variation (CNV) as well as structural variations (SVs), which have been implicated in the initiation and progression of cancer⁴⁴. In previous studies, Hi-C has been shown to detect these genomic variations in bulk tumor tissues with high accuracy and efficiency^{45, 46}. However, these bulk analyses could not resolve the heterogeneity and evolution of tumor populations in response to therapies. Droplet Hi-C overcomes this problem through detection of genomic variations in individual cancer cells. As a proof of principle, we performed Droplet Hi-C with two colorectal cancer cell lines, COLO320DM and COLO320HSR, derived from the same patient, both carrying similar copies of *MYC* amplicon. In COLO320DM, the *MYC* oncogene is located on ecDNA, while In COLO320HSR, it exists on chromosomes as tandem repeats, termed homogeneously staining regions (HSR). By inferring the copy number of each 1-Mb bin in single cells as well as aggregated pseudo-bulk profiles, we detected elevated copies of the *MYC* locus from both samples, as anticipated (Fig. 2.3a). We also observed fluctuations in the copy numbers across the entire genome (Fig. 2.3a). We used the inferred DNA copy number of each genomic region to correct bias in the contact matrices and then identified SVs. Our analysis revealed distinct SVs between these two cancer lines, and we showed one example of duplication event on chromosome 6 (Fig. 2.3b).

It has been difficult to distinguish whether the *MYC* exists on the ecDNA or as HSR from whole genome DNA sequencing alone. We ran AmpliconArchitect algorithm^{47, 48} using whole-genome sequencing (WGS) data of COLO320HSR, and AmpliconArchitect classified the HSR as a circular amplicon. While the genome-wide interaction patterns have been proposed as potential indicators of ecDNA⁴⁹, aggregate Hi-C profiles showed similar patterns of inter-chromosomal interactions between the amplified *MYC* locus and the rest of the genome in both COLO320DM

and COLO320HSR (Fig. 2.3c). Although measurement like adjusted normalized inter-chromosomal interaction frequency (adj_nTIF) has been proposed to determine ecDNA identity⁵⁰, we found that adj_nTIF was equally high at the 1-Mb bin containing *MYC* in both cell lines, indicating that this value is likely not driven by the distinct interaction properties of ecDNA compared to HSR (Fig. 2.3c).

We therefore tried to discern distinctions between *MYC* ecDNA and HSR by analyzing their interactome obtained from the Droplet Hi-C dataset. In COLO320DM, relatively uniform interactions happened within ecDNA (Extended Data Fig. 2.4a). In COLO320HSR, we found two deletions, evidenced by significantly lower numbers of sequencing reads at those regions. Single-cell profiles confirmed that these two regions appeared uniformly across all the cells (Extended Data Fig. 2.4a). Interestingly, we identified interaction hotspots for ecDNA and HSR across the whole genome (Methods), and found that ecDNA interacts with other chromosomes more evenly, while HSR displayed a preference towards certain chromosomes (Fig. 2.3d). To quantitatively measure the evenness of inter-chromosomal contacts, we devised the hub index as the Gini coefficient of inter-chromosomal contacts for the 1 Mb *MYC* bin in each single cell. COLO320DM shows a significantly smaller hub index, indicating a more uniform *trans*-interacting profile (Fig. 2.3e). We also quantified other parameters which can be obtained from single-cell Hi-C data, including inferred copy number (Fig. 2.3f) and the *trans*-to-*cis* contacting bin ratio (Fig. 2.3g). Both parameters exhibited significant differences between COLO320DM and COLO320HSR (Fig. 2.3f-g). COLO320DM cells had approximately 47% more copies than COLO320HSR cells at the *MYC* genomic bin, which aligns with previous research findings⁵¹ (Fig. 2.3f). Additionally, COLO320DM displayed a higher *trans*-to-*cis* contacting ratio for the *MYC* bin compared to COLO320HSR (Fig. 2.3g). These results collectively suggest that *MYC* ecDNA exhibits distinct

interaction patterns from *MYC* HSR at the single-cell resolution, making it a potential indicator for determining whether a genomic bin is likely to be ecDNA or not within an individual cell.

The measurement of *trans*-interaction evenness also paves the way to investigate whether multiple ecDNA tend to cluster into “hubs” in a single cell, since the agglomerated ecDNAs will show contacting preference against specific chromosomes compared to diffused ecDNAs. By shuffling the inter-chromosomal contacts across genome to generate a background distribution as reference for the 1Mb *MYC* bin in COLO320DM and COLO320HSR, we found that the hub index of inter-chromosomal contacts in COLO320HSR is significantly higher, demonstrating that this genomic bin showed stronger bias in contacts for HSR, consistent with the tandem repeat aggregation property of HSR (Extended Data Fig. 2.4b). Although COLO320DM exhibits a lower hub index than COLO320HSR, it is significantly higher compared to the random background, indicating that ecDNA has a strong genomic localization bias in COLO320DM (Extended Data Fig. 2.4b), which is consistent with previously reported ecDNA hubs in COLO320DM⁵².

The observed differences in contact profiles between ecDNA and HSR hint at potential ways to distinguish these aberrant chromosomal features from single-cell Hi-C data. To this end, we fit a multivariate logistic regression model using hub index, estimated copy number, *trans*-to-*cis* contacting bin ratio for ecDNA identification. This regression model is then used to predict ecDNA probability for each 1Mb bin that contains ecDNA among the entire cell population, as well as whether each 1Mb bin contains ecDNA in each single cell (Extended Data Fig. 2.5a). Specifically, the *MYC* locus (chr8:127Mb-128Mb) in COLO320DM cells is treated as ground truth and in COLO320HSR cells as negative in the training dataset. On the validation dataset, the model reached a 0.99 specificity, a 0.97 precision, a 0.89 accuracy and a 0.71 recall. We first applied the regression model to 1,426 COLO320DM cells and 1,535 COLO320HSR cells, which identified

two consecutive bins (chr8:126Mb-128Mb) as ecDNA. In COLO320DM, 72.7% of cells were predicted to be ecDNA positive (Extended Data Fig. 2.5b). In contrast, only 20.5% of cells were predicted to be ecDNA positive in COLO320HSR cells (Extended Data Fig. 2.5b). We also analyzed the 3,423 cells obtained from the adult mouse cortex and found no ecDNA. Taken together, these results show that our logistic regression model based ecDNA caller algorithm can identify known ecDNA from COLO320DM, although the false positive rate remains a problem in the COLO320HSR samples.

To further improve the accuracy of our ecDNA caller, we developed a deep learning model based on convolutional neural networks (CNN) (Fig. 2.3h). The deep learning model utilizes the sliced contact matrix, comprising five consecutive 1Mb genomic bins as its input. At the heart of this process, a neural network is trained using these input variables to identify ecDNA within each central bin, providing predictions for the presence of ecDNA or HSR in 1Mb bins, both at the cell population and single-cell levels (Extended Data Fig. 2.5c, Methods). Our deep learning model has demonstrated remarkable performance, with higher processing speed, achieving a recall rate of 0.80, specificity of 0.99, accuracy of 0.93, while maintaining a 0.99 precision in ecDNA prediction (Extended Data Fig. 2.5d). By applying the deep learning-based ecDNA caller to the COLO320DM and COLO320HSR validation dataset, we were able to identify *MYC* ecDNA in 96% of COLO320DM cells and HSR in 3.86% of COLO320DM cells (Fig. 2.3i top). Conversely, *MYC* HSR was identified in 0.20% of COLO320HSR cells, along with HSR in 99.54% of COLO320HSR cells (Fig. 2.3i bottom). Given that the deep learning-based ecDNA caller pipeline can directly distinguish between ecDNA and HSR, while also outperforming the regression model-based algorithm, we decided to employ the deep learning model-based ecDNA caller algorithm in all our subsequent studies.

Droplet Hi-C uncovers ecDNA heterogeneity and cancer cell evolution in response to drug treatment

Previous studies showed that glioblastoma (GBM) cells carrying extrachromosomal oncogenic variant of epidermal growth factor receptor (EGFRvIII) can develop drug resistance after treatment with tyrosine kinase inhibitors such as erlotinib (ER), which coincides with the loss of *EGFR* ecDNA⁵³. To further characterize the dynamics of *EGFR* ecDNA abundance and heterogeneity during the acquisition of drug resistance, we employed Droplet Hi-C to profile both naive GBM39 cells and cells after more than 30-days of treatment of erlotinib (GBM39-ER) (Fig. 2.4a). With 9,204 high-quality nuclei profiled, we were able to resolve five different clusters of cells based on chromatin architecture in each cell (Fig. 2.4b, Extended Data Fig. 2.6a). After erlotinib treatment, a notable evolutionary shift is observed in the GBM39 cells. Subpopulation cluster 2 (C2) nearly vanished, subpopulation cluster 0 (C0) expanded, and several new subpopulations including clusters 1, 3, and 4 (C1, C3, C4) emerged (Fig. 2.4c-d).

At the sample level, elevated inter-chromosomal contacts are observed for *EGFR* locus and, surprisingly, *MYC* locus before treatment. After erlotinib treatment, the extensive inter-chromosomal contacts at *EGFR* locus diminished, but increased at the *MYC* locus and a new locus on chromosome 12 encompassing *MDM2* (Extended Data Fig. 2.6b). We utilized our ecDNA caller to identify candidate ecDNAs in the Droplet Hi-C data from both GBM39 and GBM39-ER cells. In GBM39, we could detect ecDNA containing the *EGFR* locus (ecEGFR) and ecDNA containing the *MYC* locus (ecMYC) (Extended Data Fig. 2.6c left). We also detected a low-frequency ecDNA on chromosome 18 (1.5%, ecChr18) (Extended Data Fig. 2.6d). In GBM39-ER

cells, we discovered ecDNA harboring the *MDM2* locus (ecMDM2) and also ecMYC (Extended Data Fig. 2.6c right). Surprisingly, we also identified *MYC* HSR in GBM39-ER cells by ecDNA caller (Extended Data Fig. 2.6e) and validated it using fluorescence *in situ* hybridization (FISH) (Extended Data Fig. 2.6f).

At the cluster level, we calculated the percentage of cells containing these three ecDNA candidates in every cluster, revealing the dynamics of ecDNA-harboring cancer cell populations during development of drug resistance (Fig. 2.4e). Specifically, ecEGFR disappeared nearly completely after erlotinib treatment (C1, C3, C4) as previously reported⁵³, whereas ecMDM2 appeared only in erlotinib-resistant populations and is dominant in GBM39-ER subgroup (C3). And ecMYC showed low frequency in GBM39 population (C2) and increased in transition state clusters (C0, C1, C3). We also plotted the contact maps at 25-kb resolution for these three ecDNAs in every cluster. We observed that the structure and size of the ecMYC amplicon were altered, unlike ecEGFR and ecMDM2, suggesting a more intricate transformation at the *MYC* ecDNA during the drug treatment (Fig. 2.4e).

Surprisingly, when we analyzed the distribution of each ecDNA among all the identified clusters, we found coexistence of distinct ecDNA species in some cancer cells (Fig. 2.4f top, Extended Data Fig. 2.6g). In cluster C0, a small population of cells (4.7%) harbored both ecMYC and ecEGFR. We also validated such co-existence of ecMYC and ecEGFR by FISH imaging (Fig. 2.4f bottom). These cells may potentially serve as the drug-resistant pool of cells during erlotinib treatment, given that the gain of ecMYC may increase cellular fitness. This observation raises the intriguing question of whether GBM39-ER was selected from a small population of cells by erlotinib treatment. Since the simultaneous increase in copy number can indicate whether two adjacent genomic regions are co-present in the same ecDNA, we focused on the ecMYC and

characterized its internal structure in GBM39 cells before and after erlotinib treatment at 10-kb resolution. Although the copy number varies across different cells, the boundaries of ecMYC in GBM39 are almost identical (Fig. 2.4g left). Strikingly after treatment, the ecMYC boundaries as well as internal structure displayed huge variability, reflecting the heterogeneity of *MYC* ecDNA at single-cell resolution (Fig. 2.4g right). When performing FISH with dual color *MYC* break-apart probe set in GBM39-ER cells, while some cells showed co-amplification of both 5'*MYC* and 3'*MYC*, many cells displayed only 5'*MYC* amplification (Fig. 2.4h), which also indicates that ecMYC boundaries exhibited intricate transformations in GBM39-ER cells. Boundary variations in ecDNA did not exhibit specificity with relation to clusters (Fig. 2.4g). We next selected the same number of cells from either GBM39 or GBM39-ER within the transition cluster (C0) and compared the aggregated contacts map at ecMYC regions to see whether an intermediate state for ecMYC structure can be found. Both profiles, although from the same cluster, still resembled their original samples and showed the same differences in ecMYC structure, instead of showing a similar intermediate structure (Extended Data Fig. 2.6h). Therefore, we reasoned that for GBM39 cells, erlotinib treatment likely induced the formation of unique cell populations containing distinct *MYC* ecDNA, rather than selectively enriched a pre-existing cell population harboring ecMYC with growth advantage upon treatment.

Droplet Hi-C detects ecDNA in primary tumor samples

ecDNA is not only a biomarker for poor prognosis but also a crucial driver in GBM⁵⁴. As Droplet Hi-C successfully revealed dynamic changes of ecDNA in GBM39 cancer cells associated with drug treatment, we further sought to apply Droplet Hi-C to an IDH-wildtype GBM tumor

sample (Fig. 2.5a). By employing only chromatin structure information for cell embedding, we first segregated the cells into malignant and nonmalignant cells (Fig. 2.5b), featuring expected CNV events in malignant cells such as loss of chromosome 10 and extensive amplification of chromosome 7 (Fig. 2.5c). We found that the genomic region with the most variable copy numbers on chromosome 7 carried the oncogene *EGFR* (Fig. 2.5c-d, Extended Data Fig. 2.7a-b), and also displayed an ecDNA-like contact pattern in the contact map of the malignant population (Fig. 2.5e). Utilizing our ecDNA caller, we found significant enrichment of ecEGFR in the malignant population as expected (Fig. 2.5f). When we performed FISH of *EGFR* gene on frozen tissue sections, we observed that part of cells displayed dispersed EGFR amplicons across the nuclei (Fig. 2.5f right bottom). In addition, Droplet Hi-C also identified SVs specific to malignant population in the patient sample (Fig. 2.5g, Extended Data Fig. 2.7c). For example, based on the sequencing reads coverage and abnormally high interaction frequency, *IKZF1* gene, encoding a tumor suppressor, was found to translocate to ecEGFR without the TSS sequence, leading to its transcriptional disruption in the malignant cell population (Fig. 2.5g). Among the genes located on the ecDNA, the *VOPPI* gene exhibited the strongest interaction frequency with *IKZF1*, suggesting their fusion (Fig. 2.5g). Interestingly, this fusion event did not show any discernible effect on the transcriptional activity of *VOPPI*, based on the aggregated expression profiles from single-cell RNA-seq data (Fig. 2.5g).

At chromatin interactions level, compartment shows extensive changes in the malignant population compared to the nonmalignant population (Extended Data Fig. 2.7d). These changes show consistency with the variation in gene expression, for example, the compartment switch from inactive B to active A overlapping malignant lncRNA marker *SOX2-OT* is consistent with its increased transcription in tumor sample. Previous single-cell RNA-seq profiling of gliomas

defined four major cellular states for malignant-like cells, including neural progenitor-like (NPC-like) cells, oligodendrocyte progenitor-like (OPC-like) cells, astrocyte-like (AC-like) cells and mesenchymal-like (MES-like) cells⁵⁵. To further recapitulate different cellular states and identify intra-tumoral heterogeneity of chromatin conformation, we used k-nearest neighbors to assign malignant cells to one of the four GBM cellular states based on co-embedding with snRNA-seq data generated from the same sample (Fig. 2.5h-i). Correlation analysis using compartment scores shows similar cell states hierarchy compared to transcriptome data and previous study, where MES-like state exhibit higher similarity to AC-like state, while OPC-like and NPC-like states are more similar (Extended Data Fig. 2.7e)⁵⁶. Variation in compartment patterns across diverse cellular states is also consistent with state-specific gene expression. For instance, metastasis suppressor *CSMD1* is in A compartment in OPC-like and NPC-like states but is in B compartments in MES-like and AC-like states, echoing with the gene expression variation observed from snRNA-seq data (Extended Data Fig. 2.7f-g). To systematically compare the difference in chromatin structures across cellular states, we further focused on the differential compartments across the relatively distinct GBM states, MES-like and OPC-like, between which 1,782 switched compartments can be identified (Extended Data Fig. 2.7h). Regardless of the modest overlap between switched compartments and compartments harboring differentially expressed (DE) genes, we found that the switched compartments are significantly associated with differential gene expression (Extended Data Fig. 2.7i-j). Besides, differential genes with compartment switched are frequently associated with state-specific biological function based on the GO analysis results (Extended Data Fig. 2.7k), indicating that the variation in chromatin compartments is functionally meaningful. Disruption of chromatin insulation is known to rewire enhancer and promoter interactions and has been demonstrated to contribute to the progression of GBM⁵⁷. With Droplet

Hi-C data, we were able to observe example of insulation score alteration accompanying loss of domain boundary in malignant population (Extended Data Fig. 2.7l). In summary, with this comprehensive analysis of chromatin organization data, we showed that Droplet Hi-C has the power in dissecting the intra-tumoral heterogeneity of chromatin conformation.

To investigate heterogeneity in ecDNA, we subsequently characterized differential ecDNA features across different cellular states. We found that the fractions of cells containing ecEGFR were similar among the four cell states (Fig. 2.5j), but the cells in AC- and MES-like states contained higher copy numbers of ecEGFR (Fig. 2.5k). The increase of copy number was associated with higher *EGFR* expression level and chromatin accessibility in these states (Fig. 2.5l-m). Furthermore, by defining ecDNA boundary with Droplet Hi-C data, we found that expression levels of genes on the ecDNA showed varied level of correlation with the cellular states score, and that such correlation could not be fully explained by ecDNA copy number, indicating a complex regulatory mechanism governing ecDNA genes expression across cellular states (Fig. 2.5n). Consistent with previous reports⁵⁵, the expression level of *EGFR* exhibited highest correlation with AC-like state scores (Fig. 2.5n). Examples besides *EGFR* included the preferential expression of known *EGFR* co-amplified gene *SEC61G* in MES-like state, and the preferential expression of *LANCL2* in AC- and MES-like states (Fig. 2.5o). The abundance of ecDNA seems to be more closely related to the differentiation states (AC- and MES-like), although mechanism dictating varied expression levels of ecDNA genes besides copy number effects requires further investigation.

To further demonstrate the utility of Droplet Hi-C in clinical tumor samples, we adopted it to analyze the bone marrow mononuclear cells (BMMC) from an acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS) patient before and after treatment with azacitidine and

venetoclax. Azacitidine is a DNA methyltransferase inhibitor known to slow the growth of AML cells in part by promoting their differentiation into more mature cells (Extended Data Fig. 2.8a). Venetoclax is an inhibitor of the anti-apoptotic protein BCL2. The patient entered a remission from secondary AML after completing one cycle of therapy. The contact maps revealed that the before-treated BMMC initially harbored a strikingly large (~5 MB) ecDNA containing the *MYC* gene. After treatment, the ecMYC can no longer be detected by Droplet Hi-C (Extended Data Fig. 2.8b). To classify malignant and nonmalignant cells in BMMC, we leveraged known mutations detected by targeted bulk DNA sequencing to distinguish malignant cells. After drug treatment, the proportion of cells harboring mutations also diminished (Extended Data Fig. 2.8c). Clinical karyotype analysis showed the persistence of trisomy 8 which was present while the patient had MDS prior to AML transformation. Thus, we speculated that remaining mutant reads might originate from pre-leukemic MDS cells.

Expression of the oncogene *MYC* is up-regulated by numerous tissue-specific enhancers through long-range interactions in different cancers⁵⁸. One of the evolutionarily conserved enhancer clusters (blood enhancer cluster, BENC), which is ~1.8 Mb downstream of *MYC*, has been shown to activate *MYC* expression in hematopoietic stem cells and AML⁵⁸. The long-range interactions between *MYC* and BENC were observed in the before-treated BMMC sample. The detection of focal amplification of BENC is also consistent with a previous report⁵⁸ (Extended Data Fig. 2.8d). Specifically, both the *MYC* gene and BENC were located on the ecMYC. The interaction between *MYC* and BENC is obvious in the before-treated sample and disappears after treatment, suggesting its role in tumor initiation and progression (Extended Data Fig. 2.8e). Thus, in this patient's tumor, ecDNA may promote *MYC* expression not only through increased DNA copy number but also cancer-specific enhancer-promoter interactions.

In summary, Droplet Hi-C facilitates the detection of ecDNA dynamics in tumors before and after drug treatment. Besides ecDNA, identification of canonical chromatin structures such as SVs and chromatin loops also helps to elucidate the ectopic regulatory programs responsible for tumor progression and drug resistance. Droplet Hi-C therefore shows great potential for advancing understanding of mechanisms driving tumor clonal evolution and sensitivity to treatment.

Droplet-based joint profiling of chromatin architecture and transcriptome

Single-cell joint profiling of chromatin organization and transcriptome can facilitate the investigation of the relationships between gene expression and chromatin architecture. We modified the above Droplet Hi-C protocol to be compatible with the 10x Genomics Chromium Single Cell Multiome kit, to achieve simultaneous profiling of RNA and Hi-C from single nuclei (Fig. 2.6a, Extended Data Fig. 2.9), termed Paired Hi-C. The major modification in Paired Hi-C is that a milder SDS treatment condition is used before restriction enzymes digestion to minimize RNA degradation. Meanwhile, we used a lower formaldehyde concentration for cross-linking to increase the complexity of the chromatin contact profiles. Third, we noticed that the length of cDNA became shorter post SDS treatment, thus we revised the cDNA library preparation protocol to recover more cDNA molecules. Like Droplet Hi-C, Paired Hi-C also offers benefits in terms of throughput and hands-on time over other microwell- or combinatorial indexing- based joint Hi-C/RNA-seq single-cell assays (Extended Data Fig. 2.9).

First, we performed species mixing experiment using the same cell mixtures as Droplet Hi-C (HeLa S3 and mESC). The results showed that human and mouse cells could be successfully separated in both Hi-C (HeLa: 3928, mESC: 3301, total: 7677) and RNA (HeLa: 3999, mESC:

3293, total: 7677) modalities data (Extended Data Fig. 2.10a-b). To demonstrate the utility of Paired Hi-C, we applied it to a cortex specimen collected from an 8-week old mouse. Overall, we obtained 12,361 joint single-cell profiles of transcriptome and chromatin structure with 42,210 read pairs per cell (duplication rate 66.2%), and a median of 3,914 UMI per cell and 1,746 genes per cell (duplication rate 74%, Extended Data Fig. 2.10c). Using the published Droplet Paired-Tag RNA modality as a reference for label transfer, we successfully identified 20 subclasses in the mouse cortex Paired Hi-C RNA data, including 9 excitatory neuron subtypes, 6 inhibitory neuron subtypes, and 5 non-neuronal subtypes (Fig. 2.6b). The accuracy of subclass annotation was validated by integration with other sn- or scRNA-seq datasets and the unique expression patterns of marker genes (Extended Data Fig. 2.10d-f). While single-cell 3D genome features were relatively limited in Paired Hi-C, the compartment score computed from Paired Hi-C data for each cell type (as defined by single-nuclei RNA-seq) exhibited strong correlation with Droplet Hi-C and sn-m3C-seq profiles from the matched annotations (Fig. 2.6c-d, Extended Data Fig. 2.10g). At cell type-specific marker genes, we can also detect concordant changes of compartment and gene expression (Fig. 2.6e). As an example, *ErbB4* genes were located within the A compartment in PVGA and VIPGA GABAergic neurons, while they were placed in the B compartment in OPC non-neuronal types and ITL5GL glutamatergic neurons. The expression level of *ErbB4* was notably higher in PVGA and VIPGA than OPC and ITL5GL (Fig. 2.6c). We also utilized transcriptome information from Paired Hi-C and co-embed it with Droplet Hi-C. The results agree well with the cell annotations based on Droplet Paired-Tag (Extended Data Fig. 2.10h). As such, we can augment the number of chromatin contacts for each cell type by combining Paired Hi-C and Droplet Hi-C.

As a multiomic profiling method, Paired Hi-C provides joint information at single-cell resolution and can directly link chromatin structure with its associated gene expression, which can help us better understand the mechanism regulating gene activation or repression. Thus, we revisited the erlotinib treated GBM39 cancer cell system with Paired Hi-C to analyze how alteration in chromatin structure, especially in ecDNA affect gene expression with higher accuracy. The heterogeneity in the ecMYC boundaries at single-cell level within GBM39-ER prompted the question of whether this variable ecMYC boundary could result in varying ecDNA gene expression within individual cells. By applying Paired Hi-C to GBM39 and GBM39-ER, we were able to link ecDNA structure with expression level of its related genes in single cells. We found that in general, the expression level of genes located on ecMYC within GBM39 and GBM39-ER are positively correlated with their underlying copy number, indicating a higher probability as part of ecMYC (Fig. 2.6f). We then classified genes based on their positions on ecDNA (Fig. 2.6g). For genes located 5' ecMYC boundaries such as *CASC8* and *PCAT1*, their expression exhibited a strong correlation with the copy number of underlying DNA segments (Fig. 2.6f). In the case of ecMYC constant genes, such as *PVT1*, expression was constantly high and positively correlated with local contact numbers (Fig. 2.6f). Unexpectedly, the expression of the *CCDC26* gene, located at the 3' end of the ecMYC variable region, was high in GBM39-ER cells despite low copy number (Fig. 2.6f). When we conducted correlation analysis between gene expression levels and copy numbers for the genes at the 5' and 3' ends of the variable ecMYC boundary separately in GBM-ER cells, the findings revealed that genes located within the 5' ecMYC variable region exhibited a more pronounced increase in expression with rising copy numbers. In contrast, genes within the 3' ecMYC variable region showed much lower correlations between gene expression levels and copy

numbers (Fig. 2.6g). These results highlight the complex relationship between amplicon structure and gene expression.

Besides ecMYC structural heterogeneity and gene expression variation, we found that ecMYC displayed different *trans*-interaction patterns in GBM39 and GBM39-ER cells. These different *trans*-interaction patterns of ecDNA were also associated with distinct transcriptional responses. Specifically, we observed that GBM39-specific interacting genes showed higher expression levels in GBM39 than in GBM39-ER cells, while GBM39-ER-specific interacting genes had higher expression levels in GBM39-ER cells (Extended Data Fig. 2.11a). We also plotted the copy number of significantly interacting regions in GBM39 versus GBM39-ER (Extended Data Fig. 2.11b) to ensure that GBM39-specific or GBM39-ER-specific interacting genes showing higher expression levels were not due to their elevated copy number. This suggests a potential mobile regulatory role of ecDNA in global chromosomal transcription⁴⁹.

In the GBM patient sample, we found that ecEGFR showed higher copy numbers in AC-like and MES-like cell populations as well as higher expression levels of *EGFR*, which may render the cellular states changes under erlotinib treatment. Using previously defined methods, we utilized Paired Hi-C single-cell RNA-seq data to classify GBM39 and GBM39-ER cells into four GBM cellular states, and found a significant reduction of differentiated-like population after erlotinib treatment (Fig. 2.6h). We previously demonstrated that GBM39-ER cells were characterized by the loss of ecEGFR and the emergence of ecMDM2, and we sought to determine whether the expressions of these genes enriched in particular states within the resistant cell population. As found in the GBM tumor sample, *EGFR* expression was higher in AC-like and MES-like cells (Fig. 2.6i). Interestingly, *MDM2* expression was enriched in the progenitor states (OPC- and NPC-like), which became more prevalent after treatment (Fig. 2.6i). These results

highlight the unique interplay between ecDNA dynamics, transcriptionally defined cellular states, and drug resistance.

Besides ecDNA, chromatin architecture also undergoes a dramatic shift during erlotinib treatment, along with changes in gene expression. In GBM39 cells following erlotinib treatment, we observed 1,066 A compartments transitioned into B compartments, whereas 2,796 B compartments shifted to A compartments. The majority of compartment identities remained stable throughout the treatment, including 10,930 A compartments and 11,876 B compartments (Extended Data Fig. 2.11c). This compartment shift is also accompanied with gene expression changes. Specifically, genes situated within compartments that changed from A to B exhibited a significant decrease in expression, while those transitioning from B to A showed a significant increase in expression. In contrast, gene expression within stable compartments displayed only slight expression alterations (Extended Data Fig. 2.11d).

In summary, by incorporating transcriptomic data with chromatin structure information, Paired Hi-C enables direct interrogation of the effects of chromatin reorganization and structural alterations like ecDNA on gene expression, providing a valuable tool for characterizing gene regulatory programs in development and disease pathogenesis.

2.4 Discussion

In recent years, there has been a significant surge in the development and application of single-cell Hi-C methods. However, current single-cell Hi-C methods still face challenges when applying to heterogeneous tissues and tumor biopsies. Our Droplet Hi-C leverages the

commercially available microfluidic platform to provide high-throughput and fast single-cell Hi-C assays with minimal hands-on time, and relatively low costs. Utilizing Droplet Hi-C technology on the adult mouse cortex, we identified cell type-specific hierarchical chromatin structures and established correlations with epigenetic modifications. Additionally, our study unveiled alterations in chromatin organization, CNV, SVs, and ecDNA in cancer cells and tumor patient samples, especially during drug treatment. To promote the application of the Droplet Hi-C method, we quantified the minimum number of cells needed for accurate chromatin organization feature analysis. We down-sampled the mouse cortex cell type ITL23GL (Cortex L2/3 excitatory neurons) and calculated the compartment score (at 100-kb resolution) and insulation score (at 25-kb resolution, with a sliding window size of 125 kb) for each group (Extended Data Fig. 2.12a-c). We observed a significant drop in Spearman correlation when the cell number fell below 300, corresponding to fewer than 7.5 million long-range interactions (Extended Data Fig. 2.12b). This value aligns with previous reports on the required number of cells to observe large chromatin features like compartments⁵⁹. However, we expect that more reads are required to confidentially call and analyze chromatin loops.

We further extend Droplet Hi-C to Paired Hi-C, which achieves joint profiling of chromatin architecture and transcriptome in single cells. This enables the association of gene expression phenotypes with 3D genomic structure. The complexity of the Hi-C modality in Paired Hi-C is notably lower when compared to Droplet Hi-C. Nonetheless, we have shown that it is possible to leverage transcriptome data obtained from Paired Hi-C experiments for co-embedding with Droplet Hi-C data. This approach enabled us to resolve and annotate cell types and supply Hi-C contact information for each cell type to enhance the resolution of chromatin organization. EcDNA was first identified in 1965, described as double minutes (DM), and is prevalent in cancer as they

often harboring oncogenes^{60, 61}. Unlike the kilobase-sized circular DNA (eccDNA) found in healthy somatic tissues⁶², ecDNA varies in size from dozens of kilobases to megabases, making it 100 to 1,000 times larger⁶³. The ecDNA is characterized by high amplification and a circular structure⁶⁴. Current single-cell ecDNA detection methods can be broadly categorized into two primary groups. The first group involves the use of computational tools and WGS data obtained from NGS⁴⁷ or TGS^{65, 66} to identify circularization sites. However, distinguishing between ecDNA and HSR using this approach can be challenging, because they have the same breakpoint sequences. The second group is based on Circle-seq⁶⁷⁻⁶⁹, a strategy that utilizes exonuclease to degrade linear DNA to retain circular DNA for sequencing. This method allows for the differentiation of ecDNA from linear tandem repeats. However, Circle-seq-based strategies have been associated with a high false-positive rate due to incomplete degradation of linear DNA and multiple displacement amplification (MDA). Accurate identification of ecDNA using its multiple features at single-cell level remains challenging. Here, we have developed an innovative ecDNA detection algorithm that utilizes both *trans* and *cis* interactions data from Droplet Hi-C datasets, enabling precise ecDNA identification while effectively distinguishing it from HSR. With our deep learning-based ecDNA caller algorithm, we can reliably detect ecDNA in both cancer cell lines and clinical tumor samples in each single cell, and calculate their proportion within the cell population. Furthermore, the increased occurrence of intra-ecDNA contacts provides valuable support for pinpointing ecDNA boundaries at the single-cell level. For instance, we observed variations in the ecMYC boundary within the GBM39-ER cell population. Additionally, we are capable of detecting SVs events within both ecDNA and HSR.

Pan-cancer studies have demonstrated the detection of ecDNA in numerous human cancer types and that ecDNA is associated with inferior clinical outcomes across all amplification

classes⁷⁰⁻⁷⁴. Further, new studies have begun to highlight unique therapeutic vulnerabilities of tumors that harbor ecDNA^{75, 76}. Thus, given the independent prognostic value of ecDNA and its potential to be targeted, recent calls have been made to include ecDNA profiling in clinical specimens⁷⁷. Droplet Hi-C and deep learning-based ecDNA caller algorithm offer a reliable toolkit for the identification of ecDNA in clinical samples as demonstrated in patient GBM and AML tumors. This toolkit may be particularly useful in heterogeneous patient samples that harbor ecDNA species whose structures and copy numbers may evolve subclonally through treatment. Importantly, our ecDNA caller algorithm possesses the unique capability to differentiate between amplifications occurring on ecDNA and HSRs. Such differentiation may carry critical therapeutic and clinical implications. In our study, we observed the disappearance of ecEGFR, emergence of ecMDM2 and structural dynamics of ecMYC in GBM39 cells following treatment with erlotinib, a tyrosine kinase inhibitor commonly used in the management of *EGFR*-positive tumors. Notably, the effectiveness of erlotinib treatment in GBM was poor, possibly attributed to the ecDNA evolution. Additionally, in an AML patient sample, we noted that ecMYC could no longer be detected after treatment with azacitidine and venetoclax. Collectively, our methods can be used in ecDNA detection in clinical samples and hold promise for advancing research related to ecDNA in tumor diagnosis and treatment.

2.5 Acknowledgement

Chapter 2 is, in full, a reformatted reprint of the material as it appears as “Droplet Hi-C for Fast and Scalable Profiling of Chromatin Architecture in Single Cells” in *bioRxiv*, 2024. Lei Chang, Yang Xie, Brett Taylor, Zhaoning Wang, Jiachen Sun, Tuyet R. Tan, Rafael Bejar, Clark C. Chen,

Frank B. Furnari, Ming Hu, Bing Ren. The dissertation author was the primary investigator and co-first author of this paper.

We are grateful to Dr. Jingtian Zhou for providing suggestions in data analysis. We thank Dr. Chenxu Zhu for his invaluable suggestions on this study. The work was supported by NIH grants 1UM1HG011585 (to B.R. and M.H.), R35HG011922 (to M.H.), R56 NS080939 and R01 CA258248 (to F.B.F.), NIH training grants (5T32CA067754-27 and 5T32GM007198-47 to B.T.). Z.W. is a DDBrown awardee of the Life Sciences Research Foundation. This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a National Institutes of Health SIG grant (#S10 OD026929). This publication also includes data utilized equipment and staff support at the UCSD School of Medicine Microscopy Core with grant (P30 NS047101). We thank Nam Nguyen in UC San Diego Health Cytogenetics/Medical Genetics Laboratory for supporting in DNA FISH.

2.6 Figures

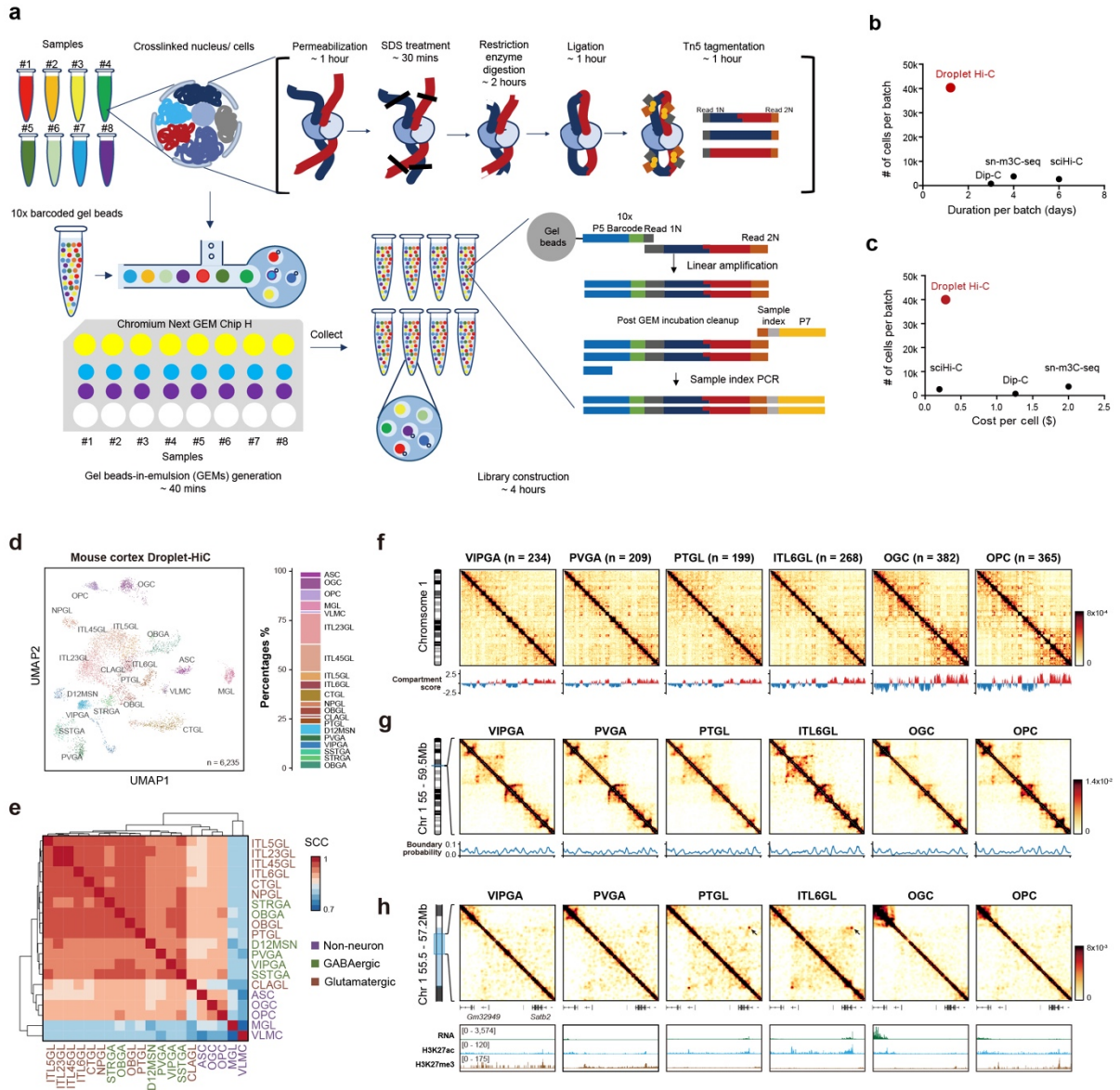
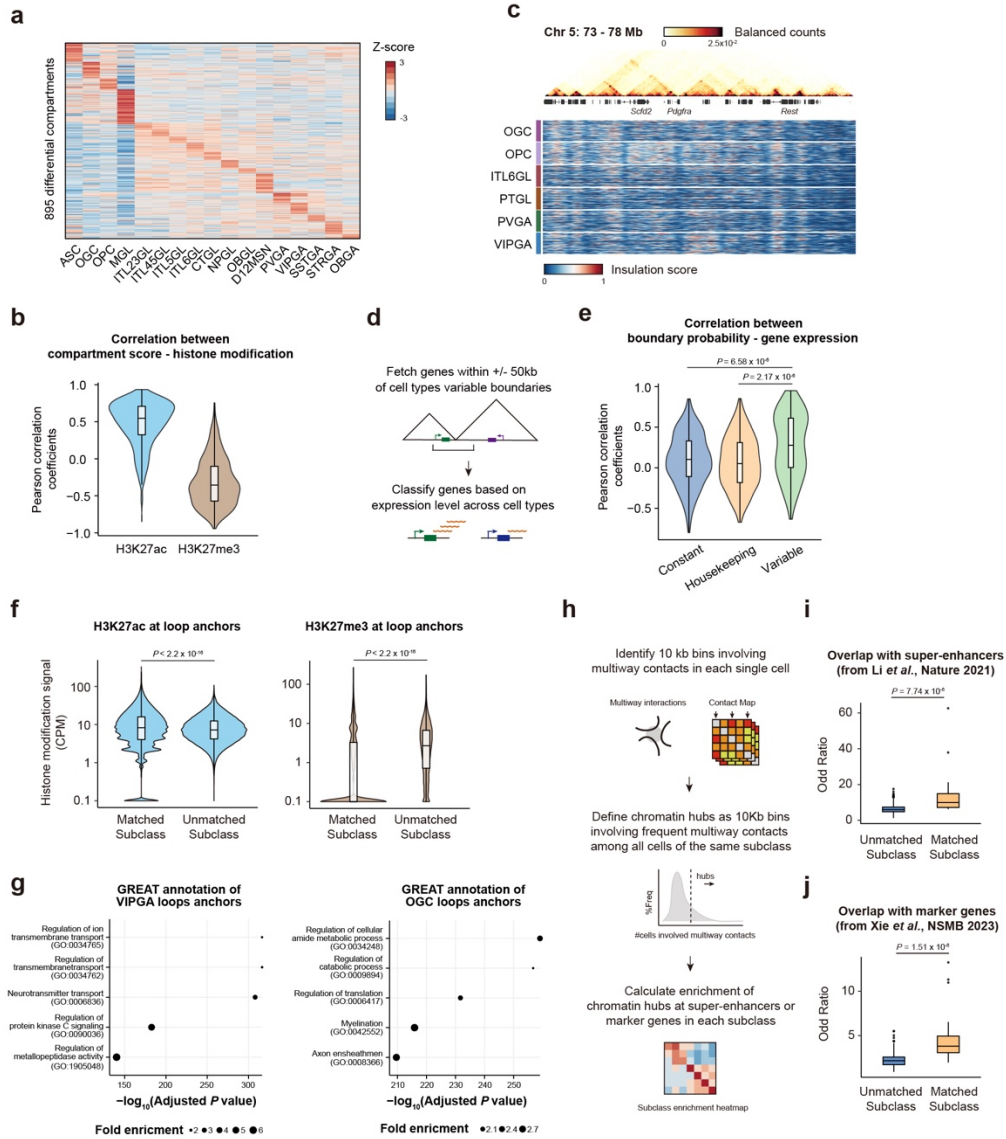


Figure 2.1: Overview and performance of Droplet Hi-C.

a, Schematic of the Droplet Hi-C workflow. **b-c**, Comparison of throughput, sample preparation time, and cost among different single-cell Hi-C methods. **d**, UMAP visualization of Droplet Hi-C data from adult mouse cortex. **e**, Genome-wide Spearman's correlation coefficients (SCC) between compartment scores from different cell types. **f**, Cell type-specific Droplet Hi-C contacts map from chromosome 1 and compartment score at 100-kb resolution. The color bar represents the imputed contact number. **g**, Cell type-specific Droplet Hi-C contacts map and boundary probability of example region (chromosome 1: 55 - 59.5Mb) at 25-kb resolution. **h**, Cell type-specific Droplet Hi-C contacts map surrounding gene *Satb2* (chr1: 55.5 - 57.2Mb) at 10-kb resolution, along with genome browser view showing transcriptome and histone modifications profiles in the same cell types from the public datasets.

Figure 2.2: Multi-scale chromatin structures are correlated with cell type-specific gene activity in adult mouse cortex.

a, Heatmap showing compartment score of differential compartments among all subclasses. **b**, Distribution of Pearson correlation coefficients between compartment score and histone modification signals at each 100-kb bin across differential compartments; $n = 895$ (H3K27ac and H3K27me3). All box plot hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median and whiskers denoting $2\times$ the interquartile range. **c**, Comparison of single-cell insulation scores from example subclasses surrounding gene *Pdgfra*. A bulk contacts map at 10-kb resolution is shown above. **d**, Schematics for correlation analysis between boundary probability and nearby gene expression level. **e**, Violin plot showing Pearson correlation coefficient between gene expression level and boundary probability at variable domain boundaries across different subclasses in mouse cortex, with genes selected for clustering being excluded for analysis. Genes are classified as constant, housekeeping, and variable based on snRNA-seq reference and literature. P values were calculated by Wilcoxon rank-sum test. All box plot hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median and whiskers denoting $2\times$ the interquartile range. **f**, Comparison of histone modification signal enrichment at loop anchors across different cell types; $n = 265,277$ (all groups); All box plot hinges were drawn from the 25th to 75th percentiles, with the middle line denoting the median and whiskers denoting $2\times$ the interquartile range. P values, Wilcoxon signed-rank test. **g**, Top enriched gene ontology terms for genes at loop anchors in selected cell types VIPGA and OGC. P values and fold enrichment were calculated by binomial test. Benjamini–Hochberg FDRs were then calculated to select overrepresented GO terms. **h**, Diagram describing the process in identifying cell type-specific chromatin hubs involving multi-way interactions from Droplet Hi-C data. **i**, Box plots showing the overlap between chromatin hubs and super-enhancers in matched versus unmatched mouse cortex subclasses. Odd ratio is calculated by Fisher exact test. **j**, Box plots showing the overlap between chromatin hubs and cell type-specific marker genes in matched versus unmatched subclasses.



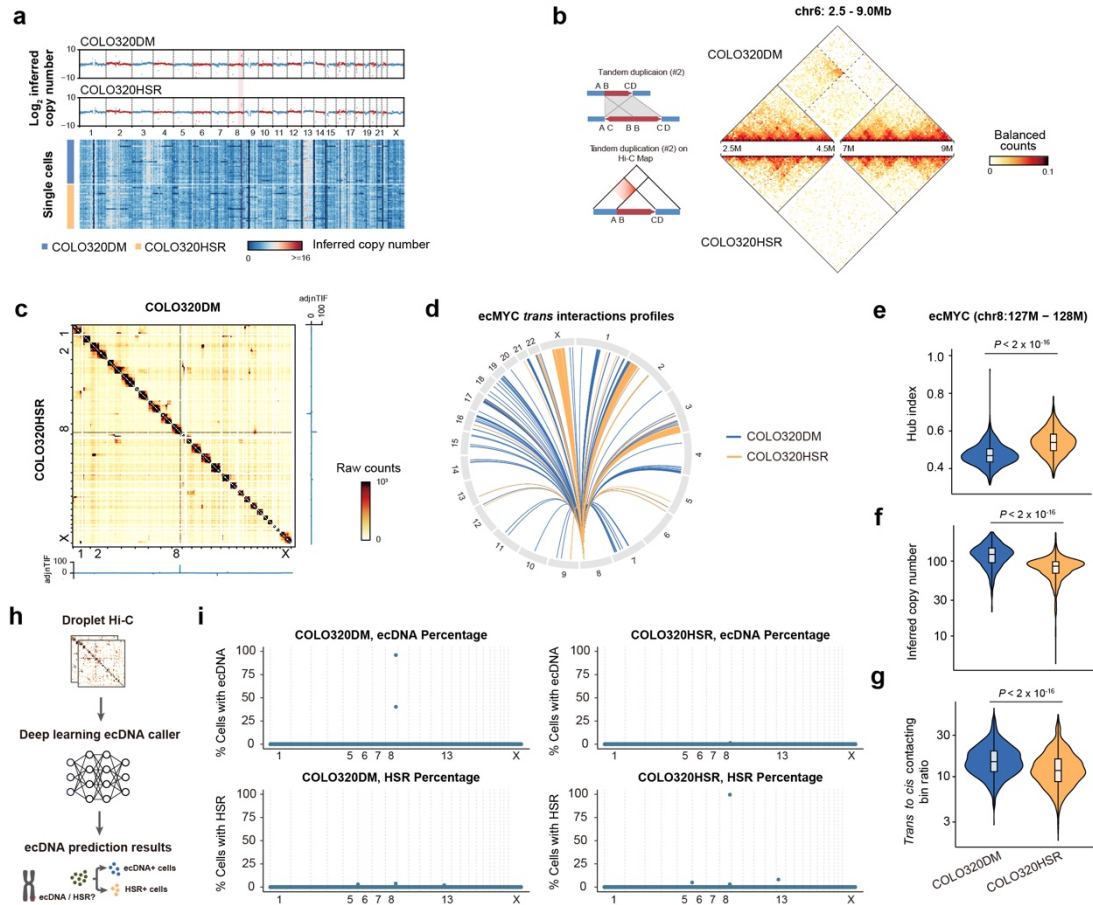


Figure 2.3: Droplet Hi-C illuminates copy number, structural variations and ecDNA in cancer cells.

a, Copy number inferred from pseudo-bulk Droplet Hi-C profiles in COLO320DM and COLO320HSR, with heatmap showing representative single-cell CNV in each sample. The ecMYC bins in COLO20DM are highlighted in pink. **b**, Example of a sample-specific SV on chromosome 6, predicted with EagleC. Cartoon illustration explaining the rearranged contacts pattern is shown aside. **c**, Comparison of genome-wide contact maps and adjusted normalized inter-chromosomal interaction frequency (adj_nTIF) between COLO320DM and COLO320HSR. **d**, Circos plot showing *trans*-interaction profiles of genomic bin containing *MYC* in COLO320DM or COLO320HSR. **e-g**, Distribution of single-cell hub index (**e**), copy number (**f**), and *trans*-to-*cis* contacting bin ratio (**g**) of ecMYC in COLO320DM and COLO320HSR. $n = 1,352$ (Hub index, COLO320DM), 1,366 (Hub index, COLO320HSR), 1,426 (Inferred copy number and *trans*-to-*cis* contacting bin ratio, COLO320DM), 1,535 (Inferred copy number and *trans* to *cis* contacting bin ratio, COLO320HSR); P values are from Wilcoxon signed-rank test. **h**, Schematics of deep learning-based ecDNA caller. **i**, Dot plot showing genome-wide ecDNA and HSR prediction results from deep learning-based ecDNA caller for COLO320DM and COLO320HSR cells.

Figure 2.4: Droplet Hi-C reveals ecDNA dynamics under drug treatment.

a, Illustration of the erlotinib treatment procedure for GBM39 cells. **b**, UMAP embedding visualization and clustering analysis of Droplet Hi-C data from GBM39 before- and after-erlotinib treatment. **c-d**, Pie charts showing the cell proportion of each cluster in GBM39 (**c**) and GBM39-ER (**d**). **e**, Comparison of contact maps and percentages of ecDNA-positive cells among all clusters for ecEGFR, ecMYC and ecMDM2. Genomic coordinates for key oncogenes are displayed aside. **f**, UMAP embedding visualization of ecDNA positive or negative cells in GBM39 and GBM39-ER. DNA FISH of the indicated ecDNA combinations in metaphase and interphase cells are shown as validation. EcDNA species are specified on top of UMAP plots Scale bar: 10 μ m. **g**, Pseudo-bulk contacts map at 10-kb resolution showing ecMYC local structure in both GBM39 and GBM39-ER samples. Heatmaps representing representative single-cell inferred CNV in the same genomic range are shown below. **h**, DNA FISH with probes targeting in 5' or 3' regions of *MYC* in GBM39 and GBM39-ER.

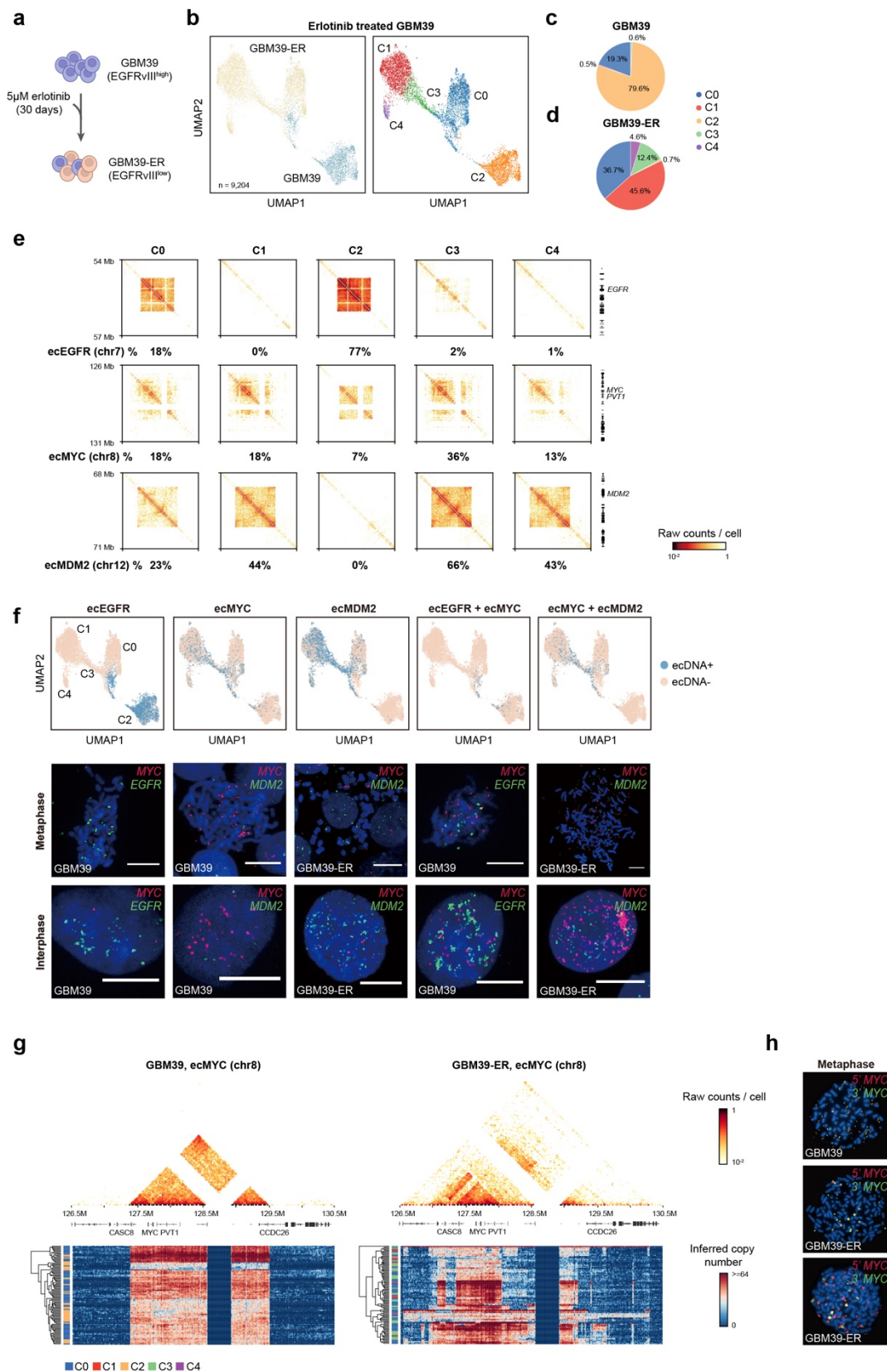
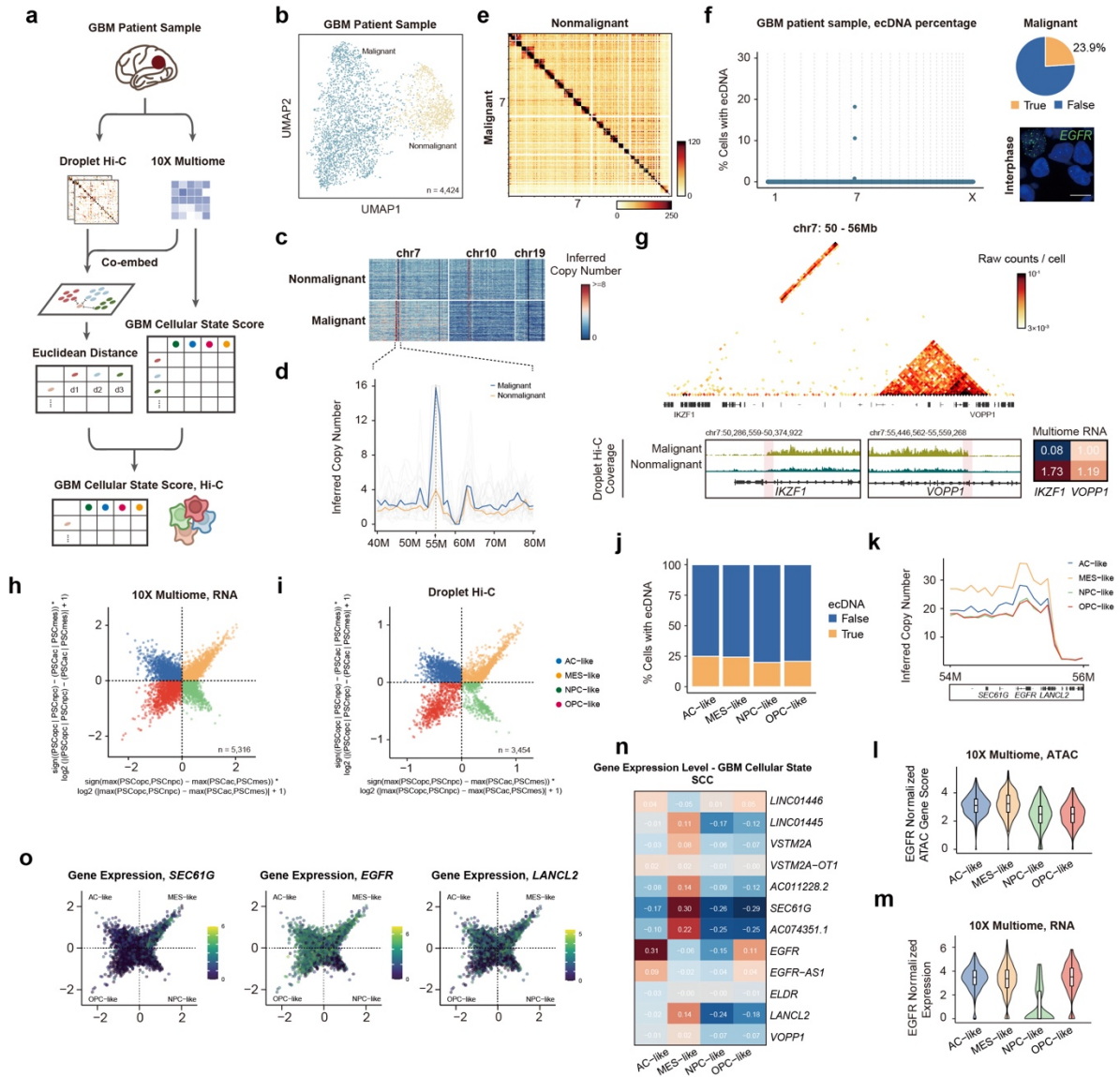


Figure 2.5: Droplet Hi-C depicts ecDNA variations in patient samples.

a, Schematics of GBM cellular state analysis for Droplet Hi-C data by co-embedding with reference 10x Multiome data. **b**, UMAP visualization of Droplet Hi-C data on GBM patient sample. **c**, Representative single-cell inferred CNV on chromosome 7, 10 and 19 in malignant and nonmalignant populations identified from Droplet Hi-C. **d**, Line plot showing copy number among single cells in malignant and nonmalignant populations. Lines with color highlighted the median profiles. The single-cell copy number profile examples are colored in gray ($n = 20$). **e**, Genome-wide contact maps from malignant and nonmalignant populations. Color bar shows the raw contact number. **f**, The ecDNA prediction results in GBM patient sample. The percentage of cells predicted to contain ecEGFR in malignant population is shown in a pie chart. DNA FISH of *EGFR* on frozen tissue sections from the GBM patient sample. Scale bar: 10 μm . **g**, An example of malignant population-specific SV. Genome browser view showing the Droplet Hi-C read coverage at associated genes displaying below. Predicted breakpoints of SV are highlighted in pink. Gene expression level (RPKM) of associated genes in different populations from 10x Multiome is also shown. **h-i**, Two-dimensional representation of cellular states based on scRNA-seq data from 10x Multiome (**h**) and Droplet Hi-C data (**i**). Each quadrant corresponds to one cellular state. **j**, The percentage of ecEGFR-positive cells across four different GBM cellular states. $n = 563$ (AC-like), 600 (MES-like), 247 (NPC-like), 346 (OPC-like). **k**, Inferred copy number at 100-kb resolution of regions on ecEGFR among different GBM cellular states. **l-m**, Violin plots showing *EGFR* ATAC gene score (**l**) and *EGFR* gene expression level (**m**) across four different GBM cellular states from 10X Multiome dataset. Only cells with GBM cellular states score pass cutoff (± 0.5) are shown. $n = 244$ (AC-like), 410 (MES-like), 125 (NPC-like), 263 (OPC-like). **n**, Heatmap showing Spearman's correlation coefficients (SCC) between GBM cellular states score and gene expression level for all the genes located on ecEGFR. **o**, Two-dimensional representation of GBM cellular states, colored by expression levels of indicated gene on ecDNA.



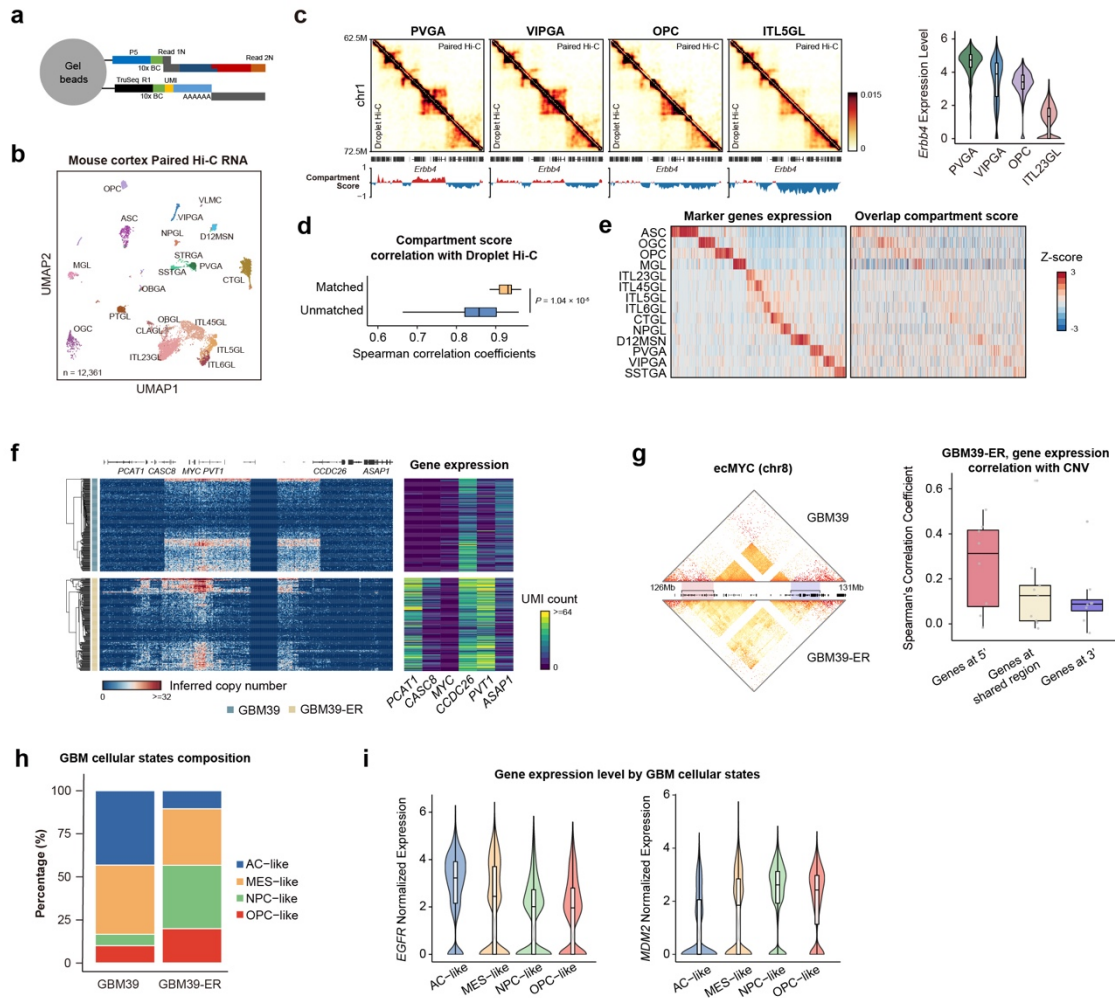


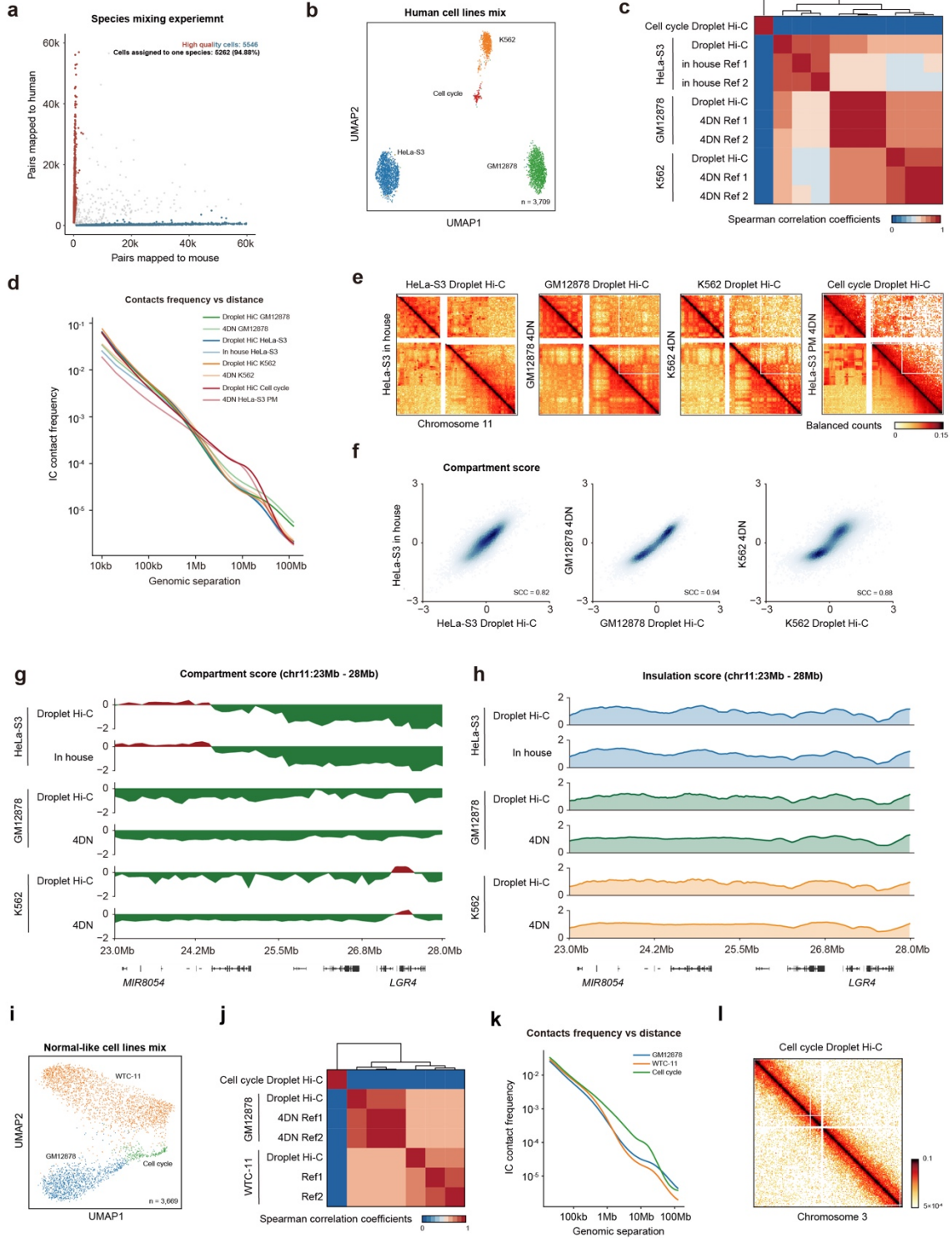
Figure 2.6: Joint profiling of single-cell Hi-C and transcriptome.

a, Schematic of the Paired Hi-C molecule barcoding step with 10X Multiome kit. **b**, UMAP visualization of mouse cortex single-cell clustering results from Paired Hi-C transcriptome data. **c**, Comparison of pseudo-bulk contact maps between Droplet Hi-C and Paired Hi-C at the region surrounding gene *Erbb4*, along with compartment score profiles from Paired Hi-C in representative subclasses. Color bar shows the imputed contact number. A violin plot of *Erbb4* expression level in representative subclasses is also shown. **d**, Box plots showing Spearman correlation coefficients (SCC) of compartment score between matched versus unmatched subclasses in Droplet Hi-C and Paired Hi-C. n = 15 (matched); 225 (unmatched). **e**, Heatmap showing marker genes expression levels and compartment scores of corresponding 100-kb bins among all subclasses. **f**, Single-cell inferred copy number heatmaps of region (chr8: 126.5 – 130.5 Mb) harboring ecMYC in GBM39 and GBM39-ER. Single-cell UMI counts heatmaps of representative genes on ecDNA are also shown. **g**, SCC between level of gene expression and copy number for ecMYC genes in 5' or 3' variable regions or shared regions in GBM39-ER. Illustration of ecMYC variable regions is shown on the left. n = 9 (5' gene); 10 (shared); 11 (3' gene). **h**, Comparison of cell proportions in four GBM cellular states between GBM39 and GBM39-ER samples. **i**, Comparison of expression levels of *EGFR* and *MDM2* among four GBM cellular states. n = 4,307 (AC-like), 6,791 (MES-like), 4,945 (NPC-like), 3,123 (OPC-like).

2.7 Supplementary Figures

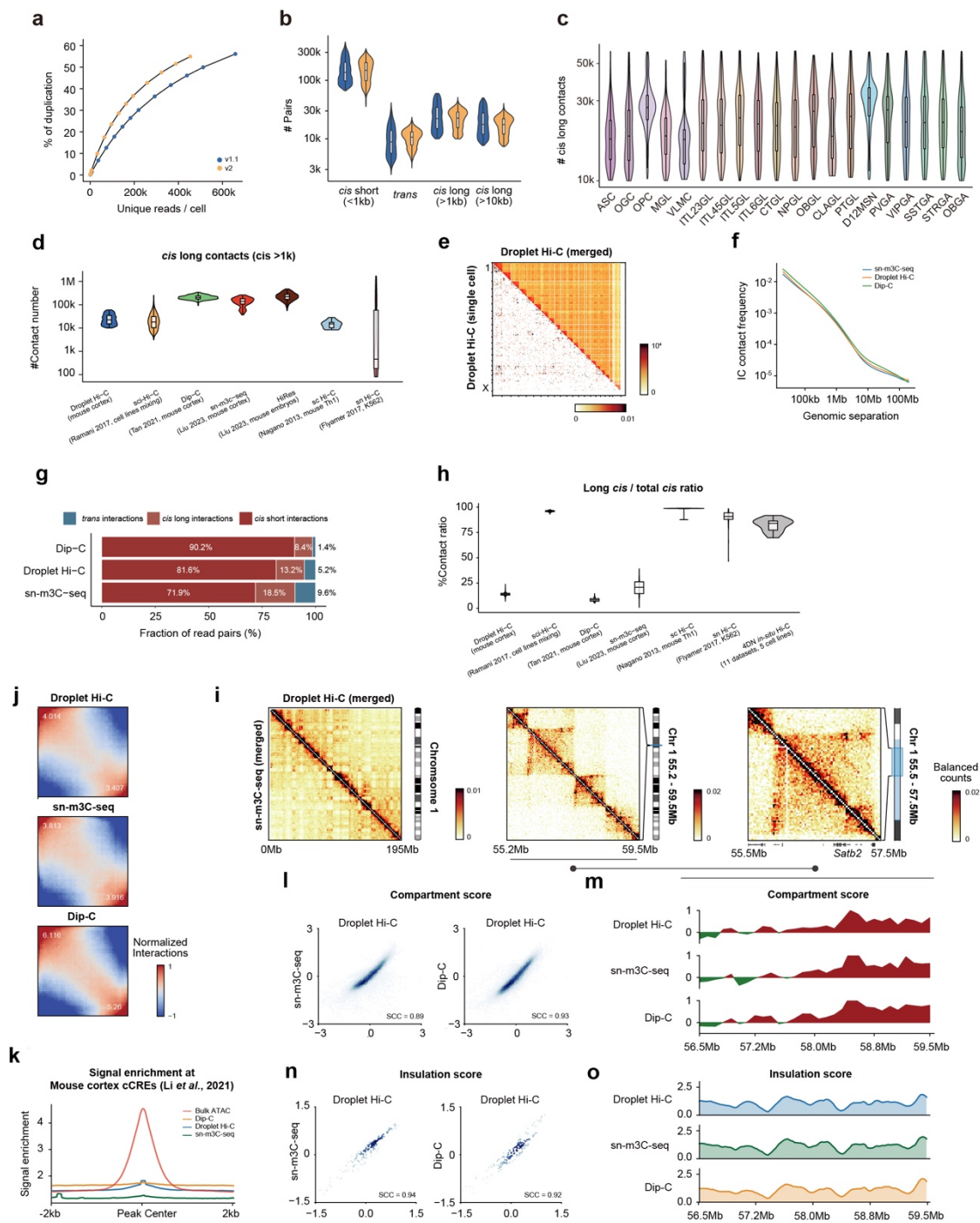
Extended Data Figure 2.1: Comparison between Droplet Hi-C and *in situ* Hi-C on cultured cells.

a, Scatter plots showing proportion of human and mouse DNA reads in each cell in the Droplet Hi-C species mixing experiment. **b**, UMAP visualization of Droplet Hi-C data on three human cell lines mixing sample. **c**, Genome-wide Spearman correlation coefficients (SCC) of compartment score between pseudobulk Droplet Hi-C data and bulk *in-situ* Hi-C datasets. **d**, Comparison of contact frequency by distance between Droplet Hi-C and *in situ* Hi-C datasets of HeLa S3, GM12878, K562 and cell cycle cells. **e**, Comparison of contact maps between Droplet Hi-C and Bulk *in situ* Hi-C for HeLa S3, GM12878, K562 and cell cycle cells on chromosome 11 at 100-kb resolution. **f**, Scatter plot showing compartment score values between Droplet Hi-C and *in situ* Hi-C for HeLa S3, GM12878 and K562 cells. **g-h**, Comparison of compartment score (**g**) and insulation score (**h**) profiles on example region (chr11: 23 - 28 Mb) from Droplet Hi-C and *in situ* Hi-C datasets for HeLa S3, GM12878 and K562 cells. **i**, UMAP visualization of Droplet Hi-C data from the cell lines mixed experiment with two normal-like human cell lines. **j**, Genome-wide spearman correlation coefficients (SCC) of compartment score between Droplet Hi-C data with reference bulk Hi-C data. **k**, Contact frequency by distance between pseudobulk Droplet Hi-C profiles in different clusters identified. **l**, Example Droplet Hi-C pseudobulk contact map on chr3 for cell cycle population.



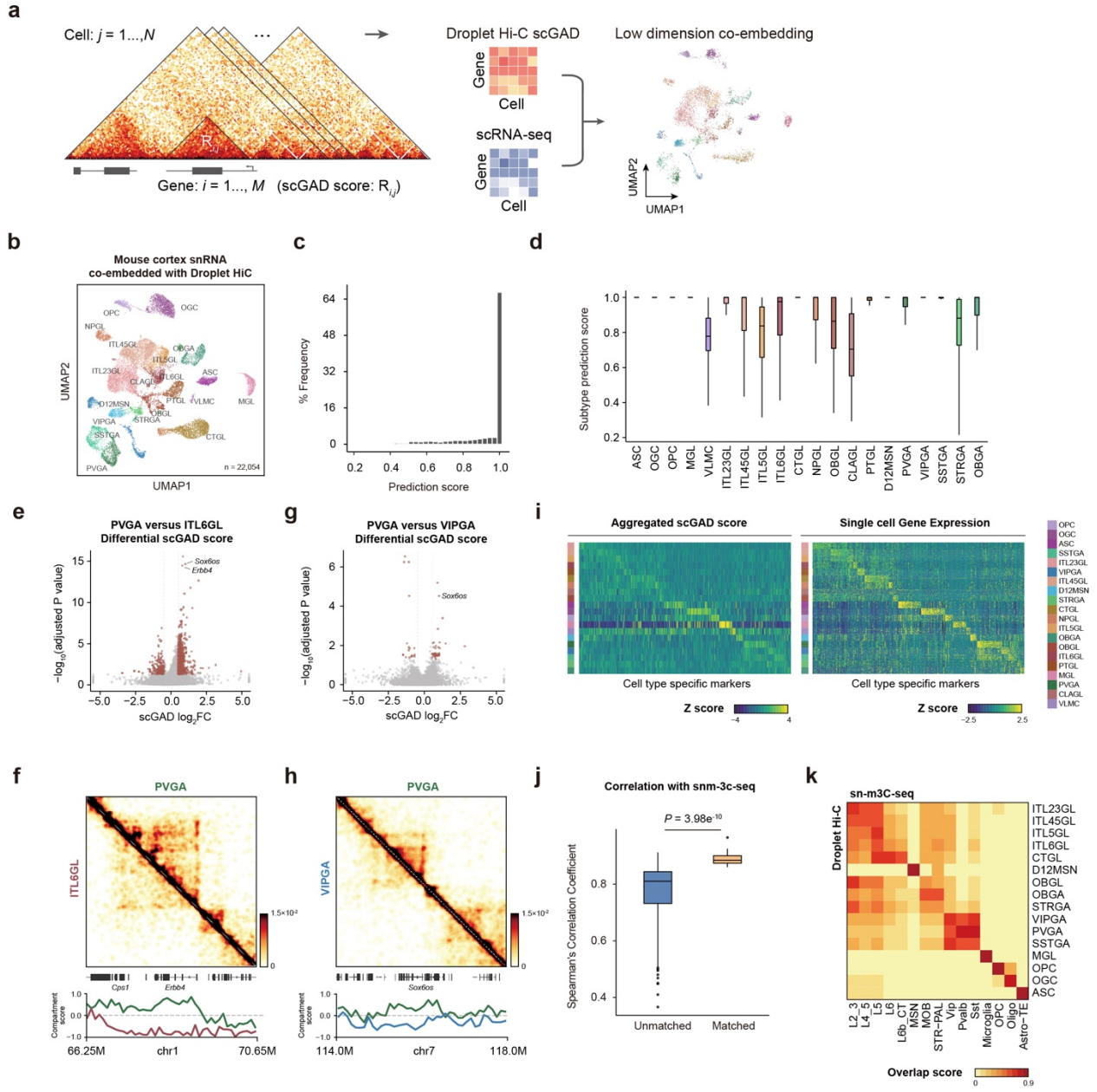
Extended Data Figure 2.2: Performance of Droplet Hi-C on adult mouse cortex.

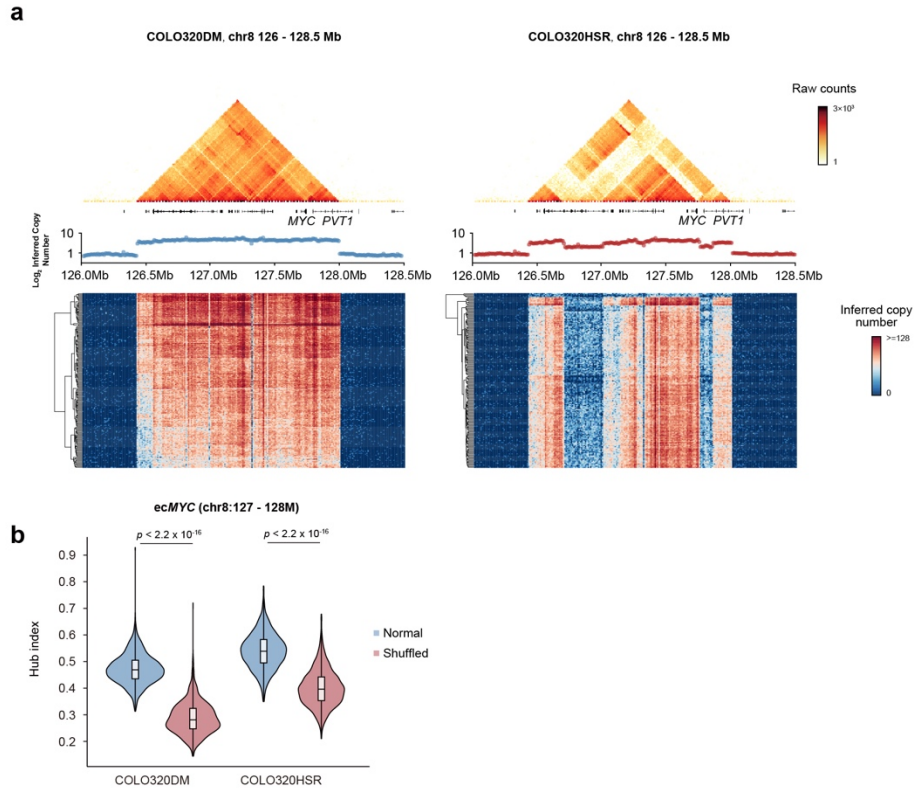
a, Comparison of library complexity by read pairs between two datasets generated with 10x Genomics Chromium Single Cell ATAC kit v1.1 and v2. **b**, Distribution of *cis*-short, *cis*-long or *trans*-contacts number per nucleus. **c**, Comparison of *cis*-long contact (>1kb) numbers distribution across different subclasses. **d**, Comparison of *cis*-long contact (>1kb) per cell across different single-cell Hi-C methods. **e**, Pseudo-bulk and representative single-cell genome-wide contact maps from mouse cortex. Color bar showed the raw contact number. **f-g**, Comparison of contact frequency by distance (**f**) and contacts ratio (**g**) on mouse cortex data among Droplet Hi-C, sn-m3C-seq and Dip-C. The *cis*-long interactions in (**g**) represent contacts separated by more than 1 kb. **h**, Comparison of *cis*-long contact ratio per cell across different single-cell methods and bulk *in situ* Hi-C. **i**, Comparison of multi-scale genome organization on chromosome 1 between Droplet Hi-C and sn-m3C-seq. **j**, The compartmentalization strength among Droplet Hi-C, sn-m3C-seq and Dip-C. Color bar showed the balanced contact number. **k**, Enrichment of sequencing reads signal over mouse cortex cCREs among Droplet Hi-C, sn-m3C-seq, Dip-C and bulk ATAC-seq. **l-o**, The compartment scores (**l-m**) and insulation scores (**n-o**) among Droplet Hi-C, sn-m3C-seq and Dip-C are also shown.



Extended Data Figure 2.3: Validation of cell type identities in mouse cortex characterized by Droplet Hi-C.

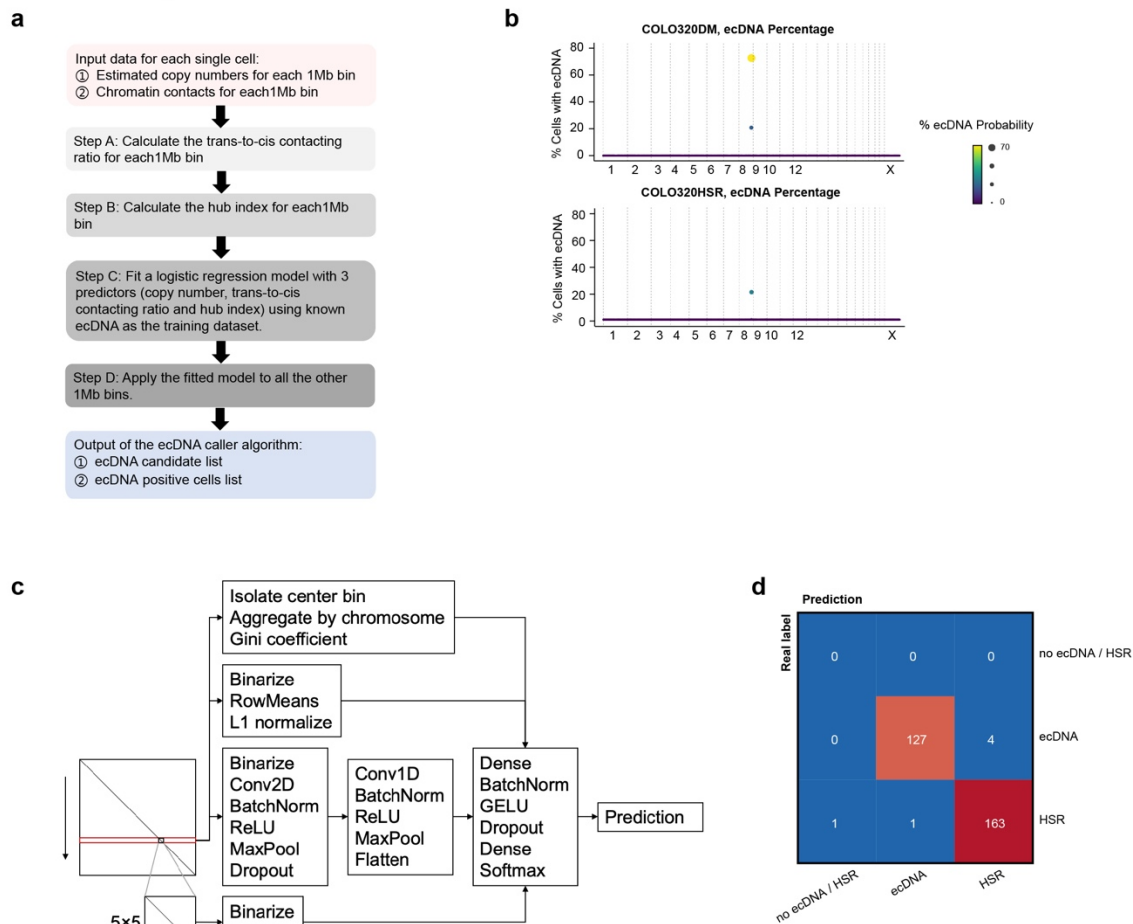
a, Schematic of the Droplet Hi-C annotation strategy using single-cell gene associating domain (scGAD) score for co-embedding. **b**, UMAP visualization of reference mouse cortex single-nucleus RNA-seq data used for co-embedding. **c**, Distribution of cell type prediction scores for Droplet Hi-C cells co-embedded with reference snRNA-seq data. **d**, Distribution of single-cell prediction scores across different annotated subclasses. **e**, Volcano plot showing differential scGAD score of genes between neuronal subtypes PVGA and ITL6GL. Genes with adjusted P value < 0.01 and fold change > 2 are colored in red. P values were calculated by Wilcoxon rank-sum test. Benjamini–Hochberg FDRs were then calculated to select significant genes. **f**, Comparison of chromatin contacts between PVGA and ITL6GL at *ErbB4* (chr1: 66.25-70.65Mb) locus. Compartment scores from the corresponding subclass are shown below the contact maps. Color bar showed the imputed contact number. **g**, Volcano plot showing the differential scGAD of genes between inhibitory neuronal subtypes PVGA and VIPGA, with significant genes colored in red. **h**, Comparison of imputed contact maps between PVGA and VIPGA around *Sox6os* gene region (chr7: 114-118Mb) at 25-kb resolution. Compartment scores from the corresponding subclass are also shown. Color bar showed the imputed contact number. **i**, Heatmap of aggregated scGAD score at cell type-specific markers across all subclasses from Droplet Hi-C data, along with single-cell gene expression heatmap of the same marker genes from snRNA-seq. **j**, Box plots showing Spearman correlation coefficients (SCC) of compartment score between matched versus unmatched subclasses in Droplet Hi-C and sn-m3C-seq. P value was calculated by Wilcoxon rank-sum test. **k**, Heatmap showing overlap score of subclass annotations between Droplet Hi-C and sn-m3C-seq. Overlap score is calculated as the overlap between the original annotations and the joint cluster.





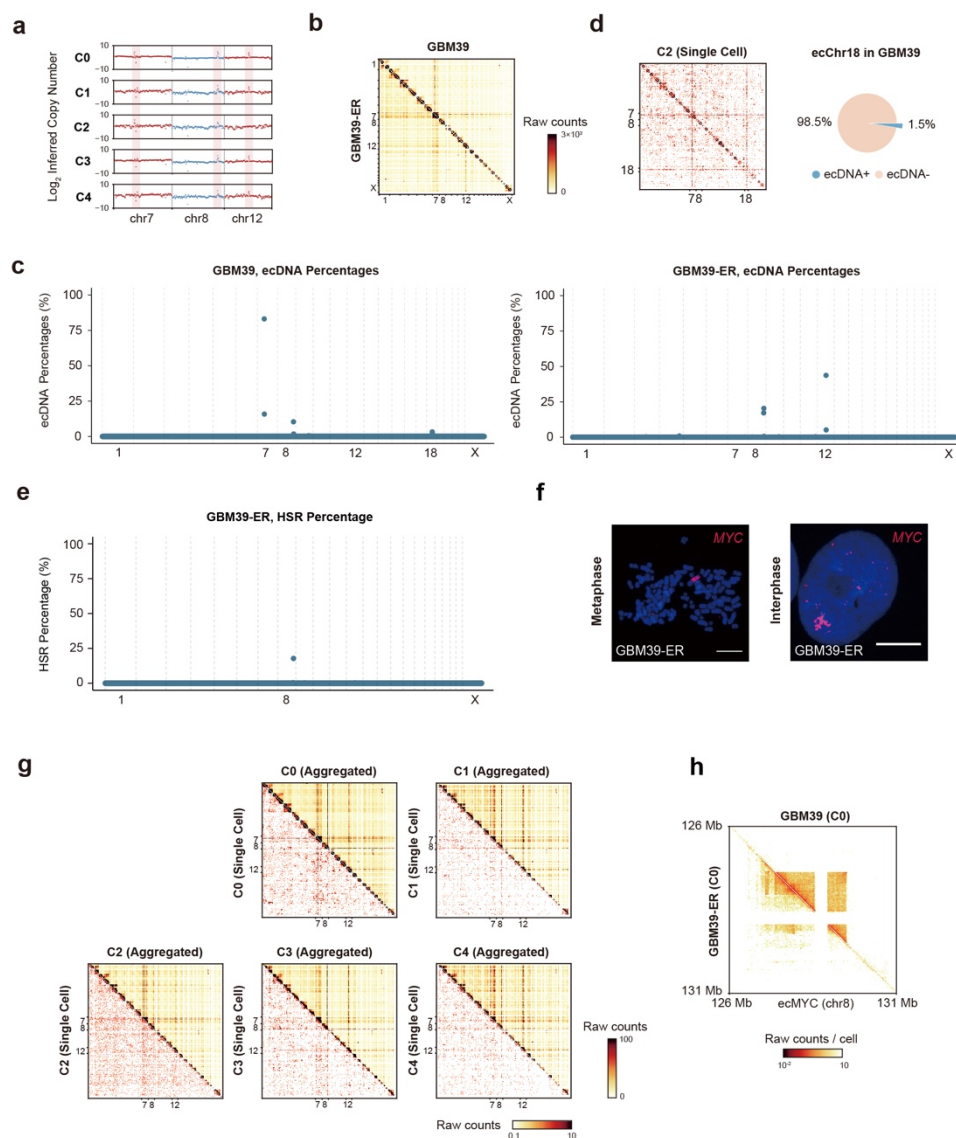
Extended Data Figure 2.4: Droplet Hi-C illuminates CNV, SVs and ecDNA in cancer cells.

a, Contact maps of regions containing ecMYC in COLO320DM and COLO320HSR. Color bar showed the raw contact number. Inferred copy number from pseudo-bulk profiles as well as representative single cells are shown below. **b**, Single-cell hub index calculated with normal versus shuffle contact profiles in COLO320DM and COLO320HSR. *P* values are calculated from Wilcoxon signed-rank test.



Extended Data Figure 2.5: Flowchart of ecDNA caller algorithms.

a, Workflow of multivariate logistic regression model-based ecDNA caller used to predict ecDNA identity for each genomic region, and cells containing the candidate ecDNA. **b**, ecDNA prediction results from the logistic regression model-based ecDNA caller in COLO320DM and COLO320HSR sample. **c**, Schematic of deep learning-based ecDNA caller used to predict HSR/ecDNA identity of genomic regions, and cells containing the candidate HSR/ecDNA. **d**, Confusion matrix of prediction results versus original label from deep learning-based ecDNA caller on COLO320DM and COLO320HSR validation datasets.

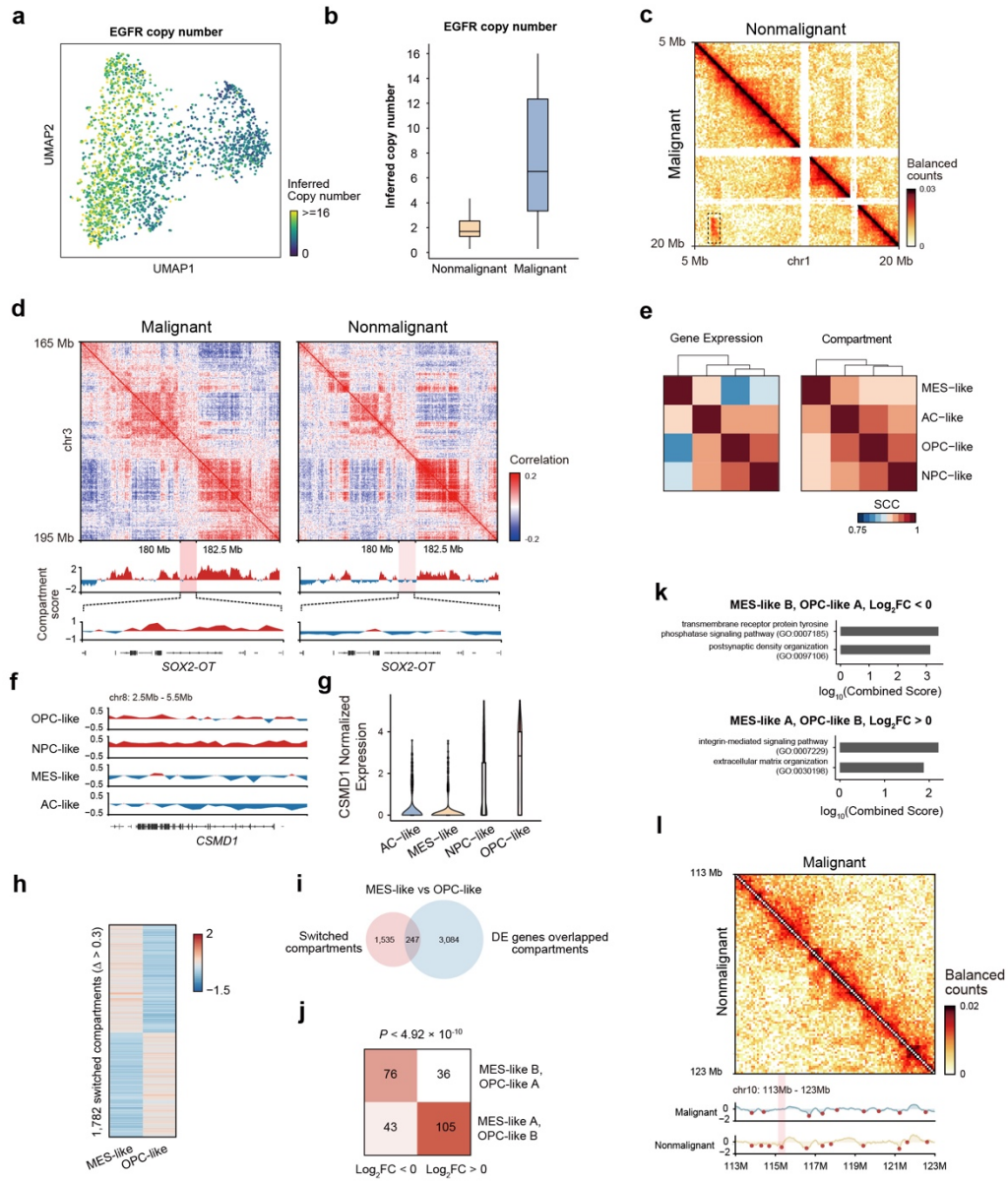


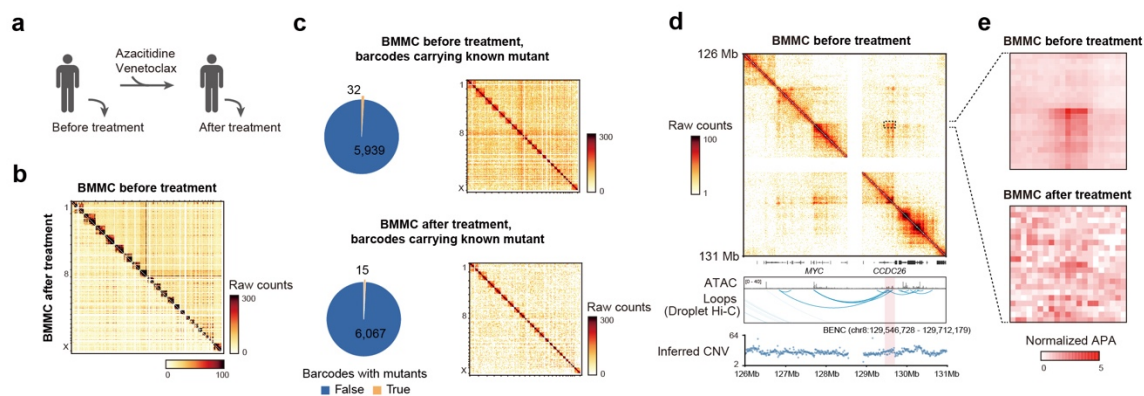
Extended Data Figure 2.6: Dynamics of ecDNA in GBM39 under erlotinib treatment.

a, Scatter plot showing inferred copy number on chromosomes containing candidate ecDNA (chromosome 7, 8, 12) across clusters in GBM39 and GBM39-ER. EcDNA-associated regions are highlighted in pink. **b**, Genome-wide contact maps for GBM39 and GBM39-ER cells. **c**, Scatter plot showing genome-wide ecDNA prediction results from deep learning-based ecDNA caller for GBM39 and GBM39-ER. Color bar showed the raw contact number. **d**, Representative single-cell contact map and percentage of cells containing a rare ecDNA (ecChr18) found by deep learning-based ecDNA caller. **e**, Scatter plot showing HSR prediction result in GBM39-ER with the deep-learning based ecDNA caller. The 1 Mb genomic bin overlap with *MYC* on chr8 is predicted to be HSR. **f**, *MYC* HSR is validated by DNA FISH in GBM39-ER metaphase and interphase cells. **g**, Genome-wide contact maps at cluster and representative single-cell levels. **h**, Comparison of ecMYC interaction profiles for ecDNA positive cells within C0 from different samples.

Extended Data Figure 2.7: Cell state-specific chromatin structure and transcriptome differences in GBM patient samples.

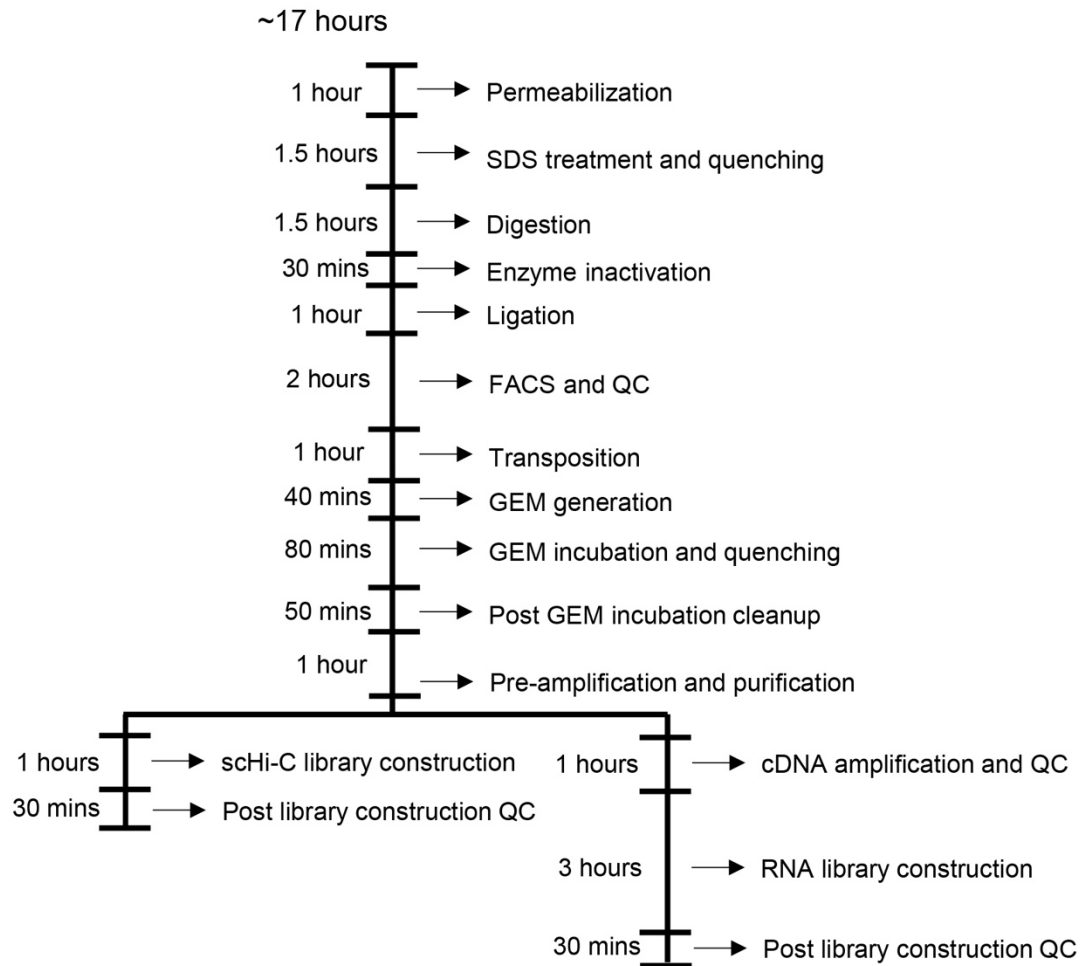
a, UMAP visualization of inferred copy number of 1Mb genomic bin (chr7: 55 - 56 Mb) harboring EGFR. **b**, Box plots showing distribution of single cell inferred copy number of 1Mb genomic bin harboring EGFR in malignant or nonmalignant populations. **c**, Example of malignant population-specific SV on chromosome 1 from GBM patient sample. SV are highlighted with black rectangle. Color bar shows the balanced contact number. **d**, Comparison of A/B compartment profiles on chromosome 3 between malignant cells and nonmalignant populations. **e**, Heatmap showing Spearman correlation coefficients (SCC) of 10X Multiome gene expression profiles and Droplet Hi-C compartment scores across different cellular states. **f**, Compartment profile of *CSMD1* gene region (chr8: 2.5 - 5.5Mb) in different GBM cellular states. **g**, Violin plots and box plots showing normalized expression of *CSMD1* in different GBM cellular states. **h**, Heatmap showing switched compartments between MES-like and OPC-like populations. **i**, Venn diagram of overlap between switched compartments and compartments with differentially expressed genes between MES-like and OPC-like populations. **j**, Enrichment of differentially expressed genes in switched compartments from (i). *P* value is calculated by Fisher exact test. **k**, Gene Ontology (GO) analysis of differentially expressed genes overlapped with switched compartments in MES-like and OPC-like populations using Enrichr. Combined score is determined as the product of the Fisher exact test *P* value and the z-score of the deviation from the expected rank. **l**, Comparison of balanced contact maps and insulation profile in example region (chr10: 113 - 123Mb) between malignant and nonmalignant populations. Example region with differential insulation score is highlighted in pink. Color bar shows the balanced contact number.





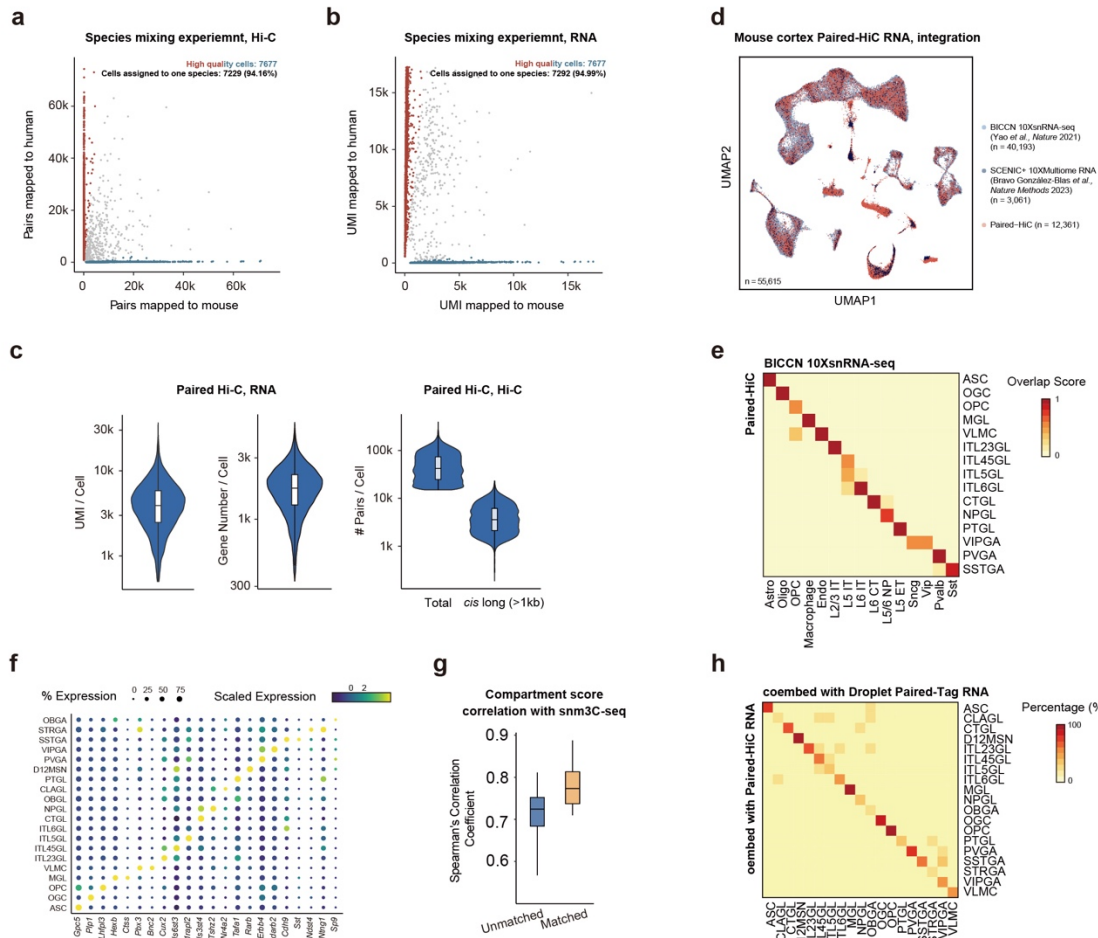
Extended Data Figure 2.8: Chromatin structure differences between before and after treatment in BMMC patient samples.

a, Illustration of AML and MDS patient treatment procedure. **b**, Pseudo-bulk genome-wide contact maps for BMMC samples from AML and MDS patient before and after treatment. **c**, Proportion of cells containing known tumor-associated mutations in BMMC samples before and after treatment. Aggregated genome-wide contact maps from mutant-carrying cells are also shown. Color bar shows the raw contact number. **d-e**, An example of a before-treated sample-specific loop between *MYC* and its enhancer region (BENC) on ecMYC. Normalized aggregate peak analysis (APA) scores of before- and after-treated samples are shown aside.



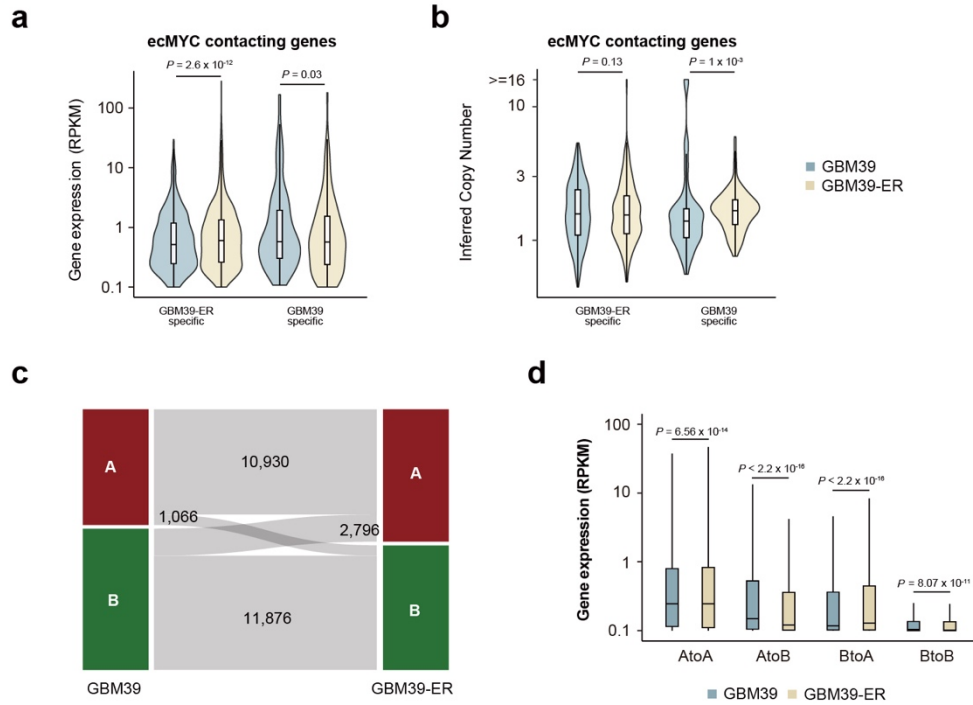
Extended Data Figure 2.9: Workflow of Paired Hi-C.

Detailed workflow describing the end-to-end procedure of Paired Hi-C and time cost for each step.



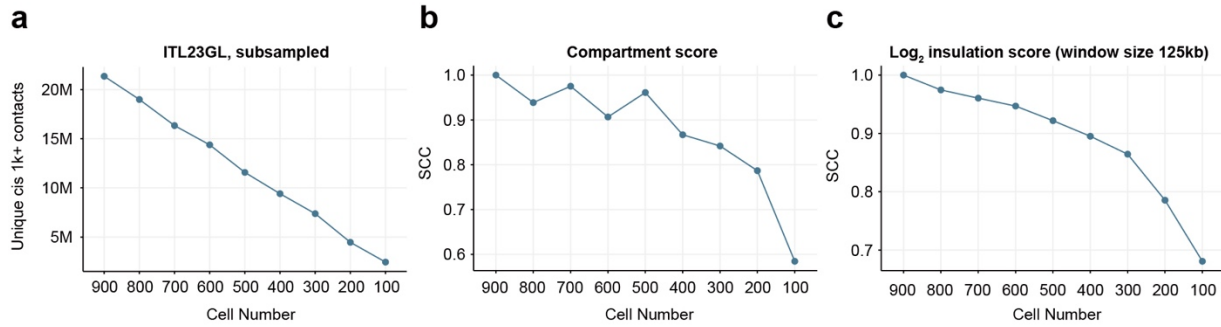
Extended Data Figure 2.10: Joint chromatin organization and gene expression analysis using Paired Hi-C.

a, Scatter plots showing proportion of human and mouse Hi-C read pairs in each cell from Paired Hi-C species mixing experiment. **b**, Scatter plots showing proportion of human and mouse RNA UMI in each cell from Paired Hi-C species mixing experiment. **c**, Distribution of UMI, gene number, and contacts per nucleus in Paired Hi-C data from adult mouse cortex. **d**, UMAP visualization of single-nucleus transcriptome profiles from Paired Hi-C experiment co-embedded with the reference BICCN and SCENIC+ datasets. **e**, The overlap scores of shared annotations between BICCN single-nucleus RNA-seq dataset and Paired Hi-C single-nucleus RNA-seq dataset. **f**, Dot plot showing expression level of marker genes in each cell type. **g**, Box plots showing Spearman correlation coefficients (SCC) of compartment score between matched versus unmatched subclasses in sn-m3C-seq and Paired Hi-C. **h**, Overlap of predicted annotations when using Droplet Paired-Tag snRNA-seq dataset as reference versus Paired Hi-C snRNA-seq dataset as reference when co-embed Droplet Hi-C data.



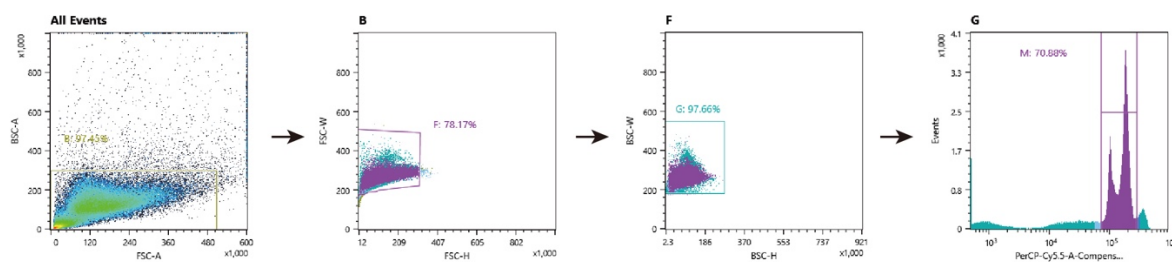
Extended Data Figure 2.11: Paired Hi-C illuminates alterations in the transcriptome associated with variations in chromatin structure.

a, Comparison of expression level for ecMYC *trans*-contacting genes between GBM39 and GBM39-ER cells. P values are calculated from Wilcoxon signed-rank test. **b**, Violin plots showing the inferred copy number of 500-kb genomic bins at ecMYC interaction hotspots, contrasts between GBM39 and GBM39-ER cells. P values are calculated from Wilcoxon signed-rank test. **c**, Summary of changes in A/B compartments in GBM39 and GBM39-ER sample. **d**, Expression of genes classified by the underlying A/B compartment transitions in GBM39 and GBM39-ER cells.



Extended Data Figure 2.12: Downsampling analysis of chromatin interactions in ITL23GL subclass.

a, Scatter plot showing the relationships between number of *cis* 1k+ contacts and subsampled cell number in ITL23GL subclass in adult mouse cortex Droplet Hi-C data. **b**, Scatter plot showing the Spearman correlation coefficients (SCC) of compartment score of subsampled pseudobulk Droplet Hi-C profiles with pseudobulk profiles from all cells. **c**, Scatter plot showing the SCC of insulation score of subsampled pseudobulk profiles with the profile from all cells.



Extended Data Figure 2.13: Nuclei gating strategy for 7'AAD stained nuclei for HeLa S3 and mESC mixing experiment.

2.8 Methods

Cell culture

HeLa S3 (human, ATCC CCL-2.2) cells were cultured according to standard procedures in high glucose DMEM (Sigma-Aldrich, D6429) supplemented with 10% fetal bovine serum (FBS; Omega Scientific, FB-02) and 1% Penicillin-Streptomycin-Glutamine (ThermoFisher Scientific, 10378016) at 37°C with 5% CO₂.

mESCs were maintained in feeder-free and serum-free 2i medium at 37 °C with 5% CO₂. To isolate nuclei, mESCs were dissociated with Accutase (Innovative Cell Technologies, AT104), collected by centrifugation.

K562 cells were cultured at 37°C, 5% CO₂ in RPMI-1640 (ATCC, 30-2001) supplemented with 1% Penicillin-Streptomycin-Glutamine and 10% FBS.

GM12878 cells were cultured at 37°C, 5% CO₂ in RPMI-1640 supplemented with 1% Penicillin-Streptomycin-Glutamine and 15% FBS.

WTC11 iPSCs containing an inducible NGN2 cassette inserted into the AAVS1 safe harbor locus were maintained in Essential 8 Flex medium (Thermo Fisher, A2858501) on culture dishes coated with Matrigel (Corning, 354277). Cells were cultured with 10 uM ROCK inhibitor (STEMCELL Technologies, Y-27632) for 24h after seeding, and fed daily with fresh E8 Flex medium

COLO320DM (ATCC, CCL-220) and COLO320HSR (ATCC, CCL-220.1) were maintained in RPMI-1640 supplemented with 10% FBS (ThermoFisher, S11550H) and 1%

penicillin-streptomycin (ThermoFisher, 15140122). Media was refreshed every 3-4 days, and cells were passaged once per week using Accutase (VWR, AT104) for dissociation.

As described previously, patient-derived xenograft (PDX) model GBM39 (Mayo Clinic Hospital) was cultured in DMEM/F12 (ThermoFisher, 11320033), 20ng/mL EGF (STEMCELL Technologies, 78006), 20ng/ml FGF (STEMCELL Technologies, 78134.1), 10% B27 Supplement without Vitamin A (ThermoFisher, 12587010), and 1% penicillin-streptomycin solution at 37°C with 5% CO₂. GBM39-ERL was under continuous treatment with 5μM erlotinib (LC Laboratories, E-4007). GBM39 cell lines were cultured as neurospheres.

Adherent cells were washed once with 1× PBS, trypsinized with 0.25% trypsin–EDTA (ThermoFisher Scientific, 25200056), spun down at 500× g for 5 min, resuspended in culture medium, and spun down again at 500 × g for 5 min. Suspension cells were spun down at 500× g for 5 min to be collected from the culture medium. The cell pellets were resuspended in 1 million cells/mL 1× PBS (pH=7.4) for crosslinking.

Nuclei preparation from the mouse cortex

All animal work described in this manuscript has been approved and conducted under the oversight of the Institutional Animal Care and Use Committee at the University of California, San Diego. Male C57BL/6J mice were purchased from the Jackson Laboratory (000664) at 7 weeks of age and were housed in the animal facility at University of California, San Diego, under a 12-h light/12-h dark cycle in a temperature-controlled room with *ad libitum* access to water and food until euthanasia and tissue collection at 8 weeks of age. The isocortex was dissected from 8-week-old male mice, snap-frozen in liquid nitrogen and stored at –80 °C before proceeding to nuclei extraction.

Single-nuclei suspensions were prepared from fresh tissues by dounce homogenization in douncing buffer (0.25 M sucrose (Sigma, S7903), 25 mM KCl (Invitrogen, AM9640G), 5 mM MgCl₂ (Invitrogen, AM9530G), 10 mM Tris-HCl (pH 7.5) (ThermoFisher Scientific, 15567027), 1 mM DTT (Sigma, D9779), 1× protease inhibitor (Roche, 5056489001), 0.5 U/μL RNaseOUT(Invitrogen, 10777019), 0.5 U/μL SUPERaseIn inhibitor (Invitrogen, AM2694) and 0.1% Triton-X100 (Sigma, 93443)). The nuclei suspension was then filtered through a 30-μm Cell-Tric filter (Sysmex) and centrifuged for 10 min at 300 x g at 4°C. Cell pellets was washed once with douncing buffer without Triton-X100, centrifuged again and resuspended in 1 million cells/mL 1× PBS (pH=7.4) for crosslinking.

Experimental protocol for Droplet Hi-C

A brief description of the Droplet Hi-C experimental procedure is provided below.

Fixation

Cells were crosslinked in 1% formaldehyde, which was diluted from 37% formaldehyde with 1× PBS (pH=7.4), and incubated at room temperature for 10 min. After crosslinking, the reactions were quenched in 200 mM glycine and incubated at room temperature for 5 min. Quenched reactions were spun down at 1,000 × g for 5 min at 4°C, resuspended using 1% BSA in 1× PBS (pH=7.4) to wash twice. One million cells were aliquoted into each tube. The cells were spun once again at 1,000 × g for 5 min, supernatant was removed, and the pellet was flash frozen in liquid nitrogen, and finally stored indefinitely at -80°C.

In situ Hi-C

Cell pellets were lysed with pre-cold 300 μ L lysis buffer (10 mM Tris-HCl, pH 8.0 (ThermoFisher Scientific, 15568025), 10 mM NaCl (Sigma, S5150), 0.2% Igepal CA630 (Sigma, I8896), 1 $\times$ protease inhibitor (Roche, 5056489001)) on ice for 45 min, then centrifuged at 1,000 $\times$ g for 5 min at 4°C to collect nuclei, and washed once with 200 μ L lysis buffer. The nuclei were resuspended with 50 μ L 0.5% SDS and incubated at 62°C for 10 min on Thermomixer. Then, we added 145 μ L of nuclease-free H₂O and 25 μ L 10% Triton X-100 (Sigma, 93443) to quench SDS and incubated samples at 37°C for 15 min on Thermomixer at 300 rpm. The samples were added 27 μ L of 10 $\times$ CutSmart Buffer and three restriction enzymes, including 50 U DpnII (NEB, R0543L), 62.5U MboII and 7.5U NlaIII (NEB, R0125L), and then incubated at 37°C for 90 min on Thermomixer at 550 rpm. The enzymes were deactivated at 65°C for 20 min and then cooled down to room temperature. The nuclei were collected by centrifugation at 1000 $\times$ g for 5 min at 4°C, washed once with 200 μ L ligation buffer (100 μ L 10 $\times$ T4 DNA ligase buffer (NEB, B0202S), 5 μ L 20 mg/mL bovine serum albumin (BSA; NEB, B9000S), 865 μ L H₂O) and ligation reaction was performed with 200 μ L ligation buffer and 20 μ L T4 DNA ligase (NEB, M0202L) at 37°C for 40 min on Thermomixer at 300 rpm.

Single-cell Hi-C sequencing Library construction

The ligated nuclei pellets were resuspended in 1 mL 1% BSA in PBS (diluted from 10% BSA in PBS (Sigma, A1595) using 1 $\times$ PBS (pH=7.4)) each tube, added 1 μ L 1,000 $\times$ 7-AAD (Invitrogen, A1310) and sorted by fluorescence-activated nuclei sorting with an SH800 cell sorter (Sony) for the isolation of single nuclei (Extended Data Fig. 2.13). Nuclei were collected in collection buffer (5% BSA in PBS) at 5°C and immediately centrifuged for 5 min at 1,000 $\times$ g and 4°C to collect nuclei. The nuclei were washed once with 1x Nuclei Buffer (diluted from 20 $\times$ Nuclei Buffer (10x Genomics, PN-2000207)), resuspended in 10 μ L 1x Nuclei Buffer, and counted

on an RWD C100-Pro fluorescence cell counter with DAPI staining. According to the targeted nuclei recovery number, 3,000-15,000 nuclei were aliquoted into PCR tubes and added Transposition Mix following user guide of Chromium Next GEM Single Cell ATAC Reagent Kits v1.1 or Chromium Next GEM Single Cell ATAC Reagent Kits v2 (10x Genomics). The incubation time for tagmentation is 60 min for both v1.1 kit and v2 kit. We also extended index PCR elongation time from 20s to 1 min. The double-sided size selection was changed to 1.14x SPRIselect to only remove small fragments.

All sequencing experiments were performed with an Illumina NextSeq 2000 or NovaSeq 6000 sequencer.

Experimental protocol for Paired Hi-C

A brief description of the Paired Hi-C experimental procedure is provided below.

Fixation

Cells were crosslinked in 0.6% formaldehyde, which was diluted from 37% formaldehyde with 1× PBS (pH=7.4), and incubated at room temperature for 10 min. After crosslinking, the reactions were quenched in 200 mM glycine and incubated at room temperature for 5 min. Quenched reactions were spun down at 1,000g for 5 min at 4°C, resuspended using 1% BSA in 1× PBS (pH=7.4) to wash twice, and then 1 million cells were aliquoted into each tube. The cells were spun once again at 1,000 × g for 5 min, supernatant was removed, and the pellet was flash frozen in liquid nitrogen, and finally stored indefinitely at -80°C.

In situ Hi-C

Cell pellets were lysed with pre-cold 300 μ L lysis buffer (10 mM Tris-HCl, pH 8.0 (ThermoFisher Scientific, 15568025), 10 mM NaCl (Sigma, S5150), 0.2% Igepal CA630 ((Sigma, I8896), 1 $\times$ protease inhibitor (Roche, 5056489001), 1 U/ μ L RNaseOUT (Invitrogen, 10777-019), 1 U/ μ L SUPERaseIn inhibitor (Invitrogen, AM2694)) on ice for 45 min, then centrifuged at 1,000 $\times$ g for 5 min at 4°C to collect nuclei, and washed once with 200 μ L lysis buffer. The nuclei were resuspended with 50 μ L 0.5% SDS, which was diluted from 10% SDS using 1 $\times$ PBS (pH=7.4) with 1 U/ μ L RNaseOUT and 1 U/ μ L SUPERaseIn inhibitor, and incubated at 37°C for 60 min on Thermomixer at 300rpm. Then, we added 197 μ L quenching buffer (137.59 μ L of nuclease-free H₂O, 25 μ L 10% Triton X-100, 27 μ L 10 $\times$ CutSmart, 1 U/ μ L RNaseOUT and 1 U/ μ L SUPERaseIn inhibitor) and incubated samples at 37°C for 15 min on Thermomixer at 300 rpm. The samples were centrifuged at 1,000 $\times$ g for 5 min at room temperature, supernatant was removed, and the pellets were resuspended with 250 μ L digestion buffer (1 $\times$ CutSmart Buffer, 1 U/ μ L RNaseOUT, 1 U/ μ L SUPERaseIn inhibitor, 50 U DpnII, 62.5U MboII and 5U NlaIII), and then incubated at 37°C for 90 min on Thermomixer at 550 rpm. The enzymes were deactivated at 65°C for 20 min and then cold to room temperature. The nuclei were collected by centrifugation at 1000 $\times$ g for 5 min at 4°C, washed once with 200 μ L ligation buffer (827.5 μ L H₂O, 100 μ L 10 $\times$ T4 DNA ligase buffer, 5 μ L 20 mg/mL BSA, 25 μ L SUPERaseIn inhibitor and 12.5 μ L RNaseOUT), and ligation reaction was performed with 200 μ L ligation buffer and 20 μ L T4 DNA ligase at 37°C for 40 min on Thermomixer at 300 rpm.

PBMCs (Allcells, LP, CR, MNC, 100M) were crosslinked in 1% formaldehyde. Then, 1 million cell pellets were lysed with 200 μ L of high-salt lysis buffer 1 (50 mM HEPES pH 7.4, 2 mM EDTA pH 8.0, 140 mM NaCl, 0.25% Triton X-100, 0.5% IGEPAL CA-630, 10% glycerol and 1 $\times$ proteinase inhibitor cocktail), and incubated on ice for 10 min. After this, cells were

pelleted at 500g for 2 min at 4 °C and then resuspended in 200 µL of high-salt lysis buffer 2 (10 mM Tris–HCl pH 8, 3 mM EDTA, 200 mM NaCl, 1× proteinase inhibitor cocktail). The solution was incubated on ice for 10 min. Following this, cells were then pelleted at 500g for 2 min at 4 °C and then resuspended in 200 µL of 1× T4 DNA ligase buffer containing 0.2% SDS. Cells were then incubated at room temperature for 1 hr. To quench the reaction, 200 µL of 1× T4 DNA ligase buffer and 10 µL 10% Triton X-100 were added to the tube. Finally, cells were spun at 500g for 4 min at 22 °C. The pellets were resuspended with 250 µL digestion buffer (1× T4 DNA ligase buffer, 1 U/µL RNaseOUT, 1 U/µL SUPERaseIn inhibitor, 80 U DpnII), and then incubated at room temperature for 2 hrs. The nuclei were collected by centrifugation at 500g for 4 min at 4 °C, washed once with 200 µL ligation buffer (857.5 µL H₂O, 100 µL 10× T4 DNA ligase buffer, 5 µL 20 mg/mL BSA, 25 µL SUPERaseIn inhibitor and 12.5 µL RNaseOUT), and ligation reaction was performed with 200 µL ligation buffer and 20 µL T4 DNA ligase at room temperature for 40 min.

Single cell sequencing library construction

The ligated nuclei pellets were resuspended in 1 mL 1% BSA in PBS (diluted from 10% BSA in PBS using 1× PBS (pH=7.4)) with 1 U/µL RNaseOUT and 1 U/µL SUPERaseIn inhibitor each tube, added 1 µL 1,000× 7-AAD (Invitrogen, A1310) and sorted by fluorescence-activated nuclei sorting with an SH800 cell sorter (Sony) for the isolation of single nuclei. Nuclei were collected in collection buffer (5% BSA in PBS with 5 U/µL Recombinant RNasin (Promega, PAN2515)) at 5°C and immediately centrifuged for 5 min at 1,000 × g and 4 °C to collect nuclei. The nuclei were washed once with 1x Nuclei Buffer (diluted from 20× Nuclei Buffer (10x Genomics, PN-2000207) with 1 mM DTT and 1 U/µL Recombinant RNasin), resuspended in 10 µL 1x Nuclei Buffer, and counted on an RWD C100-Pro fluorescence cell counter with DAPI staining. According to the targeted nuclei recovery number, 3,000-15,000 nuclei were aliquoted

into PCR tubes and added Transposition Mix following user guide of Chromium Next GEM Single Cell Multiome ATAC + Gene Expression (10x Genomics). The tagmented nuclei were loaded onto a Chromium Next GEM Chip J for droplet generation with a Chromium X microfluidic system (10x Genomics). The GEMs were collected, reverse transcription and cell barcoding were performed using a thermal cycler, and the reaction was quenched following the user guide of Chromium Next GEM Single Cell Multiome ATAC + Gene Expression (10x Genomics). After GEM cleanup and pre-amplification, the SPRIselect products were eluted into 100 μ L buffer EB (Qiagen, 19086) instead of 160 μ L. For scHi-C library construction, we extended index PCR elongation time from 20s to 1 min, and the double sided size selection was changed to 1.55x SPRIselect to only remove small fragments. For scRNA seq library construction, we used 0.8x SPRIselect to purify cDNA twice instead of 0.6x SPRIselect to keep short cDNA fragments.

All sequencing experiments were performed with an Illumina NextSeq 2000 or NovaSeq 6000 sequencer.

Bulk *in situ* Hi-C of Hela S3 cells

Replicate 1

Hi-C libraries were generated as described with minor modifications⁵. Briefly, 1 million crosslinked Hela S3 cells were lysed with 300 μ L ice cold lysis buffer (10 mM Tris-HCl (pH 8.0), 10 mM NaCl, 0.2% IPEGAL CA-630 with 1x protease inhibitors) on ice for 15 minutes, washed once with 500 μ L lysis buffer and pelleted at 4°C 2,500g for 5 min. Cell pellets were resuspended in 50 μ L 0.5% SDS and incubated for 10 min at 62 °C followed by quenching with 160 μ L 1.56% Triton X-100 for 15 min at 37 °C. The chromatin was digested overnight at 37 °C with shaking (1,000 r.p.m.) with 100 U MboI in 25 μ L 10 $\times$ NEBuffer 2. Digested fragments were filled in with

biotin-14-dATP at 37 °C 900 RPM for 2 hours and subsequently ligated with T4 DNA ligase (NEB) for 4 h at 23 °C with 300 r.p.m. gentle rotation. After the ligation step, nuclei were digested by 50 µl proteinase K (20 mg ml⁻¹) for 30 min at 55 °C. The DNA was purified by ethanol precipitation and resuspended in 130 µl 10 mM Tris-HCl (pH 8.0) and sonicated by a Covaris M220 sonicator with the parameters: duty cycle, 10%; power, 50; cycles per burst, 200; time, 70 s. Biotin-labeled free DNA ligation junctions were enriched by 150 µl myOne T1 streptavidin Dynabeads. Once bound to the streptavidin beads, the biotin-labeled DNA was subjected to end repair & biotin removal from free ends, A-tailing followed by ligation of Illumina indexed Truseq adapters. The DNA is then released from the beads by a 10 minute incubation at 98 °C. Smaller DNA fragments were removed by a 0.8X SPRI bead purification. The DNA is then PCR amplified. Final libraries were purified by SPRI bead based dual size selection (0.55:1 and 0.75:1), quantified and submitted for pair-end sequencing.

Replicate 2

Bulk Hi-C was performed as previously described¹³ with a modification involving the use of the restriction enzymes MboI and HinfI.

Metaphase spread cell preparation

One million cells were seeded into a single well of a 6-well plate coated with Matrigel (Corning, 354277). Cells were incubated in 100 ng/mL of Karyomax (Gibco, 15212-012) in standard growth media for 10 hours. Cells were washed with PBS, lifted with Accutase (Innovative Cell Technologies, AT-104), and collected into a 15 ml conical tube. Cells were spun down in the centrifuge at 1000 rpm for 5 minutes. The supernatant was aspirated, and the cell pellet was dislodged by gently flicking the conical tube. Cells were incubated in 1 mL of warm 0.075 M KCl.

200 uL of cold Carnoy's Fixative (3:1 mixture of methanol and acetic acid) was added to the cell mixture, which was mixed by gentle inversion. Cells were spun down at 1000 rpm for 5 minutes, the supernatant was aspirated, and the pellet was resuspended in 1 mL of cold Carnoy's Fixative. The cell mixture was transferred to a 1.5 mL Eppendorf tube and re-spun at 1000 rpm for 5 minutes, and the supernatant was aspirated. Fixed mitotic cells were resuspended in 40 uL of Carnoy's fixative and were either used immediately for metaphase spreads or stored at -20C. For each metaphase spread, 4 uL of cells were added to glass coverslips in a dropwise manner.

Fluorescence *In Situ* Hybridization (FISH)

Slides containing fixed cells in interphase and metaphase or 16- μ m tissue slices were briefly equilibrated in 2X SSC, followed by dehydration in 70%, 85%, and 100% ethanol for 2 minutes each. FISH probes in hybridization buffer (Empire Genomics) were then added to the slide, which was covered with a coverslip. The sample was denatured on a hotplate at 75°C for 3 minutes for fixed cell samples and 7 minutes for tissue samples, followed by hybridization at 37°C overnight. After removing the coverslip, the sample was washed once with 0.4X SSC containing 0.3% IGEPAL and once with 2X SSC containing 0.1% IGEPAL, each for 2 minutes. Finally, the sample was mounted using ProLong™ Diamond Antifade Mountant with DAPI (Invitrogen, P36966) before imaging. Images were acquired using a Zeiss LSM880 Airyscan confocal microscope with a 63 $\times$ Plan-APOChromat NA 1.4 oil lens. Representative images were selected from the Z-stack with the best brightness.

The UCSD Cytogenetics Laboratory performed FISH experiments on GBM39 and GBM39-ER cells using a *MYC* dual color break-apart probe set. This probe set detects gene rearrangement at the *MYC* locus on chromosome 8 at band q24 (*MYC*: Abbott Molecular, Inc.).

Processing of Droplet Hi-C data

Custom scripts for demultiplexing, mapping and extracting single cell contacts from Droplet Hi-C data were publicly available at <https://github.com/Xieeeee/Droplet-Hi-C>. First, depending on indexing primers used, Droplet Hi-C fastq files were demultiplexed using Illumina bcl2fastq (v2.19.0.316) or 10X Genomics cellranger-atac mkfastq (v2.0.0). After demultiplexing, a custom script was used to extract cellular barcode sequence from Read2, and barcodes are aligned to the whitelist provided by 10x Genomics using Bowtie (v1.3.0)⁷⁸. Aligned barcode sequences were appended to the beginning of the read name to record cellular identity of each read. Next, sequencing adapters were detected and trimmed with Trim-Garole (v0.6.10)⁷⁹. Cleaned reads were then mapped to the human (hg38) or mouse (mm10) reference genome using BWA-MEM (v0.7.17)⁸⁰, with arguments ‘-SP5M’ specified. Finally, after mapping, valid contacts were parsed, sorted and deduplicated by Pairtools (v0.3.0)⁸¹, with barcodes information stored in a separated column in the pairs file. Contacts were balanced and stored in cool format using Cooler (v0.8.10) for visualization and downstream analysis⁸². High-quality cells were selected based on the number of total unique contacts per cell in each library.

Analysis of Droplet Hi-C data

We here described the general analysis strategy and workflows for our Droplet Hi-C data. Dataset-specific manipulations in the manuscript, if any, will be indicated.

Visualization of Hi-C contact matrices

Bulk, pseudo-bulk, or single cell contact matrices in cool format were visualized and plotted using Cooler along with Matplotlib (v3.5.1)⁸³.

Signal enrichment calculation

Density plots and of signal enrichment over CEMBA cCREs were generated using bamCoverage from deepTools (v3.5.3). Peaks that overlapped with ENCODE blacklist (v2) regions were removed during the calculation of enrichment⁸⁴.

Embedding and clustering of single-cell Droplet Hi-C dataset on cultured cells

We used Higashi to infer low-dimensional cell embeddings for all our Droplet Hi-C datasets without imputations, including human cell lines mix (HeLa / GM12878 / K562, Extended Data Fig. 2.1b), GM12878 / WTC-11, Extended Data Fig. 2.1i), GBM39 / GBM39-ER (Fig. 2.4b), and GBM patient sample (Fig. 2.5b). For visualization, the L2 norm of cell embeddings were projected to two-dimensional space with uniform manifold approximation and projection (UMAP). To identify cells with similar identity, we performed Leiden clustering using igraph (v0.9.9) and leidenalg (v0.8.8)^{85, 86}.

Contacts frequency by distance analysis

We used the `expected\_cis` module from cooltools (v0.5.1) to calculate the average contact frequency as a function of genomic separation⁸⁷. To plot the curve, we performed smoothing and aggregation across chromosomes regions.

Imputation of single-cell chromatin contacts on adult mouse cortex Droplet Hi-C dataset

For the adult mouse cortex Droplet Hi-C data, we used scHiCluster (v1.3.4) to perform contact matrix imputation for individual cell at three different resolutions: 100 kb (for cells

embedding and visualization), 25 kb (for domain boundaries calling) and 10 kb (for cells embedding and loops calling)⁸⁸. In brief, contacts from individual cell were first masked for ENCODE blacklist regions⁸⁴. scHiCluster then performed linear convolution and random walk with restart to impute the sparse single cell contact matrices for each chromosome. Considering storage efficiency, we output the imputed results for the whole chromosome at 100kb, for 10.05Mb window at 25kb, and for 5.05Mb window at 10kb resolution for each cell. For cancer cells or tumor samples, we did not perform imputation due to copy number variations and structural variations.

Embedding and annotation of single-cell Droplet Hi-C dataset on adult mouse cortex

For the imputed adult mouse cortex datasets, we adopt a previously described method to co-embed Hi-C data with gene expression data (snRNA-seq) generated from the same tissue and to assign cell-type identities⁴⁰. Briefly, single-cell gene associating domain (scGAD) score R_{ij} is calculated as the raw number of interactions at gene body region for gene i in cell j with the 10 kb imputed matrix. This method has been implemented in scHicluster as the ‘gene_score’ module. By this, the single-cell chromatin interaction profiles can then be represented as a cell-by-gene scGAD score matrix, which serves as the input for integration with snRNA-seq data.

To perform integration, we began by normalizing and selecting the top 2,000 variable genes in reference snRNA-seq data. These genes were subjected to dimension reduction by principal-component analysis (PCA) using Scanpy (v1.7.2), retaining 30 principal components⁸⁹. The derived PCA model was then applied to transform the scGAD score matrix. Subsequent integration was conducted with a tailored version of Seurat Integration using canonical correlation analysis (CCA)⁹⁰. For the visualization of co-embedded data, we calculated the nearest neighbor graph ($k = 25$) and UMAP embedding with Scanpy.

To annotate cell identities within the Droplet Hi-C dataset, we identified the 15 nearest snRNA-seq neighbors for each cell via PyNNDescent (v0.5.6) using Euclidean distances⁹¹. These distances were scaled and converted into standardized scores so that the total score is added up to 1, and neighbor with the minimal distance would have the highest score. The identity of each Droplet Hi-C cell was inferred by nominating the cell type label that garnered the highest standardized score among its top nearest neighbors.

To validate the cell type annotations, co-embedding of Droplet Hi-C and sn-m3C-seq datasets was performed using scHiCluster, which facilitates the low-dimensional representation of chromatin contacts information. For each cell, imputed matrices at 100 kb resolution were binarized and flattened. The binarized data from individual cells across both datasets were then concatenated into a larger matrix. PCA was applied to this matrix on a per-chromosome basis. The PCA results for each chromosome were concatenated for a second round of PCA to generate the final cell embeddings. Batch effects were corrected using harmony⁹². After batch correction, Leiden clustering was employed to identify co-embedding clusters. The validity of original annotation was verified by calculating the overlap coefficients ($O_{i,j}$) between the clusters and the original annotations from Droplet Hi-C (A) and sn-m3C-seq dataset (B): $O_{i,j} = \min\left(\left[\frac{A_i \cap C_k}{A_i}\right], \max\left[\frac{B_j \cap C_k}{B_j}\right]\right)$, where i indicates query cell type from Droplet Hi-C, j indicates reference cell type from sn-m3C-seq, and k indicates the co-embedding cluster.

Analysis of A/B compartments

On sample levels, we used the ‘eigs_cis’ module from cooltools to perform eigen decomposition on balanced contact matrices and calculate the compartment score at 100 kb resolution⁸⁷. The orientation of the resulting eigenvectors was adjusted to correlate positively with

the CpG density of the corresponding 100 kb genomic bins, thereby determining the sign of the compartments (A/B). Similarity between samples is calculated as the Spearman correlation coefficient of compartment scores across all autosomes (Extended Data Fig. 2.1c, j). To illustrate chromatin interactions within and between compartments, we used the ‘saddle’ module from cooltools to calculate the average observed versus expected contact frequency, categorized by compartment scores.

To generate pseudo-bulk profiles for distinct cell types or clusters within a sample, we adopted a uniform approach outlined before to minimize bias and facilitate direct comparisons. Initially, the raw contact matrices at the sample level were normalized for distance effects. Next, Pearson correlation coefficients were computed on the distance-normalized matrices, and PCA was then performed on the correlation matrix. The model fitted at the sample level was utilized to transform the raw contact matrices of various cell subtypes or clusters. To compare similarity between cell types or clusters, we calculated the Spearman correlation coefficient of compartments score using compartment intersected with variable gene regions (Fig. 2.1e, Extended Data Fig. 2.7e). To identify differential compartments, we employed dchic (v2.1)⁹³, selecting only those with an adjusted P value < 0.01 and a Manhattan distance greater than the 2.5th percentile of the standard normal distribution for further downstream analysis.

For joint analysis of compartment and cell type-specific histone modifications, we utilized the recently published adult mouse cortex Droplet Paired-Tag dataset⁴¹. We retained only cell types that contain >100 cells, and were shared between datasets to ensure comparability across the analyses. For all differential compartments, we calculated the Pearson correlation coefficients between compartment score and signal of H3K27ac or H3K27me3 (CPM) within the same 100 kb genomic bins across various cell types.

Analysis of chromatin domains

On sample levels, we used the 'insulation' module from cooltools to compute insulation score for raw contact matrices at 25 kb resolution. To delineate cell type-specific variable domain boundaries in mouse cortex data, we employed TopDom (v0.0.2) on 25 kb resolution imputed contact matrices for each individual cell⁹⁴. We defined the 'boundary probability' for a given bin as the proportion of cells within a particular cell type that designate the bin as a domain boundary. The presence or absence of a domain boundary is summarized by n cell types into a $n \times 2$ contingency tables. We then computed the chi-square statistics and P value for each bin, and the domain boundaries were classified as variable if they exhibited a false discovery rate (FDR) < 0.001 and a boundary probability difference > 0.05.

For correlation analysis of gene expression surrounding variable domain boundaries, we first identified genes within a 100kb window centered on the variable domain boundaries. We then calculated Pearson correlation coefficients to quantify the relationship between boundary probabilities and expression levels (RPKM) of these genes from Droplet Paired-Tag data across shared cell types. The genes were further categorized into “housekeeping”, “constant”, and “variable”. In brief, the housekeeping gene list is taken from previous report⁹⁵. The top 2,500 variable genes from Droplet Paired-Tag RNA modality are selected as variable group. Genes with similar expression level as the top 2,500 variable genes but are not within the “housekeeping” or “variable” group is treated as “constant”.

Analysis of chromatin loops

For loops calling on sample levels, we adopted the 'dots' module in cooltools, which implements the principle of the commonly used HiCCUPS loop calling strategy, to call loops on the 10 kb raw contact matrices⁹⁶.

A modified version of SnapHiC workflow implemented in scHiCluster is used to perform loops calling on the 10-kb imputed contact matrices for mouse cortex cell types⁹⁷. To compare histone modification enrichment at loop anchors, we calculate H3K27ac or H3K27me3 CPM across all cell types at all loop anchors identified. When histone modification profiles and loop anchors are from the same cell types, the enrichment is classified as “matched”, otherwise classified as “unmatched”.

Gene Ontology annotation of loop anchors was performed using rGREAT (v1.26.0) in R⁹⁸. Gene Ontology biological process was selected for annotations. The result is used to generate plots in Fig. 2.2g.

Multi-way interactions analysis

Multi-way interactions are extracted with 'pairtools parse2' function in Pairtools, which is designed to rescue complex chromatin ligation events. After deduplication, only trans-interactions and cis-interactions with genomic distance >10 kb were retained. Read pairs overlapping with the ENCODE mm10 blacklist regions were removed. For each pair-end read on autosomes supporting multi-way contacts, we collected the 10-kb genomic bins containing their anchors. Finally, we defined one pair-end read to support multi-way contacts, if it contacts ≥ 3 unique 10-kb bins (can be both intra-chromosomal and inter-chromosomal).

To define contact hubs for multi-way interactions in each mouse brain cell type, we

followed the below strategy: First, the reference genome (mm10 for mouse cortex) was segmented into consecutive 10-kb bins. Next, a binary indicator is assigned to each 10-kb bin in each single cell, depending on whether or not it is involved in a multi-way contact. Then, for each 10-kb bin, we calculated the frequency of single cells involving in multi-way contacts across all cells belonging to the same cell type. We excluded the top 1% bins with the highest frequency or bins with sequence mappability score <0.8 when calculating the mean and standard deviation for frequency distribution, then converted the frequency distribution into Z-score. Finally, a 10-kb bin is defined as in multi-way contact hub if its Z-score >1.96 . Custom scripts to analyze multi-way interactions hub is available at: <https://github.com/HuMingLab/Multiway.hub>.

Enrichment of chromatin features at multi-way contact hub

After identifying the multi-way contact hub in each mouse cortex cell types, we performed enrichment analysis against candidate cis-regulatory elements (cCREs), super-enhancers and cell-type-specifically expressed genes. Specifically, cCREs and super-enhancer are calculated and converted into 10-kb genomic bins from snATAC-seq data⁹⁹. For each cell type A involving multi-way contact hub, we calculated its overlap with cCRE or super-enhancer from another cell type (B), and created a 2x2 contingency table. We then performed the Fisher's exact test to evaluate the significance of enrichment and reported the Log_2 odds ratio as enrichment score. We enumerated all pairs of cell types, and compared the Log_2 odds ratio in matched (AA) versus unmatched (AB , where $B \neq A$) cell types.

We repeated the same enrichment analysis for cell-type specific marker genes from published datasets⁴¹, except that we used the 10-kb bins overlap with marker genes transcription start site (TSS) for enrichment analysis.

Infer copy number variation (CNV) with Droplet Hi-C data

Copy number is inferred from Hi-C data using the ‘calculate-cnv’ module in NeoLoopFinder (v0.4.3)¹⁰⁰. For each single cell, the output residuals from the generalized additive model in ‘calculate-cnv’ was directly used to estimate copy number ratio. On bulk or pseudo-bulk level, an additional HMM segmentation is performed using the ‘segment-cnv’ module on the copy number ratio to determine the boundaries of copy number ratio segments. We assume all samples used in this manuscript for CNV calculation are diploid, therefore the inferred copy number is equal to $2 \times$ copy number ratio. Genome-wide CNV heatmap is plotted at 10 Mb resolution, where chromosome level CNV is plotted at 1 Mb resolution, and regional CNV profile is plotted at 10 kb resolution in R using pheatmap (v1.0.12). The inferred copy number is further used to assist identifying ecDNA candidates and the associated significant *trans*-interactions, and to correct the bias in contact matrices for structural variation identification.

Identify structural variation (SV) from Droplet Hi-C data

Structural variations (SV) were identified using the ‘predictSV’ module in EagleC (v0.1.9)¹⁰¹. In short, CNV effects on contact matrices at 5, 10, and 50 kb resolutions were removed using ‘correct-cnv’ module in NeoLoopFinder. ‘predictSV’ then uses a deep learning model to predict SV at each resolution on the corrected matrices, and combines all results to obtain a uniform, high-resolution SV list at 5 kb resolution. ‘annotate-gene-fusion’ module is applied on the final SV list to annotate gene fusion events.

Metrics to define chromatin interactions

We compiled a suite of metrics to assess the *cis*- and *trans*-interaction patterns at specific genomic loci. These metrics include: (1) contact evenness or “hub index”, quantified as the Gini coefficient for *trans*-interactions across all chromosomes excluding chromosome Y given an 1 Mb genomic bin, and was derived using `ineq` (v0.2-13) in R; (2) *trans-to-cis* contacting bin ratio (R_i), which compares the quantity of interacting bins within the same chromosome (N_C) to those across different chromosomes (N_T) for a given bin i : $R_i = \frac{N_T}{N_C}$. This ratio represents the *trans*-interaction tendency while it is not confounded by copy number variation; (3) copy number-adjusted *trans*-chromosomal interaction (*adj_nTIF*), which is also used to measure a genomic locus’s interaction activity, was calculated as described before⁴⁹.

Develop ecDNA caller for identifying ecDNA candidates

We derived two ecDNA callers to identify ecDNA candidates, one is based on the logistic regression model and the other one is based on the convolutional neural network (CNN). Both models predict genome-wide 1 Mb bins which contain ecDNA, as well as cells with presence of ecDNA population-wide. Detailed methodologies for both algorithms are delineated as follows.

Logistic regression model-based ecDNA caller

We trained a multivariate logistic regression model using the `glm` function in R with inferred copy number, hub index, and *trans-to-cis* contacting bin ratio as predictive variables. The training dataset comprised the well-defined MYC ecDNA in COLO320DM (chr8:127Mb-128Mb) and EGFR ecDNA in GBM39 (chr7:55Mb-56Mb) as positive data, with the same loci in COLO320HSR and GBM39-ER as negative control. We established a probability threshold for

classifying ecDNA presence. Specifically, in COLO320DM / COLO320HSR data, this threshold was determined to be 0.5. In GBM39 / GBM39-ER data the threshold was determined to be 0.95.

Deep learning-based ecDNA caller

Data preprocessing

To train the deep learning-based ecDNA caller, we also selected the well-defined MYC ecDNA in COLO320DM (chr8:127Mb-128Mb) and EGFR ecDNA in GBM39 (chr7:55Mb-56Mb) as positive data, with the same loci in COLO320HSR and GBM39-ER as negative control. We used all autosomes and chromosome X to create Hi-C contact matrix at 1 Mb resolution (3,044×3,044) for each cell. We then randomly selected 90% of cells as the training data and kept the remaining 10% of cells as the validation data. For each 1Mb bin, we retained its local 5 Mb neighborhood region (including the center bin itself), and used both *cis*- and *trans*- contacts (i.e., a 5×3,044 matrix) as its feature in the neural network model.

Model architecture

Our proposed model consists of two sequentially placed convolutional modules that extract features from the binarized Hi-C contact matrices. Each convolutional module consists of a multi-channel convolutional layer (8 channels in the first module, 16 channels in the second module), a batch normalization layer, a rectified linear unit (ReLU) activation function¹⁰², and a max-pooling¹⁰³ layer sequentially. Convolution kernels scan along the direction of rows in each layer. Large convolutional kernel sizes (5×45 in the first module and 1×45 in the second layer) are set to improve pattern capture because of sparsity. Max-pooling reduces the size of the matrix to half on

its second dimension to keep the most important features and thus improves learning efficiency and propagation speed. Strides and paddings of each layer were designed to balance computational efficiency and information retention and must be compatible with the matrix shapes and max-pooling layers. To enhance robustness and prevent overfitting, a dropout layer¹⁰⁴ with probability of 0.5 is inserted between convolutional modules.

Subsequently, a two-layer fully connected (dense) network integrates information from multiple sources and makes ternary predictions. The first fully connected layer with a 218 hidden size receives (A) the flattened output of the convolutional modules, (B) flattened 5×5 small contact matrix of the corresponding 5 Mb region (whose diagonal entries denotes intra-bin contacts) from binarized 5×3044 Hi-C contact matrices, (C) the hub index (used to measure inequality and heterogeneity of interactions with different chromosomes for certain bin and calculated by Python package “pygini”) of the center bin calculated from interaction counts aggregated per chromosome, and (D) L1 normalized row means of binarized 5×3044 Hi-C contact matrices. The output passes through a batch normalization layer, a Gaussian error linear unit (GELU) activation function¹⁰⁵, and a dropout layer with probability of 0.5. Finally, the second fully connected layer with hidden size 64 produces predictions with a subsequent softmax activation function, in which each prediction contains three probabilities of “None”, “ecDNA”, “HSR” sequentially.

Model training and validation

With the package PyTorch¹⁰⁶, the training process spanned 40 epochs using the mini-batch training strategy (batch size=32). To ensure robust optimization, we applied the AdamW^{107, 108} optimization algorithm with 0.001 weight decay and 0.001 learning rate, and implemented the “hard” bootstrapped cross-entropy loss with parameter β equals to 0.99³⁰, which is calculated from

one-hot-encoded class labels and predictive softmax¹⁰⁹ probabilities. To efficiently reduce false positives, we compensated the difference in quantities of different labels and in addition forced the model to favor negative prediction (“None”) using biased class weights, which is incorporated into the loss calculation. Thus, the model receives a greater loss as a penalty when generating any positive prediction (“ecDNA” or “HSR”). We mainly utilized confusion matrix and subsequently derived binary and ternary precision, recall, specificity, and accuracy as our evaluation metrics, and binary results (“None” or “ecDNA” only) from logistic regression as the baseline performance.

Prediction of ecDNA

The model scans each Hi-C contact matrix from beginning to end without strides to generate SoftMax probabilities on each 1 Mb bin (except the first two bins of chromosome 1 and the last two bins of chromosome X as no padding to the Hi-C matrix is applied) on each single cell. Argmax (the arguments of the maxima) function is used to determine the final predicted class, which is either 0 (“None”), 1 (“ecDNA”), or 2 (“HSR”). Subsequently, results are aggregated to calculate the proportion of cells with ecDNA and/or HSR among the entire cell population.

Identify significant *trans*-interactions of ecDNA candidate loci

To identify genomic regions preferentially interact with ecDNA, we first quantify the number of interactions (N_i) of 500 kb genomic interval i on different chromosomes with ecDNA candidate loci, treating each chromosome separately. The observed interaction frequency (P_i) is the proportion of N_i relative to all contacts on the same chromosome: $P_i = \frac{N_i}{\sum_{g=1}^n N_g}$. The expected interaction frequency for interval i was calculated as the ratio of the inferred copy number (CN_i) to the total copy number on the same chromosome: $E_i = \frac{CN_i}{\sum_{g=1}^n CN_g}$, based on the null hypothesis

that interaction frequency for genomic regions with ecDNA are only weighted by their underlying copy number. To identify significant interactions, observed-versus-expected P value was calculated based binomial distribution model. Multiple testing correction is done by Bonferroni adjustment. Regions with adjusted P value < 0.05 are selected as significant interacting regions.

ecDNA hub analysis

To generate a background distribution for ecDNA hub analysis, we shuffled chromosome identities for all *trans* interactions in each cell, and calculated the hub index based on the shuffled contact matrices. P value is calculated from Wilcoxon signed-rank test.

Analysis of 10X Multiome datasets

10X Multiome fastq files for the GBM patient sample were demultiplexed and pre-processed with cellranger-arc (v2.0.0). For human PBMC 10X Multiome data, the processed and filtered matrices was downloaded directly from 10X data portal. Clustering of single-nucleus transcriptomic or chromatin accessibility data were performed in R using Seurat (v4.1.0) or Signac (v1.6.0)^{110, 111}. For transcriptomic data, gene counts were normalized and scaled, and the top 2,500 variable genes were selected for dimension reduction by PCA. The first 30 principal components were used for UMAP visualization and Louvain clustering. Potential doublets were identified and removed using Scrublet¹¹². For chromatin accessibility data, cell-by-peak matrices were normalized by the two-step term frequency-inverse document frequency (TF-IDF). The top 95% of genomic bins were selected for linear dimension reduction, again followed by UMAP

visualization and Louvain clustering. Gene activity scores were computed by ATAC signal density in promoter and gene body regions.

Analysis of GBM cellular states

To classify single cells within GBM 10x Multiome or Paired Hi-C RNA datasets into one of four predefined malignant cellular states, we adopted a previously described two-dimensional visualization technique⁵⁵. Briefly, we quantified the gene enrichment score ($SC_{j(i)}$) for each cell (i) against one of the four gene sets (G_j) associated with the particular cellular state. This score was calculated as the relative averaged expression (Exp) of G_j in cell i compared to a group of genes (G_{jcont}) with a similar level of expression as control: $SC_{j(i)} = \frac{\sum_{g=1}^{N_j} Exp_{g(i)}}{N_j} - \frac{\sum_{g=1}^{N_{jcont}} Exp_{g(i)}}{N_{jcont}}$, where N_j and N_{jcont} are the number of genes in G_j and G_{jcont} . After obtaining scores for all four cellular states, the cells were stratified into OPC/NPC versus AC/MES categories using the differential score $D = \max(SC_{OPC}, SC_{NPC}) - \max(SC_{AC}, SC_{MES})$. For further refinement, OPC/NPC cells were assigned an identity value $C = \log_2(|SC_{OPC} - SC_{NPC}| + 1)$, and AC/MES cells were similarly categorized with $C = \log_2(|SC_{AC} - SC_{MES}| + 1)$. The distribution of cellular states was then plotted in the two-dimensional representation with D on the y-axis and C on the x-axis.

For cellular state identification in Droplet Hi-C data, co-embedding with the reference snRNA-seq was performed using the scGAD score. After determining the top 15 nearest snRNA-seq neighbors and their scaled distance-based similarity score (D_m) for each Droplet Hi-C cell, the Hi-C gene enrichment score ($HSC_{j(i)}$) was computed as the sum of the neighbor-weighted enrichment scores: $HSC_{j(i)} = \sum_{g=1}^{15} SC_{j(g)} \times D_{m(g)}$.

Gene Ontology (GO) enrichment

We perform Gene Ontology enrichment analysis for DE genes overlap with switched compartments in GBM cellular states using Enrichr¹¹³ in R. The gene set ‘GO_Biological_Process_2018’ was used with default parameters. Combined score is defined as the Fisher’s exact test P value multiplied with the Z-score of the deviation from the expected rank. The results are used to generate plot in Extended Data Fig 7k.

Single-cell genotyping by Droplet Hi-C data

Frequently observed mutations in AML were sequenced using Hematologic Malignancy Comprehensive Panel for the patient BMMC sample. To detect malignant cells carrying the detected mutations from Droplet Hi-C data, cellular barcode information was appended to the mapped bam files as an extra tag. To ensure comparability between pre- and post-treated samples, bam files were subsampled to a uniform sequencing depth. we employed a previously reported custom script to count reads for both the wild-type and mutant alleles, considering only barcodes that met predefined quality standards¹¹⁴. For each cell and mutational site, we then summarized the detected mutant and wild-type reads. Given the limited sequencing coverage, we defined potential malignant cells as cell carrying at least one mutant read.

Processing and Analysis of Paired Hi-C data

Preprocessing and analysis of Hi-C modality in Paired Hi-C are identical as described for Droplet Hi-C. For RNA modality, fastq were demultiplexed by cellranger-arc, but preprocessed using cellranger (v6.1.2). After obtaining the cell-by-gene matrix, clustering and visualization were performed as described for 10x Multiome RNA dataset. Since barcodes in the same gel bead for RNA and Hi-C modalities are different, we performed manual pairing to match cell barcodes based on the 10X Multiome barcodes whitelist provided in cellranger-arc (10x Genomics).

Integration of the Paired Hi-C RNA dataset with reference datasets was performed by Seurat. First, normalization of gene counts was performed, and the top 2,000 shared variable genes across datasets were selected as integration features. Subsequent canonical correlation analysis allowed projecting all nuclei into a unified embedding space. Anchors (pairs of corresponding cells from distinct datasets) were then discerned through mutual nearest neighbors searching. Anchors with low confidence were excluded, and the shared neighbor between anchors and query cells are computed. Louvain clustering was applied to the shared neighbor graph to discern co-embedded clusters. We calculated overlap coefficients as described in previous section to compare clustering results from different datasets.

Analysis of gene expression levels and copy numbers of ecDNA

To calculate the correlation between gene expression level and inferred copy number of genes on ecDNA, we first refined the range of ecDNA at 10 kb resolution. This is based on the observation of the increased intra-ecDNA interaction frequency than interactions with regions on linear genome, irrespective of their genomic distance. Specifically, we enumerated the local interactions within each 10 kb segment of the 1 Mb ecDNA candidates. Subsequently, the contact numbers were smoothed across adjacent bins. By calculating the differential contact numbers between neighboring bins, we determined the changing points of interaction with an average difference cutoff across the entire 1 Mb region. The outermost local maxima and minima were designated as ecDNA boundaries.

With the ecDNA boundaries established, we categorized genes with gene bodies residing within ecDNA, as “ecDNA genes”. In the case of MYC ecDNA (ecMYC), we further classified genes on GBM39 ecMYC as “shared genes”, and genes on non-overlapping regions between

GBM39 and GBM39-ER as “ecMYC variable genes”, considering the greater ecDNA size in GBM39-ER at pseudo-bulk level. The Spearman correlation coefficient between gene expression levels (RPKM) and the average inferred copy number over all 10 kb segments at gene body across all cells was calculated, for a comprehensive representation of the interplay between gene expression and copy number.

Ethical approval

The GBM specimen collection was approved by the Institutional Review Board (IRB) at the University of Minnesota. The AML and MDS specimen collection was approved by the IRB at the University of California, San Diego. Each patient was consented by a dedicated research coordinator prior to collection. Samples were obtained with informed consent in accordance with the Declaration of Helsinki and appropriate Ethics Committee approval from each partner institution.

Data availability

Raw and processed sequencing data generated in this study have been submitted to GEO (accession number GSE253407, with reviewer token mpgtmwcvpgfwwj). The processed data can also be accessed as supplementary files in GEO. Datasets for bulk *in situ* Hi-C on cultured cells were downloaded from the 4DN data portal with the following accession number: HeLa S3 (4DNFIIT9PKS9 and 4DNFI2XGP1BW), K562 (4DNFIIX5BNC9 and 4DNFI244AS29), GM12878 (4DNFIIS73OJN and 4DNFI3O82QVV) and HeLa S3 at prometaphase

(4DNFIW458FJD). Other external datasets were downloaded from NCBI GEO with the following accession numbers: *in situ* Hi-C on WTC-11 (GSE106690), sci-Hi-C on cell lines mixture (GSE84920), Dip-C on adult mouse cortex (GSE162511), HiRes on mouse embryos (GSE223917), single cell Hi-C on mouse Th1 cells (GSE48262), single-nuclei Hi-C on K562 (GSE80006), Droplet Paired-Tag dataset on mouse frontal cortex (GSE152020), 10X Multiome dataset on mouse cortex (GSE210749), 10X Multiome dataset on COLO320DM and COLO320HSR (GSE160148). The BICCN whole mouse brain sn-m3C-seq datasets and BICCN MOp 10x snRNA-seq data were downloaded via the NeMO archive (<https://assets.nemoarchive.org/datsig83t9>; <https://assets.nemoarchive.org/datchlnqb7>). 10X Multiome dataset on PBMC in downloaded from the 10X Genomics dataset portal (<https://www.10xgenomics.com/en/datasets>). Source data are provided with this paper.

Code availability

Scripts and code are available at <https://github.com/Xieeeee/Droplet-Hi-C>. The ecDNA callers are available at <https://github.com/HuMingLab/ecDNACaller>. Scripts to analyze multi-way interactions hub is available at: <https://github.com/HuMingLab/Multiway.hub>

2.9 References

1. Dixon, J.R., Gorkin, D.U. & Ren, B. Chromatin domains: the unit of chromosome organization. *Molecular Cell* **62**, 668-680 (2016).
2. Rowley, M.J. & Corces, V.G. The three-dimensional genome: principles and roles of long-distance interactions. *Current Opinion in Cell Biology* **40**, 8-14 (2016).
3. Dekker, J. & Heard, E. Structural and functional diversity of topologically associating domains. *FEBS Letters* **589**, 2877-2884 (2015).
4. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nature Reviews Genetics* **17**, 661-678 (2016).
5. Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
6. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
7. Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381-385 (2012).
8. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).
9. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* **14**, 390-403 (2013).
10. Gibcus, J.H. & Dekker, J. The hierarchy of the 3D genome. *Molecular Cell* **49**, 773-782 (2013).
11. Du, Z., Zheng, H., Huang, B., Ma, R., Wu, J., Zhang, X., He, J., Xiang, Y., Wang, Q., Li, Y., Ma, J., Zhang, X., Zhang, K., Wang, Y., Zhang, M. Q., Gao, J., Dixon, J. R., Wang, X., Zeng, J., & Xie, W. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232-235 (2017).
12. Ke, Y., Xu, Y., Chen, X., Feng, S., Liu, Z., Sun, Y., Yao, X., Li, F., Zhu, W., Gao, L., Chen, H., Du, Z., Xie, W., Xu, X., Huang, X., & Liu, J. 3D chromatin structures of mature

- gametes and structural reprogramming during mammalian embryogenesis. *Cell* **170**, 367-381. e320 (2017).
13. Xu, J., Song, F., Lyu, H., Kobayashi, M., Zhang, B., Zhao, Z., Hou, Y., Wang, X., Luan, Y., Jia, B., Stasiak, L., Wong, J.H., Wang, Q., Jin, Qi, Jin, Qiushi, Fu, Y., Yang, H., Hardison, R.C., Dovat, S., Plataniias, L.C., Diao, Y., Yang, Y., Yamada, T., Viny, A.D., Levine, R.L., Claxton, D., Broach, James.R., Zheng, H., & Yue, F. Subtype-specific 3D genome alteration in acute myeloid leukaemia. *Nature* **611**, 387-398 (2022).
 14. Xu, Z., Lee, D. S., Chandran, S., Le, V. T., Bump, R., Yasis, J., Dallarda, S., Marcotte, S., Clock, B., Haghani, N., Cho, C. Y., Akdemir, K. C., Tyndale, S., Futreal, P. A., McVicker, G., Wahl, G. M., & Dixon, J. R. Structural variants drive context-dependent oncogene activation in cancer. *Nature* **612**, 564-572 (2022).
 15. Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenko, V. V., Ecker, J. R., Thomson, J. A., & Ren, B. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).
 16. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology* **20**, 535-550 (2019).
 17. Ghosh, R.P. & Meyer, B.J. Spatial Organization of Chromatin: Emergence of Chromatin Structure During Development. *Annual Review of Cell and Developmental Biology* **37**, 199-232 (2021).
 18. Gorkin, D.U., Leung, D. & Ren, B. The 3D genome in transcriptional regulation and pluripotency. *Cell stem cell* **14**, 762-775 (2014).
 19. Morgan, M.A. & Shilatifard, A. Chromatin signatures of cancer. *Genes & development* **29**, 238-249 (2015).
 20. Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., Gundem, G., Van Loo, P., Martincorena, I., Ganly, P., Mudie, L., McLaren, S., O'Meara, S., Raine, K., Jones, D.R., Teague, J.W., Butler, A.P., Greaves, M.F., Ganser, A., Döhner, K., Schlenk, R.F., Döhner, H., & Campbell, P.J. Genomic classification and prognosis in acute myeloid leukemia. *New England Journal of Medicine* **374**, 2209-2221 (2016).
 21. Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D. M., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., & Ding, L. et al. Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339 (2013).
 22. Blum, A., Wang, P. & Zenklusen, J.C. SnapShot: TCGA-Analyzed Tumors. *Cell* **173**, 530 (2018).

23. Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., & Fraser, P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).
24. Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., & Tanay, A. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61-67 (2017).
25. Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., Cramard, J., Faure, A.J., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M.G., Lehner, B., Di Croce, L., Wutz, A., Hendrich, B., Klenerman, D., & Laue, E.D. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59-64 (2017).
26. Flyamer, I. M., Gassler, J., Imakaev, M., Brandão, H. B., Ulianov, S. V., Abdennur, N., Razin, S. V., Mirny, L. A., & Tachibana-Konwalski, K. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110-114 (2017).
27. Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X.S. Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924-928 (2018).
28. Li, W., Lu, J., Lu, P., Gao, Y., Bai, Y., Chen, K., Su, X., Li, M., Liu, J., Chen, Y., Wen, L., & Tang, F. scNanoHi-C: a single-cell long-read concatemer sequencing method to reveal high-order chromatin structures within individual cells. *Nature Methods* **20**, 1493-1505 (2023).
29. Li, G., Liu, Y., Zhang, Y., Kubo, N., Yu, M., Fang, R., Kellis, M., & Ren, B. Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature Methods* **16**, 991-993 (2019).
30. 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. *Nature Methods* **16**, 999-1006 (2019).
31. Liu, Z., Chen, Y., Xia, Q., Liu, M., Xu, H., Chi, Y., Deng, Y., & Xing, D. Linking genome structures to functions by simultaneous single-cell Hi-C and RNA-seq. *Science* **380**, 1070-1076 (2023).
32. Ramani, V., Deng, X., Qiu, R., Gunderson, K. L., Steemers, F. J., Disteche, C. M., Noble, W. S., Duan, Z., & Shendure, J. Massively multiplex single-cell Hi-C. *Nature Methods* **14**, 263-266 (2017).
33. Zhou, T., Zhang, R., Jia, D., Doty, R. T., Munday, A. D., Gao, D., Xin, L., Abkowitz, J. L., Duan, Z., & Ma, J. Concurrent profiling of multiscale 3D genome organization and gene expression in single mammalian cells. *bioRxiv*, 2023.2007.2020.549578 (2023).

34. Zhang, R., Zhou, T. & Ma, J. Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nature Biotechnology* **40**, 254-261 (2022).
35. Abramo, K., Valton, A. L., Venev, S. V., Ozadam, H., Fox, A. N., & Dekker, J. A chromosome folding intermediate at the condensin-to-cohesin transition during telophase. *Nature Cell Biology* **21**, 1393-1402 (2019).
36. Salik, B., Yi, H., Hassan, N., Santiappillai, N., Vick, B., Connerty, P., Duly, A., Trahair, T., Woo, A.J., Beck, D., Liu, T., Spiekermann, K., Jeremias, I., Wang, J., Kavallaris, M., Haber, M., Norris, M.D., Liebermann, D.A., D'Andrea, R.J., Murriel, C., & Wang, J.Y. Targeting RSPO3-LGR4 Signaling for Leukemia Stem Cell Eradication in Acute Myeloid Leukemia. *Cancer Cell* **38**, 263-278.e266 (2020).
37. Tan, L., Ma, W., Wu, H., Zheng, Y., Xing, D., Chen, R., Li, X., Daley, N., Deisseroth, K., & Xie, X. S. Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development. *Cell* **184**, 741-758.e717 (2021).
38. 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. Single-cell DNA methylome and 3D multi-omic atlas of the adult mouse brain. *Nature* **624**, 366-377 (2023).
39. Zhang, C., Xu, Z., Yang, S., Sun, G., Jia, L., Zheng, Z., Gu, Q., Tao, W., Cheng, T., Li, C., & Cheng, H. tagHi-C reveals 3D chromatin architecture dynamics during mouse hematopoiesis. *Cell Reports* **32** (2020).
40. Shen, S., Zheng, Y. & Keleş, S. scGAD: single-cell gene associating domain scores for exploratory analysis of scHi-C data. *Bioinformatics* **38**, 3642-3644 (2022).
41. Xie, Y., Zhu, C., Wang, Z., Tastemel, M., Chang, L., Li, Y. E., & Ren, B. Droplet-based single-cell joint profiling of histone modifications and transcriptomes. *Nature Structural & Molecular Biology* **30**, 1428-1433 (2023).
42. Zhu, Q., Zhao, X., Zheng, K., Li, H., Huang, H., Zhang, Z., Mastracci, T., Wegner, M., Chen, Y., Sussel, L., & Qiu, M. Genetic evidence that Nkx2.2 and Pdgfra are major determinants of the timing of oligodendrocyte differentiation in the developing CNS. *Development* **141**, 548-555 (2014).
43. McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., & Bejerano, G. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**, 495-501 (2010).
44. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020).

45. Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardımcı, G.G., Chakraborty, A., Bann, D.V., Wang, Y., Clark, R., Zhang, L., Yang, H., Liu, T., Iyyanki, S., An, L., Pool, C., Sasaki, T., Rivera-Mulia, J.C., Ozadam, H., Lajoie, B.R., Kaul, R., Buckley, M., Lee, K., Diegel, M., Pezic, D., Ernst, C., Hadjur, S., Odom, D.T., Stamatoyannopoulos, J.A., Broach, J.R., Hardison, R.C., Ay, F., Noble, W.S., Dekker, J., Gilbert, D.M., & Yue, F. Integrative detection and analysis of structural variation in cancer genomes. *Nature Genetics* **50**, 1388-1398 (2018).
46. Harewood, L., Kishore, K., Eldridge, M. D., Wingett, S., Pearson, D., Schoenfelder, S., Collins, V. P., & Fraser, P. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biology* **18**, 125 (2017).
47. Turner, K.M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., Li, B., Arden, K., Ren, B., Nathanson, D.A., Kornblum, H.I., Taylor, M.D., Kaushal, S., Cavenee, W.K., Wechsler-Reya, R., Furnari, F.B., Vandenberg, S.R., Rao, P.N., Wahl, G.M., Bafna, V., & Mischel, P.S. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122-125 (2017).
48. Deshpande, V., Luebeck, J., Nguyen, N. D., Bakhtiari, M., Turner, K. M., Schwab, R., Carter, H., Mischel, P. S., & Bafna, V. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nature Communications* **10**, 392 (2019).
49. Zhu, Y., Gujar, A. D., Wong, C. H., Tjong, H., Ngan, C. Y., Gong, L., Chen, Y. A., Kim, H., Liu, J., Li, M., Mil-Homens, A., Maurya, R., Kuhlberg, C., Sun, F., Yi, E., deCarvalho, A. C., Ruan, Y., Verhaak, R. G. W., & Wei, C. L. Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer cell* **39**, 694-707. e697 (2021).
50. Zhu, Y., Gujar, A. D., Wong, C. H., Tjong, H., Ngan, C. Y., Gong, L., Chen, Y. A., Kim, H., Liu, J., Li, M., Mil-Homens, A., Maurya, R., Kuhlberg, C., Sun, F., Yi, E., deCarvalho, A. C., Ruan, Y., Verhaak, R. G. W., & Wei, C. L. Oncogenic extrachromosomal DNA functions as mobile enhancers to globally amplify chromosomal transcription. *Cancer Cell* **39**, 694-707.e697 (2021).
51. Liang, Z., Gilbreath, C., Liu, W., Wang, Y., Zhang, M. Q., Zhang, D. E., Wu, S., & Fu, X. D. Chromatin-associated RNA Dictates the ecDNA Interactome in the Nucleus. *bioRxiv*, 2023.2007.2027.550855 (2023).
52. Hung, K.L., Yost, K.E., Xie, L., Shi, Q., Helmsauer, K., Luebeck, J., Schöpflin, R., Lange, J.T., Chamorro González, R., Weiser, N.E., Chen, C., Valieva, M.E., Wong, I.T.-L., Wu, S., Dehkordi, S.R., Duffy, C.V., Kraft, K., Tang, J., Belk, J.A., Rose, J.C., Corces, M.R., Granja, J.M., Li, R., Rajkumar, U., Friedlein, J., Bagchi, A., Satpathy, A.T., Tjian, R., Mundlos, S., Bafna, V., Henssen, A.G., Mischel, P.S., Liu, Z., & Chang, H.Y. ecDNA hubs drive cooperative intermolecular oncogene expression. *Nature* **600**, 731-736 (2021).
53. Nathanson, D.A., Gini, B., Mottahedeh, J., Visnyei, K., Koga, T., Gomez, G., Eskin, A., Hwang, K., Wang, J., Masui, K., Paucar, A., Yang, H., Ohashi, M., Zhu, S., Wykosky, J.,

- Reed, R., Nelson, S.F., Cloughesy, T.F., James, C.D., Rao, P.N., Kornblum, H.I., Heath, J.R., Cavenee, W.K., Furnari, F.B., & Mischel, P.S. Targeted Therapy Resistance Mediated by Dynamic Regulation of Extrachromosomal Mutant EGFR DNA. *Science* **343**, 72-76 (2014).
54. Purshouse, K., Friman, E. T., Boyle, S., Dewari, P. S., Grant, V., Hamdan, A., Morrison, G. M., Brennan, P. M., Beentjes, S. V., Pollard, S. M., & Bickmore, W. A. Oncogene expression from extrachromosomal DNA is driven by copy number amplification and does not require spatial clustering in glioblastoma stem cells. *eLife* **11**, e80207 (2022).
 55. Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., Dewitt, J., Gritsch, S., Perez, E.M., Gonzalez Castro, L.N., Lan, X., Druck, N., Rodman, C., Dionne, D., Kaplan, A., Bertalan, M.S., Small, J., Pelton, K., Becker, S., Bonal, D., Nguyen, Q.-D., Servis, R.L., Fung, J.M., Mylvaganam, R., Mayr, L., Gojo, J., Haberler, C., Geyeregger, R., Czech, T., Slave, I., Nahed, B.V., Curry, W.T., Carter, B.S., Wakimoto, H., Brastianos, P.K., Batchelor, T.T., Stemmer-Rachamimov, A., Martinez-Lage, M., Frosch, M.P., Stamenkovic, I., Riggi, N., Rheinbay, E., Monje, M., Rozenblatt-Rosen, O., Cahill, D.P., Patel, A.P., Hunter, T., Verma, I.M., Ligon, K.L., Louis, D.N., Regev, A., Bernstein, B.E., Tirosh, I., & Suvà, M.L. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835-849. e821 (2019).
 56. Chaligne, R., Gaiti, F., Silverbush, D., Schiffman, J.S., Weisman, H.R., Kluegel, L., Gritsch, S., Deochand, S.D., Gonzalez Castro, L.N., Richman, A.R., Klughammer, J., Biancalani, T., Muus, C., Sheridan, C., Alonso, A., Izzo, F., Park, J., Rozenblatt-Rosen, O., Regev, A., Suvà, M.L. & Landau, D.A. Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nature Genetics* **53**, 1469-1479 (2021).
 57. Rahme, G. J., Javed, N. M., Puorro, K. L., Xin, S., Hovestadt, V., Johnstone, S. E., & Bernstein, B. E. Modeling epigenetic lesions that cause gliomas. *Cell* **186**, 3674-3685. e3614 (2023).
 58. Lancho, O. & Herranz, D. The MYC enhancer-ome: long-range transcriptional regulation of MYC in cancer. *Trends in cancer* **4**, 810-822 (2018).
 59. Lafontaine, D.L., Yang, L., Dekker, J. & Gibcus, J.H. Hi-C 3.0: Improved Protocol for Genome-Wide Chromosome Conformation Capture. *Current Protocols* **1**, e198 (2021).
 60. Cox, D., Yuncken, C. & Spriggs, A. MINUTE CHROMATIN BODIES IN MALIGNANT TUMOURS OF CHILDHOOD. *The Lancet* **286**, 55-58 (1965).
 61. Spriggs, A.I., Boddington, M.M. & Clarke, C.M. Chromosomes of human cancer cells. *British medical journal* **2**, 1431-1435 (1962).
 62. Møller, H. D., Mohiyuddin, M., Prada-Luengo, I., Sailani, M. R., Halling, J. F., Plomgaard, P., Maretty, L., Hansen, A. J., Snyder, M. P., Pilegaard, H., Lam, H. Y. K., & Regenberg, B. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nature Communications* **9**, 1069 (2018).

63. Paulsen, T., Kumar, P., Koseoglu, M.M. & Dutta, A. Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. *Trends in Genetics* **34**, 270-278 (2018).
64. Wu, S., Turner, K.M., Nguyen, N., Raviram, R., Erb, M., Santini, J., Luebeck, J., Rajkumar, U., Diao, Y., Li, B., Zhang, W., Jameson, N., Corces, M.R., Granja, J.M., Chen, X., Coruh, C., Abnousi, A., Houston, J., Ye, Z., Hu, R., Yu, M., Kim, H., Law, J.A., Verhaak, R.G., Hu, M., Furnari, F.B., Chang, H.Y., Ren, B., Bafna, V., & Mischel, P.S. Circular ecDNA promotes accessible chromatin and high oncogene expression. *Nature* **575**, 699-703 (2019).
65. Fan, X., Yang, C., Li, W., Bai, X., Zhou, X., Xie, H., Wen, L., & Tang, F. SMOOTH-seq: single-cell genome sequencing of human cells on a third-generation sequencing platform. *Genome Biology* **22**, 195 (2021).
66. Chang, L., Deng, E., Wang, J., Zhou, W., Ao, J., Liu, R., Su, D., & Fan, X. Single-cell third-generation sequencing-based multi-omics uncovers gene expression changes governed by ecDNA and structural variants in cancer cells. *Clinical and Translational Medicine* **13**, e1351 (2023).
67. Chamorro González, R., Conrad, T., Stöber, M.C., Xu, R., Giurgiu, M., Rodriguez-Fos, E., Kasack, K., Brückner, L., van Leen, E., Helmsauer, K., Dorado Garcia, H., Stefanova, M.E., Hung, K.L., Bei, Y., Schmelz, K., Lodrini, M., Mundlos, S., Chang, H.Y., Deubzer, H.E., Sauer, S., Eggert, A., Schulte, J.H., Schwarz, R.F., Haase, K., Koche, R.P., & Henssen, A.G. Parallel sequencing of extrachromosomal circular DNAs and transcriptomes in single cancer cells. *Nature Genetics* **55**, 880-890 (2023).
68. Møller, H.D., Parsons, L., Jørgensen, T.S., Botstein, D. & Regenberg, B. Extrachromosomal circular DNA is common in yeast. *Proceedings of the National Academy of Sciences* **112**, E3114-E3122 (2015).
69. Koche, R.P., Rodriguez-Fos, E., Helmsauer, K., Burkert, M., MacArthur, I.C., Maag, J., Chamorro, R., Munoz-Perez, N., Puiggròs, M., Dorado Garcia, H., Bei, Y., Röefzaad, C., Bardin, V., Szymansky, A., Winkler, A., Thole, T., Timme, N., Kasack, K., Fuchs, S., Klironomos, F., Thiessen, N., Blanc, E., Schmelz, K., Künkele, A., Hundsdörfer, P., Rosswog, C., Theissen, J., Beule, D., Deubzer, H., Sauer, S., Toedling, J., Fischer, M., Hertwig, F., Schwarz, R.F., Eggert, A., Torrents, D., Schulte, J.H., & Henssen, A.G. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nature Genetics* **52**, 29-34 (2020).
70. Kim, H., Nguyen, N. P., Turner, K., Wu, S., Gujar, A. D., Luebeck, J., Liu, J., Deshpande, V., Rajkumar, U., Namburi, S., Amin, S. B., Yi, E., Menghi, F., Schulte, J. H., Henssen, A. G., Chang, H. Y., Beck, C. R., Mischel, P. S., Bafna, V., & Verhaak, R. G. W. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nature Genetics* **52**, 891-897 (2020).
71. Wu, S., Bafna, V., Chang, H.Y. & Mischel, P.S. Extrachromosomal DNA: an emerging hallmark in human cancer. *Annual Review of Pathology: Mechanisms of Disease* **17**, 367-386 (2022).

72. Verhaak, R.G., Bafna, V. & Mischel, P.S. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nature Reviews Cancer* **19**, 283-288 (2019).
73. Wu, S., Bafna, V. & Mischel, P.S. Extrachromosomal DNA (ecDNA) in cancer pathogenesis. *Current Opinion in Genetics & Development* **66**, 78-82 (2021).
74. Curtis, E.J., Rose, J.C., Mischel, P.S. & Chang, H.Y. Extrachromosomal DNA: Biogenesis and Functions in Cancer. *Annual Review of Cancer Biology* **8**, null (2024).
75. Chowdhry, S., Garcia, S., Nguyen, N.-P., Celeste, A., Tse, E., Milutinovic, S., Wilkinson, D., Hassig, C., & Kasibhatla, S. Abstract 1520: Replication stress and the inability to repair damaged DNA, the potential “Achilles' heel” of ecDNA+ tumor cells. *Cancer Research* **82**, 1520-1520 (2022).
76. Bei, Y., Bramé, L., Kirchner, M., Fritsche-Guenther, R., Kunz, S., Bhattacharya, A., Köppke, J., Proba, J., Wittstruck, N., Sidorova, O.A., González, R.C., Garcia, H.D., Brückner, L., Xu, R., Giurgiu, M., Rodriguez-Fos, E., Koche, R., Schmitt, C., Schulte, J.H., Eggert, A., Haase, K., Kirwan, J., Hagemann, A.I.H., Mertins, P., Dörr, J.R., & Henssen, A.G. Amplicon structure creates collateral therapeutic vulnerability in cancer. *bioRxiv*, 2022.2009.2008.506647 (2022).
77. Noorani, I., Mischel, P.S. & Swanton, C. Leveraging extrachromosomal DNA to fine-tune trials of targeted therapy for glioblastoma: opportunities and challenges. *Nature Reviews Clinical Oncology* **19**, 733-743 (2022).
78. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
79. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10-12 (2011).
80. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
81. Open2C, Abdennur, N., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., Imakaev, M., & Venev, S. V. Pairtools: from sequencing data to chromosome contacts. *bioRxiv*, 2023.2002.2013.528389 (2023).
82. Abdennur, N. & Mirny, L.A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311-316 (2019).
83. Hunter, J.D. Matplotlib: A 2D graphics environment. *Computing in science & engineering* **9**, 90-95 (2007).
84. Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* **9**, 9354 (2019).

85. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 (2019).
86. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, complex systems* **1695**, 1-9 (2006).
87. Open2C, Abdennur, N., Abraham, S., Fudenberg, G., Flyamer, I. M., Galitsyna, A. A., Goloborodko, A., Imakaev, M., Oksuz, B. A., Venev, S. V., & Xiao, Y. Cooltools: enabling high-resolution Hi-C analysis in Python. *BioRxiv*, 2022.2010. 2031.514564 (2022).
88. Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T. J., Dixon, J. R., & Ecker, J. R. Robust single-cell Hi-C clustering by convolution- and random-walk-based imputation. *Proceedings of the National Academy of Sciences* **116**, 14011-14018 (2019).
89. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
90. 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).
91. Dong, W., Moses, C. & Li, K. in Proceedings of the 20th international conference on World wide web 577-586 (2011).
92. 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. *Nature Methods* **16**, 1289-1296 (2019).
93. Chakraborty, A., Wang, J.G. & Ay, F. dcHiC detects differential compartments across multiple Hi-C datasets. *Nature Communications* **13**, 6827 (2022).
94. Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., & Zhou, X. J. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Research* **44**, e70-e70 (2015).
95. Hounkpe, B.W., Chenou, F., de Lima, F. & De Paula, Erich V. HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. *Nucleic Acids Research* **49**, D947-D955 (2020).
96. Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95-98 (2016).

97. Yu, M., Abnoui, A., Zhang, Y., Li, G., Lee, L., Chen, Z., Fang, R., Lagler, T. M., Yang, Y., Wen, J., Sun, Q., Li, Y., Ren, B., & Hu, M. SnapHiC: a computational pipeline to identify chromatin loops from single-cell Hi-C data. *Nature Methods* **18**, 1056-1059 (2021).
98. Gu, Z. & Hübschmann, D. rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics* **39** (2022).
99. 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).
100. Wang, X., Xu, J., Zhang, B., Hou, Y., Song, F., Lyu, H., & Yue, F. Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nature methods* **18**, 661-668 (2021).
101. Wang, X., Luan, Y. & Yue, F. EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Science Advances* **8**, eabn9215 (2022).
102. Nair, V. & Hinton, G.E. in Proceedings of the 27th international conference on machine learning (ICML-10) 807-814 (2010).
103. Krizhevsky, A., Sutskever, I. & Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012).
104. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **15**, 1929-1958 (2014).
105. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
106. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. & Desmaison, A. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).
107. Kingma, D.P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
108. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
109. Bridle, J.S. in Neurocomputing: Algorithms, architectures and applications 227-236 (Springer, 1990).

110. Cao, Y., Fu, L., Wu, J., Peng, Q., Nie, Q., Zhang, J., & Xie, X. Integrated analysis of multimodal single-cell data with structural similarity. *Nucleic Acids Research* **50**, e121-e121 (2022).
111. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nature Methods* **18**, 1333-1341 (2021).
112. Wolock, S.L., Lopez, R. & Klein, A.M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems* **8**, 281-291. e289 (2019).
113. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research* **44**, W90-W97 (2016).
114. van Galen, P., Hovestadt, V., Wadsworth Ii, M. H., Hughes, T. K., Griffin, G. K., Battaglia, S., Verga, J. A., Stephansky, J., Pastika, T. J., Lombardi Story, J., Pinkus, G. S., Pozdnyakova, O., Galinsky, I., Stone, R. M., Graubert, T. A., Shalek, A. K., Aster, J. C., Lane, A. A., & Bernstein, B. E. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell* **176**, 1265-1281.e1224 (2019).

DISCUSSION

Single-cell epigenetic tool is akin to the Hubble Telescope, not for exploring outer space but for navigating gene regulation at the cellular resolution that was not reached before. The resolution achieved by these methods has enabled the identification of disease-relevant cell types, and characterization of regulatory elements that overlap with disease-associated variants, and the construction of cell type-specific gene regulatory networks that are critical in the developmental process^{1,2}. However, despite the identification of a large pool of candidate cis-regulatory elements, the validation of their activities and specificities lags significantly behind the pace at which new data is generated. Conventional exogenous methods to assess enhancer activity do not recapitulate the native chromatin context or the cell-type-specific activity of enhancers, although recent advancements have started to fill in the gaps^{3,4}. Endogenous methods, such as CRISPR-based enhancer assays, attempt to overcome these limitations by evaluating enhancer activity within their native genomic context but are currently limited by low throughput and efficiency^{5,6}. The limitations highlight the necessity for continued innovation in the enhancer validation strategies for more efficient, scalable, and context-specific methods. Understanding the real activity and specificity of enhancers in their native context will have significant implications for applications like targeted gene therapies, where precise control over gene expression is essential.

Epigenetics serves as both a historical record and a future guide for cellular identity and function. The epigenetic landscape of a cell encodes rich information snapping a wide range from past exposures to developmental history, and to specific genes that are going to be activated or repressed^{7,8,9}. The integration of single-cell epigenetic data with other omics layers, such as transcriptomics or proteomics, makes it possible to construct predictive models of cellular behavior.

Such models could, in theory, predict which cells and how they will respond to various stimulations or perturbations¹⁰.

However, even though predictive models have been proposed to forecast future cellular states, the true trajectory of a cell remains elusive without data validation. The next frontier in single-cell epigenetics lies in the ability to capture temporal dynamics. While providing detailed snapshots of epigenetic states, traditional sequencing methods inherently lack the temporal resolution necessary to fully grasp the dynamic and transient nature of regulatory processes. Several approaches have been developed to track the temporal activity or fate of cells, offering unique insights into different aspects of cellular dynamics. One method involves labeling newly synthesized RNA, allowing researchers to capture transcriptional responses to stimuli, such as immune activation, within a short time window^{11,12}. The other approach involves real-time sampling RNA from the same cells at multiple time points. This longitudinal sampling provides a continuous view of the transcriptomic landscape, offering valuable insights into how gene expression evolves within individual cells^{13,14}. On the other hand, lineage tracing methods provide another dimension of temporal analysis by marking the progression of cells from their origins to their differentiated states, therefore allowing researchers to reconstruct the differentiation trajectories of cells^{15,16,17}. Despite these advances, there is still a relative scarcity of studies that specifically investigate the temporal dynamics of epigenetic regulation. Future research should prioritize the development of tools capable of capturing the temporal progression of epigenetic modifications in individual cells. This could involve the integration of live-cell imaging with sequencing, or novel biochemical tools could be designed to label and track specific epigenetic marks in real time, providing a dynamic view of how these changes influence gene expression and cellular behavior over time.

The quest to explore epigenetic regulation at the single-cell level will persist, much like the unending exploration of the universe by astronomers. The emergence and rapid development of single-cell omics tools are revolutionizing our ability to create a "star map" of any biological system, capturing not only the static states of cells but also their dynamic transitions across temporal and spatial dimensions. By leveraging single-cell technologies, scientists can now observe how individual cells within tissue respond to internal and external cues, tracing the pathways that lead to specific cellular outcomes. Furthermore, the ability to capture and validate these regulatory processes across temporal and spatial dimensions will be crucial for understanding how cells evolve and respond to their environment. These advancements will ultimately propel us toward a deeper understanding of the causal relationships between epigenetic changes and cellular outcomes, paving the way for new insights into development, disease progression, therapeutic interventions and beyond.

Reference

1. Li, Y.E., Preissl, S., Miller, M., Johnson, N.D., Wang, Zihan, Jiao, H., Zhu, C., Wang, Zhaoning, Xie, Y., Poirion, O., Kern, C., Pinto-Duarte, A., Tian, W., Siletti, K., Emerson, N., Osteen, J., Lucero, J., Lin, L., Yang, Q., Zhu, Q., Zemke, N., Espinoza, S., Yanny, A.M., Nyhus, J., Dee, N., Casper, T., Shapovalova, N., Hirschstein, D., Hodge, R.D., Linnarsson, S., Bakken, T., Levi, B., Keene, C.D., Shang, J., Lein, E., Wang, A., Behrens, M.M., Ecker, J.R., Ren, B., (2023). A comparative atlas of single-cell chromatin accessibility in the human brain. *Science*, 382(7044)
2. Kamimoto, K., Stringa, B., Hoffmann, C. M., Jindal, K., Solnica-Krezel, L., & Morris, S. A. (2023). Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949), 742–751.
3. Zhao, S., Hong, C. K. Y., Myers, C. A., Granas, D. M., White, M. A., Corbo, J. C., & Cohen, B. A. (2023). A single-cell massively parallel reporter assay detects cell-type-specific gene regulation. *Nature genetics*, 55(2), 346–354.
4. Mangan, R. J., Alsina, F. C., Mosti, F., Sotelo-Fonseca, J. E., Snellings, D. A., Au, E. H., Carvalho, J., Sathyan, L., Johnson, G. D., Reddy, T. E., Silver, D. L., & Lowe, C. B. (2022). Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell*, 185(24), 4587–4603.e23.
5. 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. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1-2), 377–390.e19.
6. Xie, S., Duan, J., Li, B., Zhou, P., & Hon, G. C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular cell*, 66(2), 285–299.e5.
7. Kanherkar, R. R., Bhatia-Dey, N., & Csoka, A. B. (2014). Epigenetics across the human lifespan. *Frontiers in cell and developmental biology*, 2, 49.
8. Nashun, B., Hill, P. W., & Hajkova, P. (2015). Reprogramming of cell fate: epigenetic memory and the erasure of memories past. *The EMBO journal*, 34(10), 1296–1308.
9. Mooijman, D., Dey, S. S., Boisset, J. C., Crosetto, N., & van Oudenaarden, A. (2016). Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nature biotechnology*, 34(8), 852–856.
10. 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. (2020). Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, 183(4), 1103–1116.e20.
11. Muhar, M., Ebert, A., Neumann, T., Umkehrer, C., Jude, J., Wieshofer, C., Rescheneder, P.,

- Lipp, J. J., Herzog, V. A., Reichholf, B., Cisneros, D. A., Hoffmann, T., Schlapansky, M. F., Bhat, P., von Haeseler, A., Köcher, T., Obenauf, A. C., Popow, J., Ameres, S. L., & Zuber, J. (2018). SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science*, 360(6390), 800–805.
12. Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., Gagneur, J., & Cramer, P. (2016). TT-seq maps the human transient transcriptome. *Science*, 352(6290), 1225–1228.
 13. Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., Vorholt, J. A., & Deplancke, B. (2022). Live-seq enables temporal transcriptomic recording of single cells. *Nature*, 608(7924), 733–740.
 14. Horns, F., Martinez, J. A., Fan, C., Haque, M., Linton, J. M., Tobin, V., Santat, L., Maggiolo, A. O., Bjorkman, P. J., Lois, C., & Elowitz, M. B. (2023). Engineering RNA export for measurement and manipulation of living cells. *Cell*, 186(17), 3642–3658.e32.
 15. Wagner, D. E., & Klein, A. M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. *Nature reviews. Genetics*, 21(7), 410–427.
 16. Chen, W., Choi, J., Li, X., Nathans, J. F., Martin, B., Yang, W., Hamazaki, N., Qiu, C., Lallanne, J. B., Regalado, S., Kim, H., Agarwal, V., Nichols, E., Leith, A., Lee, C., & Shendure, J. (2024). Symbolic recording of signalling and cis-regulatory element activity to DNA. *Nature*, 10.1038/s41586-024-07706-4.
 17. Choi, J., Chen, W., Minkina, A., Chardon, F. M., Suiter, C. C., Regalado, S. G., Domcke, S., Hamazaki, N., Lee, C., Martin, B., Daza, R. M., & Shendure, J. (2022). A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature*, 608(7921), 98–107.