UNIVERSITY OF CALIFORNIA SAN DIEGO

Interplay Between RNA and Chromatin: From Genome-Wide Mapping to Functional Implications

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

in

Bioinformatics and Systems Biology

by

Xingzhao Wen

Committee in charge:

Professor Sheng Zhong, Chair
Professor Bing Ren, Co-Chair
Professor Vineet Bafna
Professor Christopher Benner
Professor Joseph R. Ecker

2024

The Dissertation of Xingzhao Wen 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 husband, Bojing, my anchor, and constant source of strength and joy.

To my loving parents, my first inspiration in science and engineering, who always love and support me unconditionally.

In memory of my grandfather, your wisdom continues to echo in my heart.

To mentors, friends, and peers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

TF	Transcription factor
H3K27ac	Acetylation of histone H3 lysine 27
H3K4me2	Di-methylation of histone H3 lysine 4
H3K4me3	Tri-methylation of histone H3 lysine 4
iMARGI	In situ mapping of RNA–genome interactome
MUSIC	Multinucleic acid interaction mapping in single cells
ChIP-seq	Chromatin immunoprecipitation sequencing
TAD	Topologically associating domain

ACKNOWLEDGEMENTS

Five years have flown by in what feels like an instant, yet so much has changed. I'm both happy and proud to have dedicated the most creative and dynamic years of my life to exploring questions that genuinely excite me. This journey would not have been so pleasant and meaningful without the support and companionship of so many people.

First, I want to extend my tremendous thanks to my PhD advisor, Dr. Sheng Zhong. When I first approached him about a research assistant position, he not only welcomed me but also proposed writing a book on the 3D genome—a topic I was initially unfamiliar with. After three months of intensive literature review, I completed an ebook that laid the foundation for my understanding of the field of 3D genomics and since then guided my PhD research. Sheng's pure passion for science, a keen eye for detail, and the amazing ability to transform my pessimistic conclusions into exciting hypotheses have been a constant source of inspiration. His mentorship, knowledge, and encouragement to let me explore freely have been pivotal to my PhD journey.

The day-to-day life of a graduate student is enriched by lab members. I thank Dr. Zhangming Yan and Dr. Weizhong Chen for introducing me to RNA-chromatin interaction data analysis and teaching me good bioinformatics practices. Riccardo Callendrelli has been a wonderful friend and trustworthy collaborator on many projects. I am also grateful to our talented wet lab scientists Dr. Zhifei Luo, Dr. Kara Johnson, Dr. Junchen Chen, Dr. Tri Nyuen, Wenxin Zhao, Dr. Shuanghong Xue, Dr. Ming Xu, Tianyang Xu, and many others who patiently shared their knowledge, answered my questions, and brainstormed new ideas with me. Most importantly, I'm so grateful they never said no or pushed me away for my constant unnoticed interruptions.

I want to extend my gratitude to my committee members Dr. Bing Ren, Dr. Christopher Benner, Dr. Joseph Ecker, and Dr. Vineet Bafna, whose foundational works in epigenetics,

bioinformatics, chromatin organization have profoundly influenced my research. Special thanks to Dr. Chris Benner, who shared a lot of wisdom on grad school and offered helpful suggestions for my thesis. Professors Trey Ideker, Vineet Bafna, Glenn Tesler, Lawrence Saul, and Julian McAuley have also been instrumental in teaching and inspiring many concepts that resonate in my research. My undergraduate mentors, Professors Shuaicheng Li, Xiao Sun, Jing Tu, Ian Holmes, and Indika Rajapakse, ignited my interest in research and encouraged me to tackle exciting questions for graduate school.

Being part of the NIH-funded 4DN Nucleome Consortium throughout my PhD has been an amazing opportunity. Weekly Zoom calls and annual meetings allowed me to interact with leading scientists and talented trainees in 3D genomics. Special thanks to Dr. Job Dekker, Dr. Jian Ma, Dr. Wenbo Li, and others for their insightful questions and suggestions during presentations and discussions.

Outside the lab, my support system has been invaluable. Xiaochen, Alex, Wenliang, Ruochi, Hsuan-lin, and Jie have always been there to offer advice and encouragement during both academic and personal challenges.

My family has been the cornerstone of my journey. My dear mom and dad allowed me to make my own decisions from a very young age, supporting and loving me unconditionally as I pursued my dreams.

Last but not least, I want to thank my husband, Bojing Blair Jia, my dearest and nearest friend, and the most wonderful thing that has ever happened to me. We met as classmates in the graduate program and grew together as scientists. His fervent passion for science, unique preference for ‘tough and long’ challenges, and unbeatable strong will in the face of adversity have brightened and shaped my life day by day.

Chapter 1, in full, is a reprint of the material as it appears in Nature Communications 2023. Calandrelli*, R, Wen*, X, Richard, JL Charles, Luo, Z, Nguyen, TC, Chen, CJ, Qi, Z., Xue, S., Chen, W., Yan, Z., Wu, W., Zaleta-Rivera, K., Hu, R., Yu, M., Wang, Y., Li, W., Ma, J., Ren, B., & Zhong, S. Genome-wide analysis of the interplay between chromatin-associated RNA and 3D genome organization in human cells. Nature Communications, 14(1), 6519. The dissertation author was co-first author of this paper.

Chapter 2, in full, is currently being prepared for submission of the material. Xingzhao, Wen; Sheng, Zhong. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Nature 2024. Wen*, X., Luo*, Z., Zhao, W., Calandrelli, R., Nguyen, T. C., Wan, X., Charles Richard, J. L., & Zhong, S. (2024). Single-cell multiplex chromatin and RNA interactions in ageing human brain. Nature, 1-9. The dissertation author was co-first author of this paper.

VITA

2018	Bachelor of Engineering, Southeast University
2019-2020	Teaching Assistant, University of California San Diego
2024	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Wen*, X., Luo*, Z., Zhao, W., Calandrelli, R., Nguyen, T.C., Wan, X., Charles Richard, J.L., Zhong, S., 2024. Single-cell multiplex chromatin and RNA interactions in ageing human brain. Nature 628, 648–656.

Calandrelli*, R., **Wen*, X**., Charles Richard, J.L., Luo, Z., Nguyen, T.C., Chen, C.-J., Qi, Z., Xue, S., Chen, W., Yan, Z., Wu, W., Zaleta-Rivera, K., Hu, R., Yu, M., Wang, Y., Li, W., Ma, J., Ren, B., Zhong, S., 2023. Genome-wide analysis of the interplay between chromatin-associated RNA and 3D genome organization in human cells. Nat. Commun. 14, 6519.

Wen, X., Zhong, S., 2023. Alu transposable elements rewire enhancer-promoter network through RNA pairing. Mol. Cell 83, 3234–3235.

Johnson, K.L., Qi, Z., Yan, Z., **Wen, X**., Nguyen, T.C., Zaleta-Rivera, K., Chen, C.-J., Fan, X., Sriram, K., Wan, X., Chen, Z.B., Zhong, S., 2021. Revealing protein-protein interactions at the transcriptome scale by sequencing. Mol. Cell 81, 4091–4103.e9.

Lindsly, S., Jia, W., Chen, H., Liu, S., Ronquist, S., Chen, C., **Wen, X.,** Stansbury, C., Dotson, G.A., Ryan, C., Rehemtulla, A., Omenn, G.S., Wicha, M., Li, S.C., Muir, L., Rajapakse, I., 2021. Functional organization of the maternal and paternal human 4D Nucleome. iScience 24, 103452.

Huang*, M., Yang*, Y., **Wen*, X**., Xu, W., Lu, N., Sun, X., Tu, J., Lu, Z., 2021. Inferring single cell expression profiles from overlapped pooling sequencing data with compressed sensing strategy. Nucleic Acids Res. 49, 7995–8006.

Pan, X., Liu, B., **Wen, X.,** Liu, Y., Zhang, X., Li, S., Li, S., 2019. D-GPM: A Deep Learning Method for Gene Promoter Methylation Inference. Genes 10.

Yan, Z., Huang, N., Wu, W., Chen, W., Jiang, Y., Chen, J., Huang, X., **Wen, X.,** Xu, J., Jin, Q., Zhang, K., Chen, Z., Chien, S., Zhong, S., 2019. Genome-wide colocalization of RNA-DNA interactions and fusion RNA pairs. Proc. Natl. Acad. Sci. U. S. A. 116, 3328–3337.

* These authors contributed equally to this work.

FIELD OF STUDY

Major Field: Bioinformatics
Studies in Genomics
Professor Sheng Zhong

ABSTRACT OF THE DISSERTATION

Interplay Between RNA and Chromatin: From Genome-Wide Mapping to Functional Implications

by

Xingzhao Wen

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2024

Professor Sheng Zhong, Chair
Professor Bing Ren, Co-Chair

The spatial interactions between chromatin, RNA, and proteins inside the cell nucleus underpin the exquisite execution of cellular functions. RNA, one of the most versatile molecules, plays crucial roles in forming membrane less nuclear bodies, modulating chromatin architecture,

and regulating gene expression. Despite a few well-studied ncRNAs, the roles of many other RNAs in gene regulation and chromatin organization remain unclear. This thesis aims to provide a systematic approach to delineate functionally important chromatin-associated RNAs (caRNAs).

Chapter 1 establishes a baseline model of caRNA localization relative to chromatin structure and the influence of caRNAs on chromatin organization. By jointly comparing chromatin structure with genome-wide RNA-chromatin interactions in human embryonic stem cells, I found that caRNA localization is largely confined and shaped by local chromatin structures such as compartments, topologically associated domains (TADs), and chromatin loops. Chromatin loop anchors are hotspots for caRNA interactions, and depletion of loop anchor caRNAs leads to the formation of new chromatin loops. Chapter 2 addresses the challenge of identifying functional caRNA-chromatin interactions, where nascent transcripts significantly contribute to RNA-chromatin interactions, complicating the identification of functional caRNAs. I proposed a modified version of the iMARGI technique that introduces RNase A treatment before crosslinking. This method effectively removes non-protected caRNAs, such as nascent transcripts, leaving only stringent and protected interactions. We found that RNase-inaccessible caRNAs exhibit higher sequence conservation and are enriched at transcription factor binding sites, histone modifications, and other functional genomic regions. Additionally, we identified specific RNA species associated with transcription factor and histone modification sites. Chapter 3 presents the development of a technology enabling the joint profiling of higher-order nucleic acid interactions at single-cell resolution. My colleagues and I developed MUSIC, a technique that allows simultaneous profiling of DNA-DNA, RNA-DNA, and RNA-RNA interactions along with the transcriptome at single-cell resolution. Applying MUSIC to complex human tissues, such as the frontal cortex of control and Alzheimer's disease (AD) patients, we identified that transcriptomically aged cells exhibit

diminished local chromatin contacts and are enriched in AD patients. Female brain cells show high levels of XIST-chromatin interaction heterogeneity, and the loss of XIST-chrX enrichment is associated with increased chrX gene expression. Overall, this thesis provides valuable tools and insights into the functional roles of chromatin-associated RNAs. It offers new technologies that enable novel modalities and directions for studying the interplay between RNA and DNA in gene expression regulation.

INTRODUCTION

The cell nucleus hosts the genetic material of each cell within a highly confined space. This confinement promotes the intricate 3D architecture of chromatin along with nuclear RNA transcripts.

Remarkably, around 98% of the human genome comprises non-coding regions, sequences that do not encode proteins. While the coding regions are highly conserved across species, the non-coding portions drive much of the heterogeneity observed between organisms. And with rapid advancements in genomics technologies, it has become evident that 60% to 85% of the non-coding regions are transcribed into RNAs (Djebali et al., 2012) . To decode the information embedded in the human genome, studying its transcriptional output is essential. Beyond a few RNA species involved in translation machinery (mRNA, rRNA, tRNA), most transcripts remain within the nucleus throughout their lifetime (Dhanoa et al., 2018; Guo et al., 2020; Palazzo & Lee, 2018). Among these nuclear RNAs, chromatin-associated RNAs (caRNAs) are of particular interest due to their proximity to chromatin and their significant roles in gene transcription regulation (Calandrelli et al., 2020; Li et al., 2021; Sigova et al., 2015), chromatin spatial organization (Bonetti et al., 2020; Calandrelli et al., 2023; Quinodoz et al., 2021) and certain disease pathogenesis (Allou et al., 2021; Nalavade et al., 2013).

Despite exciting discoveries regarding certain well-studied non-coding RNAs, a systematic understanding of caRNA localization patterns across the genome and their impact on genome organization is still lacking. In Chapter 1, I systematically examine the genome-wide distribution of caRNA at various levels of chromatin structure, including chromosomal domains, chromosome compartments, topologically associating domains (TADs), and chromatin loops. I found that caRNA attachment peaks at transcription sites and decreases with increasing genomic distance,

while chromatin structure significantly influences caRNA localization patterns, generally confining the spread of caRNAs.

Although a general pattern of caRNA localization relative to chromatin structure can be established by comparing their signals, identifying regulatory or functional caRNAs remains challenging. This difficulty arises because the majority of caRNAs are nascent transcripts and observed caRNA-chromatin interactions often result from promiscuous interactions. To address this issue, in Chapter 2, I proposed an enhanced version of the iMARGI technology, incorporating RNase treatment at an early stage to remove non-stringent RNA-chromatin interactions, typically involving nascent transcripts. This method allowed us to identify groups of caRNAs protected from RNase digestion, revealing sequences with distinct functional motifs. We found that these caRNAs are usually enriched in regulatory genomic regions, including open chromatin, transcription factor binding sites, and histone modifications. This chapter marks a pioneering effort to eliminate diffusive nascent transcripts and uncover stable, regulatory caRNAs along with their target genomic regions.

Nuclear structures and gene transcription levels exhibit extreme heterogeneity within and between cell populations. Single-cell Hi-C techniques have revealed that previously considered stable TAD patterns are dynamic across individual cells, and genomic structures vary significantly throughout the cell cycle. Recent advances have highlighted variations in genome structure between and within cell types, but the relationship between these variations and cellular function remains poorly understood. Regulatory RNAs and their localization patterns result from a combination of transcription profiles and genome organization, with many RNAs functioning by leveraging the 3D genome structure (Engreitz et al., 2013; Pandey et al., 2008). Understanding the regulatory roles of caRNAs requires examining heterogeneous conditions through a multi-omic

approach that can simultaneously measure genome structure and function in single cells. In Chapter 3, I developed a full data processing pipeline and analysis methods for MUSIC (Multinucleic acid interaction mapping in single cells), a breakthrough sequencing technique that enables simultaneous profiling of genome organization, transcriptional landscape, and RNA-chromatin interactions at single-cell resolution. Applying MUSIC to human embryonic cells and human frontal cortex tissues from healthy donors and Alzheimer's patients, we uncovered abnormal loss of local chromatin structures in cells from Alzheimer's disease patients, alongside cell aging. We also observed significant heterogeneity in XIST-chrX interactions in neurons, where only a fraction of cells exhibited enriched XIST localization on chrX, while in other cells, XIST interacted broadly across the genome. The decreased XIST attachment level on chrX correlated with increased chrX gene expression, suggesting a critical role for XIST in gene expression regulation and implicating its disruption in disease conditions.

Collectively, this thesis provides a comprehensive view of the global localization patterns of caRNAs on chromosomes and how they are shaped by chromatin topology at different scales. It also introduces novel technologies with computational analysis methods to delineate the regulatory caRNA-chromatin interaction landscape and the nucleic acid interaction network at single-nucleus resolution.

Chapter 1 Genome-wide analysis of the interplay between chromatin-associated RNA and 3D genome organization in human cells

Abstract

The interphase genome is dynamically organized in the nucleus and decorated with chromatin-associated RNA (caRNA). It remains unclear whether the genome architecture modulates the spatial distribution of caRNA and vice versa. Here, we generate a resource of genome-wide RNA-DNA and DNA-DNA contact maps in human cells. These maps reveal the chromosomal domains demarcated by locally transcribed RNA, hereafter termed RNA-defined chromosomal domains. Further, the spreading of caRNA is constrained by the boundaries of topologically associating domains (TADs), demonstrating the role of the 3D genome structure in modulating the spatial distribution of RNA. Conversely, stopping transcription or acute depletion of RNA induces thousands of chromatin loops genome-wide. Activation or suppression of the transcription of specific genes suppresses or creates chromatin loops straddling these genes. Deletion of a specific caRNA-producing genomic sequence promotes chromatin loops that straddle the interchromosomal target sequences of this caRNA. These data suggest a feedback loop where the 3D genome modulates the spatial distribution of RNA, which in turn affects the dynamic 3D genome organization.

Introduction

The interphase genome is highly organized (Dekker et al., 2017). The multiscale organizational features of the genome have been characterized, including A/B compartments (Lieberman-Aiden et al., 2009), topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012), and chromatin loops (Rao et al., 2014). This multiscale organization begs the question of what the functions of such an intricate architecture are. Transcriptional regulation is

one of the possible functions and the most extensively studied function. In this direction, the genome architecture is shown to regulate the transcription of specific genes (Dekker et al., 2017; Lupiáñez et al., 2015; Sun et al., 2018), but it remains debatable whether the genome architecture has a widespread role in modulating the transcription of many genes (Rao et al., 2017a). Moreover, it remains unclear if the 3D genome's regulatory roles are limited to transcriptional regulation. Other possible functions have rarely been tested. Here, we test another possible function, namely regulating spatial localization of chromatin-associated RNA (caRNA) (Holmes et al., 1972).

After initial debates (Holmes et al., 1972), caRNA has been recognized as an integral component of interphase chromosomes rather than passive degradation products (Hall and Lawrence, 2016; Nozawa et al., 2017; Nozawa and Gilbert, 2019; Pederson, 2019; Rodríguez-Campos and Azorín, 2007). Growing evidence confirms that caRNA regulates gene transcription and RNA splicing (Calandrelli et al., 2020; Li et al., 2017a; Miao et al., 2018; Morris et al., 2004; Penny et al., 1996; Place et al., 2008; Yang et al., 2015). These regulatory roles often depend on caRNA's spatial localization within the nucleus (Chen et al., 2018; Quinodoz et al., 2021; Yang et al., 2015; Yin et al., 2020). Depending on their spatial localizations, caRNA can orchestrate the organization of nuclear bodies and compartments (Chen et al., 2018; Chu et al., 2017; Quinodoz et al., 2021) and foster the formation of transcriptionally silent or active chromosomal domains (Lee, 2011; Quinodoz et al., 2021; Rinn and Chang, 2012). However, it remains unclear how the caRNAs are spatially organized in the context of the multiscale genome architecture; whether there is any specificity in the spatial distribution of caRNAs; if there is, how is such specificity regulated; and in turn, whether the spatial localization of caRNA modulates the dynamic organization of the genome?

Guided by these questions, we generate high-resolution genome-wide RNA-DNA contact maps (Bell et al., 2018; Bonetti et al., 2020; Li et al., 2017a; Sridhar et al., 2017; Wu et al., 2019) in human cells using in situ Mapping of RNA-Genome Interaction (iMARGI) (Sridhar et al., 2017; Wu et al., 2019). iMARGI captures RNA-genome associations by jointly sequencing caRNA and their associated genomic sequences with paired-end sequence reads. iMARGI can differentiate the sequencing reads originating from RNA (iMARGI RNA-end reads) or genomic DNA (iMARGI DNA-end reads). We also use in situ Hi-C (Hi-C) (Akgol Oksuz et al., 2021; Rao et al., 2014) to map genome-wide chromatin interactions. These maps reveal most caRNAs are associated with the genomic sequences within several megabases of their transcription sites.

To dissect any causal relationships between the 3D genome organization and caRNA, we generate RNA-DNA contact maps in the genetically engineered human cells where the TAD boundaries are deleted or inserted. Comparisons of these maps reveal the ability of TAD boundaries to constrain the spreading of caRNA on the chromosomes. These data demonstrate the 3D genome's functions in regulating the spatial localization of the caRNA. Moreover, we generate RNA-DNA and DNA-DNA contact maps in human cells undergoing either acute RNA depletion or deletion of a specific caRNA-producing sequence. These data reveal a suppressive role of between-anchor caRNA, i.e., the caRNA associated with the genomic region between the loop anchors, on chromatin looping. Thus, the spatial localization of the caRNA, in turn, modulates the dynamic 3D genome organization.

Results

Localized RNA-genome association and RNA-defined chromatin domains

We generated iMARGI data from human embryonic stem (H1), foreskin fibroblast (HFF), and chronic myelogenous leukemia (K562) cells in duplicates (Table S1). These data revealed the relative level of any gene's RNA attached to any genomic region (target region), hereafter called

the RNA attachment level (RAL) of this gene and target region, defined as the number of iMARGI read pairs with the RNA ends mapped to this gene and the DNA ends mapped to this target region (Sridhar et al., 2017). For example, in H1 ES cells the coding gene Jumonji and AT-Rich Interaction Domain Containing 2 (JARID2) exhibited large RAL in an approximately 5 Mb region containing the JARID2 gene (Figure 1.1a). Additionally, the non-coding gene Pvt1 Oncogene (PVT1) exhibited large RAL in an approximately 7 Mb region containing the PVT1 gene (Figure 1.1b). Overall, the average RAL of all the genes decreases as the genomic distance between the gene and the target region increases (Supplementary Figure 1-1a).

We represented iMARGI data as a contact matrix, where the rows represent the RNA ends of iMARGI read pairs, and the columns represent the corresponding DNA ends (Wu et al., 2019) (Figure 1.1c). A notable difference to Hi-C's symmetric contact matrix is that iMARGI's contact matrix is asymmetric. This is because RNA-DNA contacts are not necessarily reciprocal. Rectangular blocks of high-value entries emerged as a recurring pattern from iMARGI's contact matrix (Figure 1.1c). We identified the rectangular blocks using HOMER (Heinz et al., 2010) to call peaks on the rows of the contact matrix (row peaks), and in each row peak using HOMER to call one strongest peak in the columns (column peak). A pair of row peak and column peak defines a rectangular block. We identified 3,217, 2,019, and 2,468 rectangular blocks from H1, HFF, and K562 iMARGI data (Figure 1.1d). All the identified rectangular blocks overlap with the diagonal entries of iMARGI's contact matrix, suggesting that they represent localized RNA-genome associations where a RNA's target regions are near the transcription site of this caRNA. Each rectangular block corresponds to a unique chromatin domain, characterized by extensive genomic association of the RNA transcribed from within this domain. Hereafter we term such domains "RNA-association domains". The size of an RNA-association domain, represented by the width of

a rectangular block, can reach tens of megabases (Figure 1.1e). In summary, RNA-association domains emerged as a main feature of the genome-wide distributions of caRNA.

Correlation between 3D genome compartmentalization and RNA-chromatin association

The 3D genome is organized on different scales, including compartments, TADs, and chromatin loops (Dekker et al., 2017). We asked whether the RNA association on any genomic region correlates with this genomic region's 3D compartmentalization. To this end, we generated Hi-C data in H1, HFF, and K562 cells in duplicates and compared them with our iMARGI data (Table S1). We calculated the cumulative RAL (cRAL), the sum of the RAL of all the RNA, on every genomic region, defined as the number of iMARGI read pairs with the DNA ends mapped to this genomic region (Sridhar et al., 2017). The A/B compartments as indicated by Hi-C contact matrix's first eigenvector (PC1) (Mourad & Cuvier, 2015) exhibited a genome-wide correlation with cRAL (p-value < 2e-16, one way ANOVA), revealing a correlation between 3D genome compartmentalization and RNA-chromatin association.

We asked if the higher cRAL in the A compartment is completely attributable to a higher level of local transcription. To this end, we compared JARID2 and PVT1's RALs with A/B compartments (Mourad & Cuvier, 2015) (PC1 track, Figure 1.1a, b). Both JARID2 and PVT1 exhibited small but non-zero RALs in several A compartment genomic regions that are tens of megabases away from the JARID2 and PVT1 genes (Figure 1.1a, b). However, the B compartment genomic regions that are closer to the JARID2 and PVT1 genes did not exhibit association of JARID2 or PVT1 RNA (Figure 1.1a, b), suggesting an enrichment of target regions of long-range RNA-chromatin contacts in the A compartment. Thus, the higher cRAL in the A compartment is not completely due to a higher level of local transcription.

TAD boundaries insulate RNA-DNA contacts

TADs, where DNA sequences interact with each other more frequently than with the sequences outside, are important 3D genome features that are strongly correlated with transcriptional regulation (Dixon et al., 2012; Nora et al., 2012). We separately analyzed the RNA transcribed from within a TAD or the other regions of the same chromosome outside of this TAD. The chromatin attachment level of any RNA transcribed from within a TAD sharply decreases at the two boundaries of this TAD (p-value = 6.5e-16, Wilcoxon rank-sum test) (Figure 1.2a). Conversely, the attachment level of any RNA transcribed from outside of a TAD exhibits drastic changes at the TAD boundaries in the opposite direction (p-value = 2.6e-12, Wilcoxon rank-sum test) (Supplementary Figure 1-1b). These changes at TAD boundaries cannot be completely explained by the 1-dimensional genomic distance to the caRNA's transcription site. They suggest the possibility that a TAD boundary can insulate RNA-DNA contacts from the two sides of this boundary (cross-over RNA-DNA contacts).

We asked if altering the genomic sequence within a TAD boundary can affect the cross-over RNA-DNA contacts. First, we leveraged our previous finding that a CRISPR-mediated deletion of a HERV-H element (Chr13:55,578,227-55,584,087) (KO) within a TAD boundary from H9 human ES cells (WT) abolishes this TAD boundary (Zhang et al., 2019). We carried out iMARGI experiments on the KO and WT cells. We counted the numbers of cross-over and non-cross-over iMARGI read pairs in WT and KO. Compared to WT, KO exhibited an increased proportion in the cross-over read pairs (OR = 1.3, p-value = 0.013, Chi-square test) (Figure 1.2b, c, g). Thus, deleting a fraction of a TAD boundary reduced its insulation to cross-over RNA-DNA contacts.

Second, we previously created an insertion cell line (KI) using piggyBac transposon-mediated genomic insertion of this HERV-H sequence and identified seven insertion sites in KI42

(Columns, Figure 1.2d). Four of the seven insertion sites exhibited small increases in insulation (unlikely-de_novo-boundary sites), as measured by the difference in directionality index (delta_DI < 20) (Columns 4-7, Figure 1.2d), whereas the other three insertion site exhibited large increases in insulation (delta_DI > 20, likely-de_novo-boundary sites) (Zhang et al., 2019) (Columns 1-3, Figure 1.2d). Only one insertion site, that has the largest increase in insulation (delta_DI = 66.3), reached the significance level to be detected as a de novo TAD boundary, i.e., a boundary called in the KI Hi-C but not called in the WT Hi-C (de_novo-boundary site) (Column 1, Figure 1.2d). We note the de_novo-boundary site is one of the three likely-de_novo-boundary sites.

To test any impact of any insertion site on RNA-DNA contacts, we carried out iMARGI in KI and WT cells. For every insertion site, our null hypothesis is that whether any RNA-DNA contact is a cross-over or a non-cross-over contact is independent of whether this RNA-DNA contact is detected in KI or WT. The three likely-de_novo-boundary sites (delta_DI > 20) all led to some degrees of decrease in the odds ratio (OR) of the cross-over RNA-DNA contacts in KI (OR < 0.90), in which the decreases on two of the three likely-de_novo-boundary sites were significant (p-value < 1.0e-4, Chi-square test, stars in Figure 1.2d). In particular, the de_novo-boundary site exhibited a significant decrease in the OR of the cross-over RNA-DNA contacts in KI (OR = 0.75, p-value = 3.7e-5, Chi-square test, Figure 1.2e-g). Thus, the de novo creation of a TAD boundary suppressed cross-over RNA-DNA contacts.

In contrast, none of the four unlikely-de_novo-boundary sites (delta_DI < 20) led to a detectable decrease in the OR of the cross-over RNA-DNA contacts in KI (OR > 0.99, p-value > 0.63, Chi-square test, Figure 1.2d). Thus, inserting the same DNA sequence without sufficient subsequent changes in TAD structure did not suppress the cross-over RNA-DNA contacts. Taken together, these data show an impact of the 3D genome structure to the distribution of caRNA.

Induction of transcription locally suppresses chromatin looping

Our next question is whether RNA has an impact on the 3D structure of the genome. We approached this question in three steps. First, we depleted RPB1, the largest subunit of RNA Polymerase II (RNAP II), in HCT116 cells. RPB1 depletion resulted in increases in loop number and strengths as measured by Hi-C (Extended Text: RPB1 depletion, Supplementary Information). This result is consistent with the recent report in another cell line (DLD-1), where depletion of RPB1 led to the emergence of chromatin loops (Shu Zhang et al., 2021). These data implicate the transcriptional machinery, especially the presence of RNAPII on chromatin in suppressing chromatin looping.

Second, we tested if the induction of the transcription of a specific gene suppresses chromatin looping. To this end, we leveraged that there is a ~55 kb loop straddling the AAVS1 locus (AAVS1 loop), and nearby, there is a non-overlapping loop with a similar size to the AAVS1 loop (Krietenstein et al., 2020) (Nearby Ctrl loop, Figure 1.3a). We applied doxycycline to engineered H1 ES cells with a dCas9-KRAB knock-in transgene at the AAVS1 locus (Genga et al., 2019) to induce transcription of the transgene at the AAVS1 locus (Dox+), and subsequently tested the loops with chromosome conformation capture (3C) (Dekker et al., 2002). Compared to without doxycycline (Dox-), Dox+ weakened the AAVS1 loop (star, Figure 1.3b) but had little impact on the Nearby Ctrl loop (Figure 1.3b). Thus, inducing the expression of a gene can suppress a chromatin loop that straddle across this gene. Taken together, the transcription of a gene can locally suppress chromatin looping. It remains unclear whether it is the transcriptional machinery, the process of transcription, or the product of transcription, i.e., RNA, that affects chromatin looping.

RNA has a genome-wide impact on chromatin looping

We asked whether it is the transcription (including the association of the transcriptional machinery on chromatin and the process of transcription) or the RNA that impacts chromatin looping. We recognized that we could not answer this question by only testing with a specific genomic locus. This is because the answer at one genomic locus cannot necessarily rule out the alternative answer in other genomic regions. Thus, we recognized that the perhaps more important question is whether transcription or RNA has a genome-wide impact on chromatin looping. Furthermore, we recognized that if RNA impacts chromatin looping, then the transcriptional process must be implicated. However, if the transcriptional process is the cause, the causal chain does not necessarily involve RNA. With these considerations, we revised our question to whether RNA can impact chromatin looping genome-wide? To answer this question, we compared chromatin looping in control, transcription-inhibited (Chao & Price, 2001; Engreitz et al., 2014; Rahl et al., 2010), and RNase-treated cells (Barutcu et al., 2019). If the primary cause is the transcription process, we expect to see a widespread impact in transcription inhibition but not in acute RNase treatment. However, if transcription inhibition and acute RNase treatment lead to overlapping changes in chromatin loops genome-wide, the data would suggest RNA is involved in modulating chromatin looping. Additionally, we included another experimental condition where electrostatic molecular interactions are inhibited to test if any observed impacts are attributable to charge-driven condensates or phase separation (Jain & Vale, 2017; Saha & Hyman, 2017).

In our third step, we subjected H1 cells with ammonium acetate (NH4OAc) to disrupt electrostatic molecular interactions (the interactions due to electric charges) (Aumiller & Keating, 2016; Jain & Vale, 2017; Saha & Hyman, 2017), flavopiridol (FL) to suppress transcription elongation without displacing RNAPII from chromatin (Brahma & Henikoff, 2023; Chao & Price, 2001; Engreitz et al., 2014; Rahl et al., 2010), and acute RNase treatment to reduce RNA in the

nuclei (10-minute RNase treatment before fixing the cells) (Barutcu et al., 2019) based on established protocols (NH4Oac (Jain & Vale, 2017), FL (Engreitz et al., 2014), RNase A (Barutcu et al., 2019)). NH4OAc disrupts molecular electrostatic interactions in living cells by providing monovalent cations without perturbing intracellular pH (Saha & Hyman, 2017). To check the expected effects of the three treatments, we immunostained nuclear speckle-associated proteins SON (Ilik et al., 2020) and SC35 (Fu & Maniatis, 1990) in control and each treatment. NH4OAc reduced the numbers of SON and SC35's foci (p-value = 0.001 for SON, 0.009 for SC35, Wilcoxon test) (Supplementary Figure 1-2a, b, e, f, i, l), consistent with the role of RNA's electrostatic interactions in maintaining nuclear speckles (Smith et al., 2020). Conversely, FL made SON and SC35 foci larger and more distinct (J. Kim et al., 2019) (Supplementary Figure 2c, g, j, m). RNase A increased the numbers of SON and SC35's foci (p-value = 0.034 for SON, 0.010 for SC35, Wilcoxon test) (Supplementary Figure 2d, h, k, n), consistent with the observations that "low RNA/protein ratios promote phase separation into liquid droplets" (Sigova et al., 2015) and condensate formation (Henninger et al., 2021).

We generated Hi-C after each treatment in duplicates (Table S1) and analyzed these data together with those of the unperturbed H1 cells (control). We called chromatin loops from our Hi-C data in each of the four conditions that have comparable sequencing depths (Table S1) using HiCCUPS (Durand, Shamim, et al., 2016). The loop numbers were similar in control (2,473 loops) and NH4OAc (2,437 loops) (p-value = 0.55, paired t test) and were increased in FL (5,039 loops) (p-value < 1.1e-8, paired t test) and RNase (4,963 loops) (p-value < 2.3e-9, paired t test) (Figure 1.3c). These loop number differences cannot be attributed to different sequencing depths or batch effects because the samples were prepared in the same batch and sequenced to comparable depths (600 – 650 million read pairs per condition, Table S1b). Most of the emerged loops in FL

colocalized with the emerged loops in RNase (first column, Figure 1.3c). For example, a loop linking ATF7 and KRT18 genes that was absent in control and NH4OAc emerged in both FL and RNase (arrows, Figure 1.3d, Supplementary Figure 3).

The overall loop strength was similar in control and NH4OAc, but stronger in FL and RNase, as reflected by both Peak to Lower Left (P2LL) (Figure 1.3e) and Z-score Lower Left (ZscoreLL) scores (Rao et al., 2014) (Figure 1.3f). We repeated these analyses based on the union of the loops in the four conditions and quantified every loop's strength by Peak to Mean (P2M) in each condition. P2Ms were greater in FL and RNase than in control (p-value < 2.2e-16, Wilcoxon test), whereas NH4OAc's P2Ms were not different from the control's (p-value = 0.41, Wilcoxon test) (Figure 1.3g). The consistent increases of loop strengths in FL and RNase as compared to control support our detected increases of loop numbers in FL and RNase and suggest that our conclusion of loop number increases does not depend on the threshold of loop calls. Taken together, FL and RNase both resulted in an increase of chromatin loops and these emerged loops often co-localize. As opposed to the null hypothesis that RNA does not have a genome-wide impact on chromatin looping, these data are in favor of a suppressive effect of RNA to chromatin looping genome-wide.

RNA's genomic target regions correlate with the suppressed chromatin loops

Our next question is what RNA has an impact on which chromatin loops. Although we cannot analyze every aspect of a select RNA, we can analyze the chromatin-associated fraction of this RNA, in terms of this RNA's genomic target regions (target region) and the RNA attachment level (RAL) of this RNA on any target region. We can compare the target region with the genomic location of any chromatin loop. Thus, we asked whether the change of any caRNA, in terms of changes of target regions or RAL, correlates with the change of any chromatin loop. Answering

this question can inform us which RNA could impact which chromatin loop, although we will miss those impacts that are independent of RNA-chromatin association.

We generated iMARGI in each treatment condition (NH4OAc, FL, RNase) in duplicates (Table S1) and analyzed these data together with those of the unperturbed H1 cells (control). As expected, FL exhibited the largest reduction of the heights of the rectangular blocks in iMARGI's contact matrix (p-value < 3e-104, Wilcoxon rank-sum test) (Supplementary Figure 4e), consistent with FL's inhibitory effect on transcription elongation48. RNase exhibited the largest reduction of caRNA domains' number (3,217 in control and 357 in RNase, p-value < 3e-9, paired t test) (Supplementary Figure 4d) and sizes (widths of the rectangular blocks) (p-value < 5e-210, Wilcoxon rank-sum test) (Supplementary Figure 4f).

We analyzed two groups of caRNA, namely those associated with loop anchors (anchor caRNA), and those between loop anchors (between-anchor caRNA). We asked if the changes in the level of between-anchor caRNA correlates with the changes of chromatin loops across our treatment conditions. To answer this question, we analyzed the union of the loops (Union loops) detected in every condition (Control, NH4OAc, FL, RNase). These Union loops represent all possible loop locations, including those detected as loops in the Control (control loop) or in an RNA perturbation experiment (emergent loop). We used the ratio of between-anchor caRNA and anchor caRNA levels (Inside-loop To Anchor ratio (ITA ratio)) to represent the relative level of between-anchor caRNA for any Union loop.

First, we tested whether the detected loops in control (control loops) tend to locate at the genomic locations with a low level of between-anchor caRNA in the control. We carried out this test using Gene Set Enrichment Analysis (GSEA) (Subramanian et al., 2005). According to GSEA's procedure, we sorted the Union Loops by increasing levels of between-anchor caRNA,

i.e., increasing ITA ratios, creating a ranked list (Figure 1.4a). We then plotted the corresponding GSEA score at every rank (Figure 1.4b), where a positive/negative GSEA score indicates an enrichment/depletion of the control loops in the subset of top-ranked Union Loops. Here, top-ranked means from rank #1 to the current rank on which GSEA score is reported. The GSEA scores stayed positive in the top portion (~30%) of this rank list (Figure 1.4b), suggesting that the control loops are enriched in those Union loops that exhibit lower levels of between-anchor caRNA in the control condition than in the other conditions. In other words, among all the locations where loops have been detected, the control loops tend to appear at those locations where the relative level of between-anchor caRNA is low.

Second, RNase reduced caRNA levels in all Union loops (Figure 1.4c) and nearly doubled the number of detected loops as compared to Control (Figure 1.3c). We tested whether the loops in RNase (RNase loops) appeared at the locations where the between-anchor caRNA is most rigorously depleted in RNase. To this end, we re-ordered the Union loops by increasing levels of between-anchor caRNA, i.e., the ITA ratio in RNase (Figure 1.4c). As expected, all the GSEA scores are positive in this analysis (Figure 1.4d), which is because the RNase loops comprise the majority (~75%) of the Union loops, and therefore a majority in any top ranked subset. The GSEA scores in this rank list of Union loops first increased and then decreased, which means that the RNase loops are enriched in the higher ranked subset, which are the Union loops with low levels of between-anchor caRNA in RNase. This enrichment means that the emerged loops in RNase often appeared at the genomic locations where the between-anchor caRNA is most rigorously removed by RNase. Taken together, we observed a genome-wide negative correlation between between-anchor caRNA and the chromatin loops that stride across the target region of these caRNA.

Reducing select RNA creates specific chromatin loops

We wondered if we could apply the aforementioned correlation to identify which RNA has an impact on what chromatin loops. To this end, we tested whether we could create a particular chromatin loop by reducing a specific RNA, which is the RNA that exhibits a strong level of chromatin attachment to the genomic region between the anchors of this chromatin loop. We chose the ZMYND8 RNA for this test. We chose the ZMYND8 RNA because (1) FL reduced the RAL of the ZMYND8 RNA in an approximately 90 kb genomic region (Figure 1.4e); (2) RNase also removed the ZMYND8 caRNA in this (and a larger) genomic region; (3) a chromatin loop straddling across this 90 kb region was detected by Hi-C in both FL and RNase (arrows, Figure 1.4e), hereafter called the "straddling loop". We note that FL and RNase reduce the RALs of many RNAs, and thus from these data we cannot conclude that the emergence of this straddling loop in FL and RNase is due to the reduction of any specific RNA.

RNA knockdown without affecting transcription can reduce nucleoplasmic RNA and suppress long-range RNA-chromatin interactions, however, it cannot effectively remove nascent RNA that are associated with the chromatin near the transcription locus (Calandrelli et al., 2020). Therefore, we do not have a method to effectively remove ZMYND8 caRNA near the ZMYND8 gene without affecting the transcription of the ZMYND8 gene. We employed two approaches to address this issue. First, we asked whether suppression of ZMYND8 transcription has the same effect as RNase in creating the "straddling loop". Second, we will describe in subsequent sections the analysis of inter-chromosomal RNA-chromatin interactions, where we can better distinguish between impacts of the RNA from the transcriptional process.

We suppressed ZMYND8 by CRISPR interference (CRISPRi) in an H1 ES cell line with doxycycline-inducible dCas9-KRAB (Genga et al., 2019; González et al., 2014). Compared to scrambled gRNA control, our gRNA targeting ZMYND8's promoter reduced ZMYND8's

transcription level to approximately 25% (Figure 1.4f). We designed chromosome conformation capture (3C) primers[46] for (1) a negative control "loop" (Negative Ctrl) that is located 200 kb upstream of the emerged loop and has approximately the same size as the emerged loop, which is not detected as a loop in any Hi-C experiment, (2) a positive control loop (Positive Ctrl) detected by Hi-C in both control and FL, which is not on the same chromosome as ZMYND8, and (3) the straddling loop (also termed the "to-be-tested loop"). We carried out 3C after treating the cells with doxycycline without supplying gRNA (gRNA:None Ctrl), supplying with a scrambled gRNA (gRNA:Scramble Ctrl), and with gRNA targeting ZMYND8's promoter (gRNA:ZMYND8). The Negative Ctrl primers did not yield any product in any experiment (the first 3 lanes), and the Positive Ctrl primers yielded products at the expected sizes in all three experiments (the last 3 lanes, Figure 1.4g). In contrast, the primers for the to-be-tested loop yielded a unique product with ZMYND8 gRNA (arrow, Figure 1.4g), which is absent from the gRNA:None and gRNA:Scramble controls. In summary, acute reduction of RNA induced many chromatin loops including the straddling loop, and suppression of the ZMYND8 expression can re-create the emergence of the straddling loop. Thus, the negative correlation of between-anchor caRNA and chromatin loops can help to identify which RNA has an impact on which chromatin loop. We note that the CRISPRi experiment by itself cannot distinguish whether the loop was created by suppression of transcription or reduction of ZMYND8 RNA. This CRISPRi experiment demonstrates that a loop created by acute depletion of RNA (RNase) can be re-created by suppression of the expression of a specific gene.

Acute RNA reduction increases the average strength of the loops with convergent CTCF binding sites in their loop anchors

Our next question is whether an RNA can impact the chromatin loops located far from the transcription locus of this RNA (distal loops). We recognize that our previously mentioned

correlation is not sufficient to connect a specific RNA to specific distal loops. This is because an RNA can associate with many distal genomic regions, often at low levels. Thus, we proceeded to identify additional correlational rule(s) to between RNA and chromatin loops.

Convergent CTCF binding sites (CBS) in the loop anchors is a characteristic of the loops created by loop extrusion (Henninger et al., 2021) . We tested whether the convergent CBS are enriched in the anchors of the loops with increased loop strengths in RNase. To this end, we categorized the Union Loops (the union of the loops detected in any treatment condition) into three groups based on the orientations of the CTCF binding sites at their anchors, namely the loops with convergent CBS, non-convergent CBS, or no CBS. We used Peak to Lower Left (P2LL) to quantify the strength of each loop (Rao et al., 2014) . Compared to control, RNase treatment increased P2LL in the Union Loops with convergent CBS (p-value < 1.6E-9, Wilcoxon test, Figure 5a). In comparison, RNase did not increase P2LL in the Union Loops with non-convergent CBS (p-value = 0.4663, Wilcoxon test, Figure 5a) or in the loops without CBS (p-value = 0.6277, Wilcoxon test, Figure 5a). Thus, acute RNA reduction increased the average strength of those loops with convergent CBS in their loop anchors, suggesting an enrichment of convergent CBS in anchors of RNA suppressed loops. In summary, we have observed two genome-wide correlations, which are (1) a negative correlation of between-anchor caRNA and chromatin loops and (2) an enrichment of convergent CBS in RNA suppressed loops. Hereafter we call these correlations the "correlational rules".

Removal of select RNA increases the strengths of a subset of distal chromatin loops

We wondered if we could apply the correlational rules to identify which RNA may have an impact on what distal chromatin loops. To this end, we tested whether removing specific RNA can increase the strengths of certain distal loops. We chose the HERV-H RNA for this test for the following reasons. First, we identified the HERV-H caRNA-associated genomic sequences

(HERV-RNA target regions) in Control and compared them with the locations of the loops emerged in RNase (RNase emergent loops). The RNase emergent loops are enriched at the locations that exhibit between-anchor HERV-H caRNA in Control (odds ratio = 1.38, p-value = 5.621e-5, Chi-square test, Figure 5b), suggesting that those loops that stride across between-anchor HERV-H caRNA are suppressed in Control. Second, we analyzed the subset of RNase emergent loops that stride across HERV-H caRNA-attached genomic sequences in Control. Hereafter, we call this subset of RNase emergent loops as "candidate HERV-H caRNA insulated loops" (CHRI-loops). CHRI-loops are enriched with convergent CBS in their loop anchors as compared to (1) control loops striding across HERV-H caRNA-attached genomic sequences (OR = 1.34, p = 0.0068, Chi-square test), and to (2) the other control loops not striding across HERV-H caRNA-attached genomic sequences (OR = 1.44, p = 5.8e-6, Chi-square test), and to (3) the RNase emergent loops not striding across any HERV-H caRNA-attached genomic sequences in Control (OR = 1.60, p = 4.1e-9, Chi-square test, Figure 5c). Thus, convergent CBS are enriched in the loop anchors of CHRI-loops.

We tested whether deleting an HERV-H element from the human genome can lead to increase the loop strength of any distal CHRI-loop. To this end, we re-used our Chr13:55.5MB_HERV KO (CRISPR-mediated deletion of a HERV-H element at Chr13:55,578,227-55,584,087) human ES cells (Zhang et al., 2019). We identified the caRNA transcribed from this Chr13:55.5MB_HERV element and its target genomic sequences (Chr13:55.5MB_HERV targets) in the WT. We call the loops that stride across any Chr13:55.5MB_HERV targets as "target-crossing loops". We compared the loop strength changes of all target-crossing loops between Chr13:55.5MB_HERV KO and WT based on Hi-C data. No target-crossing loop exhibited detectable decrease in loop strength in Chr13:55.5MB_HERV KO,

whereas two target-crossing loops exhibited increased loop strengths in the Chr13:55.5MB_HERV KO (Figure 5d). Both Chr13:55.5MB_HERV KO-induced target-crossing loops contain convergent CBS in their loop anchors (Figure 5d). Neither Chr13:55.5MB_HERV KO-induced target-crossing loop locates on Chromosome 13, where the HERV-H element is deleted (Figure 5d). Thus, removal of specific RNA increased the loop strengths of a subset of chromatin loops that stride across this RNA's interchromosomal target regions and contain convergent CBS in their loop anchors. These data suggest specific RNA can modulate a subset of chromatin loops. Furthermore, the correlational rules help to identify which RNA modulates what chromatin loops.

Discussion

We presented a resource composed of genome-wide RNA-DNA and DNA-DNA contact maps in three human cell lines. The iMARGI and Hi-C experimental protocols and data processing pipelines used for generating this resource were proven by the 4D Nucleome (4DN) Consortium Omics Standards Working Group and the 4DN Steering Committee (https://www.4dnucleome.org/protocols/). The three human cell lines for data generation were nominated by the 4DN Joint Analysis Working Group and cultured under the 4DN Cell Working Group approved protocols (https://www.4dnucleome.org/cell-lines/). All the data are accessible through the 4DN Data Portal (see Data Availability and Table S1).

The initial challenges to caRNA as a distinct class of RNA were focused on whether these RNAs are exclusively nascent transcripts (Holmes et al., 1972). Such a concern was alleviated by the discoveries of long-range RNA-chromatin interactions (Hall and Lawrence, 2016; Nozawa et al., 2017; Nozawa and Gilbert, 2019; Pederson, 2019; Rodríguez-Campos and Azorín, 2007), suggesting that caRNA does not completely overlap with nascent transcripts. Our genome-wide analyses reveal two features of RNA-genome association. First, RNA is preferentially associated

with its transcription site and up to several megabases of flanking genomic sequence. Second, TAD boundaries insulate RNA-DNA contacts, evidencing the impact of 3D genome on the spatial distribution of caRNA.

It remains unclear how RNA may affect the 3D genome. Because several 3D features of the genome can be reproduced by computational models without considering RNA (Liang et al., 2023; Zufferey et al., 2018) and in vitro experiments to recapitulate loop extrusion without RNA, RNA (Wen & Zhong, 2023) was not expected to affect the genome's 3D organization. Furthermore, previous work found that acute reduction of RNA had subtle impacts to the 3D genome at the compartment and the TAD levels (Barutcu et al., 2019). Our analyses led to similar findings. At the compartment level, Hi-C's PC1 in FL and RNase exhibited strong correlations with Hi-C's PC1 in control, suggesting these perturbations had little impact to A/B compartments. At the TAD level, FL and RNase exhibited "highly concordant" TADs with the control, based on the "Measure of Concordance (MoC)" (Zufferey et al., 2018) (pairwise MoCs = 0.93 and 0.90, well above 0.75, the threshold for being "highly concordant" (Zufferey et al., 2018)). These data confirm that the impacts of RNA to the 3D genome are subtle, at least at the scales of A/B compartments and TADs.

It has not been tested whether an acute reduction of RNA exerts systematic impacts to chromatin loops. Our data reveal either transcription inhibition or acute RNA reduction induced chromatin loops. Most induced loops are shared between transcription inhibition and acute RNA reduction, indicating that the impact on chromatin looping cannot be completely attributed to transcription or the presence of RNAPII on chromatin. Indeed, suppressing a specific caRNA created a chromatin loop, with the loop anchors striding across the genomic sequence associated with this caRNA (Figure 1.4e-g). Furthermore, deleting the genomic sequence of a caRNA

(Chr13:55.5MB_HERV) strengthened the chromatin loops on other chromosomes (Figure 5d). These inter-chromosomal effects argue against that loop strengths are modulated by the transcription of the deleted sequence. They support the idea that the caRNA at specific locations, i.e. between-anchor caRNA, suppresses chromatin looping. Of note, these experiments were not meant to establish an exclusive role of RNA in modulating chromatin looping. While these data establish RNA's role, they do not exclude transcription or RNAPII's role in modulating chromatin looping.

What remains to be addressed is whether there is any rule that links specific RNA with specific loops that this RNA can modulate. Disrupting electrostatic interactions by NH4OAc did not lead to significant changes in chromatin loops, withholding us from exploring possible rules based on charge-mediated condensates or phase separation. Instead, we investigated RNA's target regions, because the genomic locations of the target regions can be compared with the genomic locations of loops. We found the caRNA located between the two anchors of a chromatin loop often weakens this loop. This suppressive role of between-anchor caRNAs is compatible with the recently reported promotive role of the caRNAs with complementary sequences to chromatin looping (Liang et al., 2023; Wen & Zhong, 2023). When two caRNAs, that are attached to two genomic sequences, have complementary sequences, they tend to bring their associated genomic sequences to spatial proximity, thus promoting loop formation (Liang et al., 2023). Following the same idea, any sequence complementarity between any between-anchor caRNA and any other caRNA associated with any outside-of-loop genomic sequences could promote spatial proximity of this between-anchor genomic sequence and those outside-of-loop genomic sequences, thus reducing the spatial proximity of the loop anchors. Thus, a unified model can explain the

suppressive role of between-anchor caRNA and the promotive role of the caRNAs with sequence complementarity on chromatin looping.

Recent work revealed a suppressive role of RNAPII's presence on chromatin to chromatin looping (Shu Zhang et al., 2021). Without underestimating RNAPII's role, our experiments were designed to test if there are any effects of the RNA as well. FL treatment does not displace RNAPII from chromatin (Brahma & Henikoff, 2023), making the FL emergent loops unlikely due to a change of RNAPII's presence on chromatin. Furthermore, our analysis focused on the shared loops that are created by both FL or RNase treatment. These shared loops are even more unlikely attributable to the loading of RNAPII on chromatin. Our data suggest that in addition to RNAPII, RNA should be considered toward obtaining a complete picture on the interplay between transcription and genome organization.

Figure

Figure 1.1: Localized RNA-genome association.
a) iMARGI's contact matrix between the RNA of the JARID2 gene (rows, bin size = 10 kb) to the genomic sequence of chromosome 6 (columns, bin size = 500 kb). The RNA association level (RAL) of JARID2 RNA (RAL track), the truncated version of RAL showing the small values (trunc RAL track), and the cumulative RAL of all RNA (cRAL track) exhibit a correlation with the first principal component of Hi-C's contact matrix (PC1 track). (b) iMARGI's contact matrix, RAL, and truncated RAL of the PVT1 RNA on chromosome 6. (c) An RNA-DNA contact matrix in a 2M bp sequence on Chromosome 6. Each entry in this contact matrix represents the number of iMARGI read pairs with the RNA-end mapped to the corresponding row and the DNA-end mapped to the corresponding column. The box marks an identified RNA-association domain, an approximately 1 Mb region containing the ATXN1 gene. (d) Upset plot of the numbers of the detected RNA-association domains in H1, HFF, and K562. (e) Box plots of RNA-association domains' sizes (blue) corresponding to the widths of the detected rectangular blocks, and the lengths of caRNA-producing genomic sequences (red) for each RNA-association domain corresponding to the heights of the detected rectangular blocks in iMARGI's contact matrix (n = 3,217 RNA-association domains). The center line of the boxplots is the median. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. Source data are provided as a Source Data file.

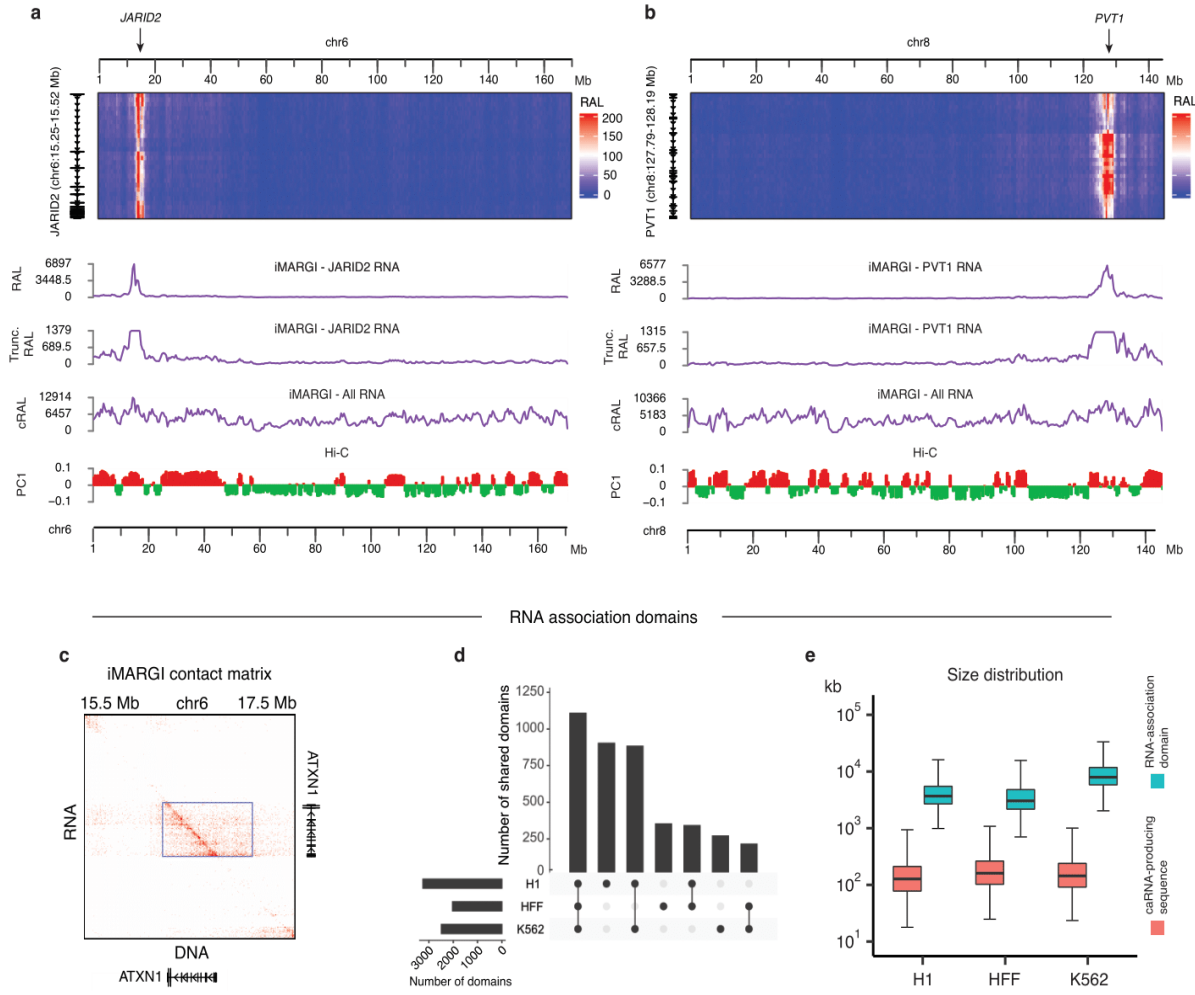

a
JARID2
chr6
JARID2 (chr6:15.25-15.52 Mb)
RAL
iMARGI - JARID2 RNA
Trunc. RAL
cRAL
iMARGI - All RNA
Hi-C
PC1
b
PVT1
chr8
PVT1 (chr8:127.79-128.19 Mb)
iMARGI - PVT1 RNA
RNA association domains
c
iMARGI contact matrix
15.5 Mb
17.5 Mb
RNA
DNA
ATXN1
d
Number of shared domains
H1
HFF
K562
Number of domains
e
Size distribution
kb
RNA-association domain
caRNA-producing sequence

Figure 1.2: Boundaries suppress cross-over RNA-DNA contacts.
(a) The RNA association level (RAL, color-coded) of the RNA transcribed within each TAD (row) on this TAD (center block) and its equal-length flanking regions (x axis). Curve at the bottom: the average RAL of all TADs (rows) with the 95% confidence interval (band). (b) Comparison of normalized RNA-DNA contact matrices in WT and KO cell lines. The arrowhead points to the HERV-H element in WT that is deleted in KO. KO-WT: The contrast of the KO and WT contact matrices where red indicates an increase of RNA-DNA contacts in KO. The increased RNA-DNA contacts in KO are enriched with cross-over contacts (in the box at the upper right corner). (c) The 2x2 contingency table for an association test based on the data in panel b. (d) The seven previously identified insertion sites (columns) are ranked by delta_DI, where a smaller delta_DI (on the right) indicates a smaller increase in the ability to insulate cross-over DNA-DNA contacts (a weaker putative boundary). The rows mark whether each insertion site is a "de_novo boundary site" (Column 1), a "likely-de_novo-boundary site" (Columns 1-3), or an "unlikely-de_novo-boundary site" (Columns 4-7), based on the comparison of Hi-C data in KI and WT. A Chi-square test is performed on each insertion site (column) based on the iMARGI data in KI and WT. A smaller p-value (y axis) represents stronger evidence against the null hypothesis that there is no association between the RNA-DNA cross-over contacts and KI. *: p-value = 3.724e-5. (d) (e) Comparison of normalized RNA-DNA contact matrices in WT and KI cell lines at the de_novo boundary site (arrowhead). KI-WT: The contrast of the KI and WT contact matrices where red indicates an increase of RNA-DNA contacts in KI. The increased RNA-DNA contacts in KI are enriched in the non-crossover contacts (in the boxes at the upper left and lower right corners). (f) The 2x2 contingency table for an association test based on the data in panel e. (g) Significance levels for the deletion site (panel b, c) and the de_novo-boundary insertion site (panel e, f). *: p-value = 0.013, **: p-value = 3.724e-5, Chi-square test. Data are presented as odds ratios, with the error bar whiskers at exp(log(OR) -/+ SELOR), where SELOR is the standard error of the log odds ratio (n = 3,761 and 6,927 RNA-DNA contacts at the deletion and insertion site respectively). Source data are provided as a Source Data file.

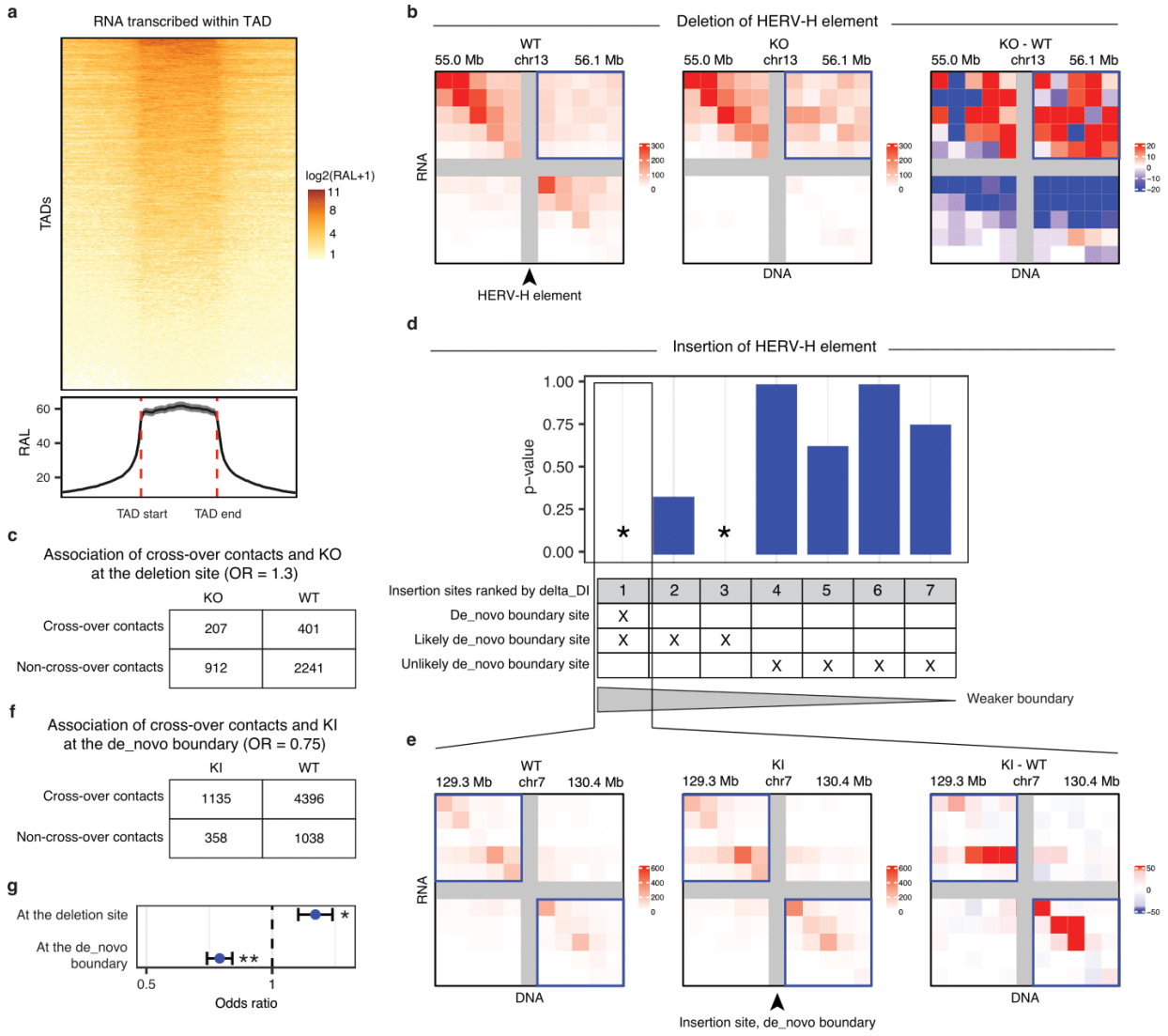

a
RNA transcribed within TAD
TADs
log2(RAL+1)
11
8
4
1
RAL
60
40
20
TAD start
TAD end
b
Deletion of HERV-H element
WT
KO
KO - WT
55.0 Mb
chr13
56.1 Mb
RNA
DNA
HERV-H element
c
Association of cross-over contacts and KO at the deletion site (OR = 1.3)
KO
WT
Cross-over contacts
207
401
Non-cross-over contacts
912
2241
d
Insertion of HERV-H element
p−value
1.00
0.75
0.50
0.25
0.00
Insertion sites ranked by delta_DI
1
2
3
4
5
6
7
De_novo boundary site
Likely de_novo boundary site
Unlikely de_novo boundary site
Weaker boundary
e
KI
KI - WT
129.3 Mb
chr7
130.4 Mb
Insertion site, de_novo boundary
f
Association of cross-over contacts and KI at the de_novo boundary (OR = 0.75)
KI
WT
Cross-over contacts
1135
4396
Non-cross-over contacts
358
1038
g
At the deletion site
At the de_novo boundary
0.5
1
Odds ratio

Figure 1.3: RNA-related loop changes.

(a) Transcription induction of a gene suppresses a loop straddling this gene. Genomic coordinates of the AAVS1 locus, the loop straddling the AAVS1 locus (AAVS1 loop), and a nearby loop with a similar size (Nearby Ctrl loop). (b) 3C products without doxycycline (Dox: -) and with transcription induction by doxycycline (Dox: +), based on primers against the AAVS1 loop (Lanes 3, 4), the Nearby Ctrl loop (Lanes 5, 6), and a size-matched control region without any Hi-C detected loop (Negative Ctrl, Lanes 7, 8). *: Difference in 3C products between Dox- and Dox+. Lane 1: E-Gel™ 1 kb DNA Ladder. Lanes 2 and 9: E-Gel™ 50 bp DNA Ladder. Each experiment was repeated independently 3 times. (c) Upset plot of the loop numbers in the four conditions, control, NH4OAc, FL, and RNase (rows). (d) An example of loop changes. Hi-C contact matrix of every replicate (row). Arrows: a shared loop in FL and RNase that is absent in control and NH4Oac. (e-g) FL and RNase increase loop strengths. (e, f) Aggregate loop strength represented by P2LL (e) or ZscoreLL (f) (y axis) in each condition (column). Color bars: the loops detected in each condition (red) or their union (blue). (g) Box plots of the strengths of individual loops (P2M) in every condition (column). ****: p-value < 2.2e-16, Wilcoxon test (n = 6,783 loops). The center line of the boxplots is the median. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. Source data are provided as a Source Data file.

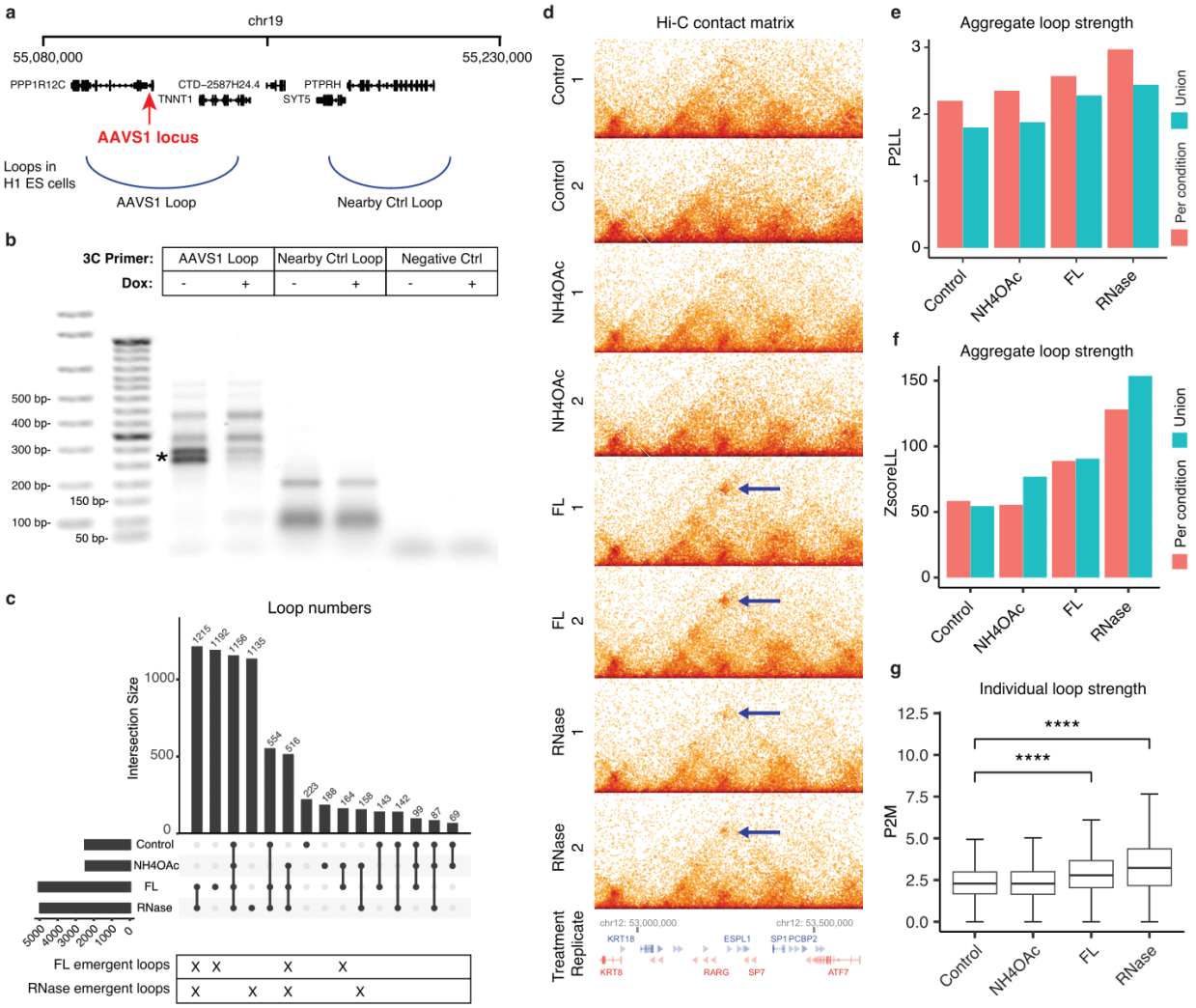

a
chr19
55,080,000
55,230,000
PPP1R12C
CTD−2587H24.4
PTPRH
TNNT1
SYT5
AAVS1 locus
Loops in H1 ES cells
AAVS1 Loop
Nearby Ctrl Loop
b
3C Primer:
AAVS1 Loop
Nearby Ctrl Loop
Negative Ctrl
Dox:
- + - + - +
500 bp-
400 bp-
300 bp-
200 bp-
150 bp-
100 bp-
50 bp-
c
Loop numbers
Intersection Size
1215
1192
1156
1135
554
516
223
188
164
158
143
142
89
87
69
Control
NH4OAc
FL
RNase
FL emergent loops
RNase emergent loops
d
Hi-C contact matrix
Control 1
Control 2
NH4OAc 1
NH4OAc 2
FL 1
FL 2
RNase 1
RNase 2
Treatment Replicate
chr12: 53,000,000
chr12: 53,500,000
KRT18
KRT8
ESPL1
SP1
PCBP2
RARG
SP7
ATF7
e
Aggregate loop strength
P2LL
Union
Per condition
Control
NH4OAc
FL
RNase
f
Aggregate loop strength
ZscoreLL
Union
Per condition
Control
NH4OAc
FL
RNase
g
Individual loop strength
P2M
****
****
Control
NH4OAc
FL
RNase

Figure 1.4: Between-anchor caRNA anticorrelates with chromatin looping.

(a-b) The loops in the Control (control loops) are depleted with between-anchor caRNA. (a) The caRNA levels in the control (Control RAL) on loop anchors (two sides) and between the anchors (middle) is color-coded (blue: low, yellow: high) for every loop detected in any condition (Union loops, rows). Loops are ranked by the relative level of their between-anchor caRNA (Inside-loop To Anchor (ITA) ratio) from low (top) to high (bottom). (b) The enrichment/depletion level (GSEA score, x axis) of the control loops in the subset of loops from the top-ranked loop (first row) and the currently ranked loop (current row, y axis). A positive/negative GSEA score indicates an enrichment/depletion of the control loops in this subset of loops. The control loops are enriched in the top-ranked loops, i.e., those with low levels of between-anchor caRNA (blue bar on the right), and are depleted in the bottom-ranked loops, i.e., those with high levels of between-anchor caRNA (yellow bar). (c-d) RNase emergent loops are those with low levels of between-anchor caRNA. (c) The caRNA levels in the RNase (RNase RAL) on loop anchors (two sides) and between the anchors (middle) is color-coded (blue: low, yellow: high) for every loop detected in any condition (Union loops, rows). The union loops (rows) are ordered by the relative level of their between-anchor caRNA (the ITA ratio calculated in RNase) from low (top) to high (bottom). (d) The enrichment level (GSEA score, x axis) of the RNase loops (x axis) in the subset of loops from the top-ranked loop (first row) and the currently ranked loop (current row, y axis). The RNase-specific loops are enriched in the top-ranked loops, i.e., the loops with low levels of between-anchor caRNA in RNase, as indicated by the increasing GSEA scores (blue bar on the right). In contrast, the loops detected in other conditions are enriched in the bottom-ranked loops, i.e., the loops with high levels of between-anchor caRNA in RNase, as indicated by the decreasing GSEA scores (yellow bar). (e-g) Transcriptional suppression of the ZMYND8 induces a specific chromatin loop. (e) Changes in iMARGI RNA-DNA contact maps in Control (left panel), FL (central panel), and RNase (right panel). FL reduced the caRNA in the upstream region of the ZMYND8 gene (blue dox) and induced a chromatin loop near the caRNA-depleted region (curve at the bottom). RNase reduced the caRNA from a wider genomic region and induced the same chromatin loop as that in FL. CBS: CTCF binding site. Arrowheads point to CBSs' directions. Blue dots: Hi-C derived loops that are superimposed on this iMARGI contact map. Arrow: the emergent loop in FL. (f) Comparison of normalized ZMYND8's expression levels (y axis) in CRISPRi experiments with the scrambled gRNA (Scramble Ctrl) and ZMYND8-targeting gRNA (ZMYND8 gRNA). Data are presented as mean values +/- SEM (n = 3 replicates). (g) 3C products from the Negative Ctrl primers (the first 3 lanes), the primers for the To-be-tested loop (3 middle 3 lanes), and the Positive Ctrl primers (the last 3 lanes) in CRISPRi experiments without a gRNA (gRNA: None), with a scrambled gRNA control (gRNA: Scramble), or with the ZMYND8-targeting gRNA (gRNA: ZMYND8). The Negative Ctrl primers did not yield any product in any experiment (Lanes 3-5). The Positive Ctrl primers yielded products of the same sizes in all three experiments (Lanes 9-11). The primers for the to-be-tested loop yielded a product with ZMYND8-targeting gRNA (arrow) but not with a scrambled gRNA or without gRNA (Lanes 6-8), confirming that a loop is created by ZMYND8 CRISPRi. Lane 1: E-Gel™ 1 kb DNA Ladder. Lanes 2: E-Gel™ 50 bp DNA Ladder. Each experiment was repeated independently 3 times. Source data are provided as a Source Data file.

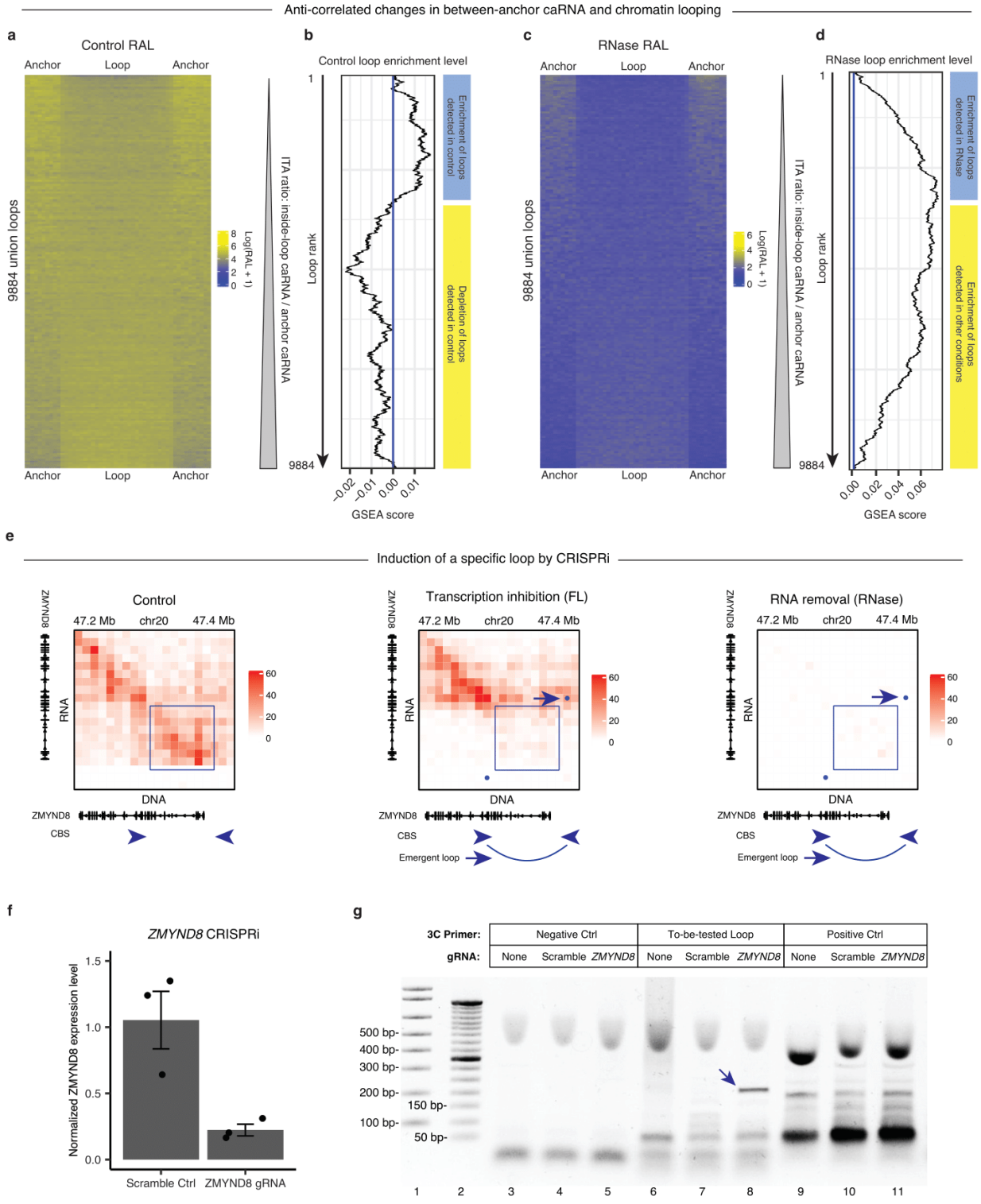
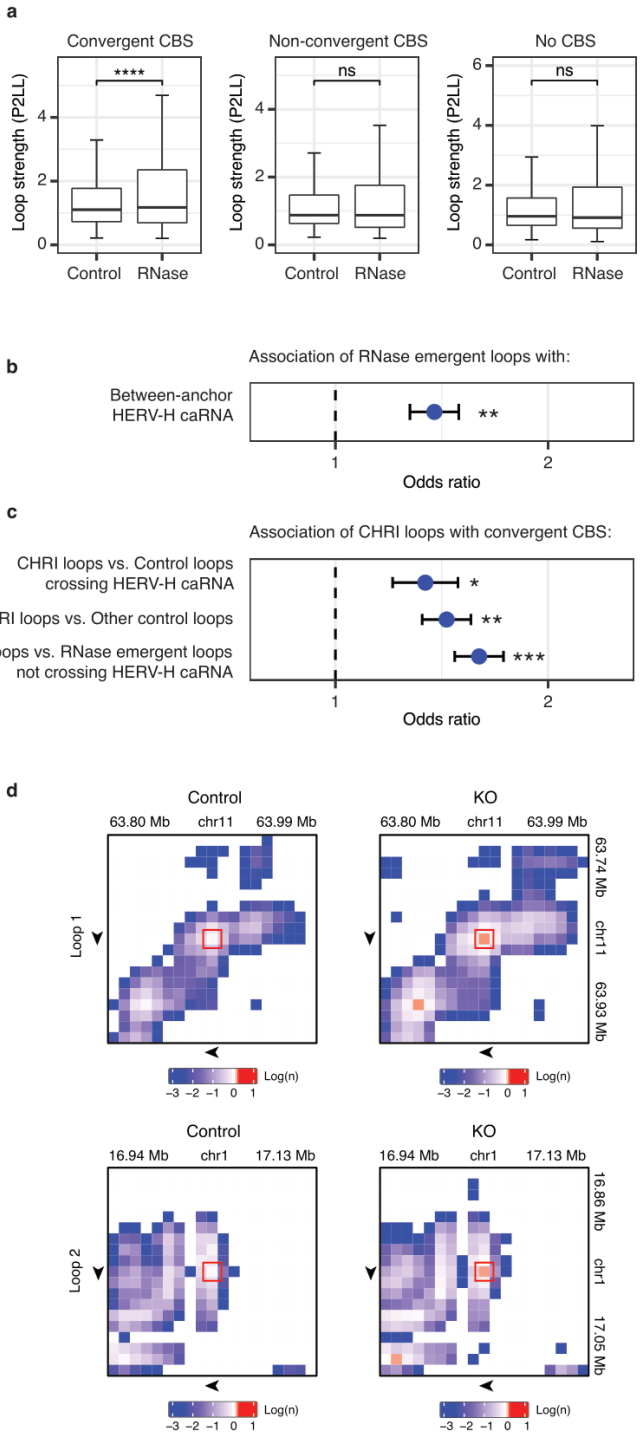

Anti-correlated changes in between-anchor caRNA and chromatin looping
a
Control RAL
Anchor
Loop
Anchor
9884 union loops
Log(RAL + 1)
ITA ratio: inside-loop caRNA / anchor caRNA
b
Control loop enrichment level
Loop rank
Enrichment of loops detected in control
Depletion of loops detected in control
GSEA score
c
RNase RAL
d
RNase loop enrichment level
Enrichment of loops detected in RNase
Enrichment of loops detected in other conditions
e
Induction of a specific loop by CRISPRi
Control
Transcription inhibition (FL)
RNA removal (RNase)
47.2 Mb
chr20
47.4 Mb
ZMYND8
RNA
DNA
CBS
Emergent loop
f
ZMYND8 CRISPRi
Normalized ZMYND8 expression level
Scramble Ctrl
ZMYND8 gRNA
g
3C Primer:
gRNA:
Negative Ctrl
To-be-tested Loop
Positive Ctrl
None
Scramble
ZMYND8
500 bp-
400 bp-
300 bp-
200 bp-
150 bp-
100 bp-
50 bp-

Figure 1.5: Enrichment of convergent CTCF binding sites (CBS) in the anchors of RNA-affected loops.
(a) Comparison of loop strengths (P2LL, y axis) in the loops with convergent CBS, non-convergent CBS, and without CBS in control and RNase (columns). ****: p-value = 1.571e-9. ns: not significant (p-values = 0.466 and 0.628 for Non-convergent CBS and No CBS respectively), Wilcoxon test. The center line of the boxplots is the median. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. (b) Enrichment of RNase emergent loops with between-anchor HERV-H caRNA in control (odds ratio, x axis). **: p-value = 5.621e-5, Chi-square test (n = 8,112 chromatin loops). Odds ratio > 1 means enrichment. Data are presented as odds ratios, with the error bar whiskers at exp(log(OR) -/+ SELOR), where SELOR is the standard error of the log odds ratio. (c) Enrichment of "candidate HERV-H caRNA insulated loops" (CHRI-loops) with convergent CBS in their loop anchors as compared to control loops striding across HERV-H caRNA-attached genomic sequences (first row, n = 1,653 chromatin loops), the other control loops not striding across HERV-H caRNA-attached genomic sequences (second row, n = 4,059 chromatin loops), and the RNase emergent loops not striding across any HERV-H caRNA-attached genomic sequences (third row, n = 4,219 chromatin loops). *: p = 6.884e-3, **: p = 5.857e-6, ***: p < 4.056e-9, Chi-square test. Data are presented as odds ratios, with the error bar whiskers at exp(log(OR) -/+ SELOR), where SELOR is the standard error of the log odds ratio. (d) The two target-crossing loops (rows) with increased Hi-C contacts in HERV-H KO (KO column) as compared to control (Control column). Hi-C data were denoised using the DeepLoop software. The denoised Hi-C contact maps were shown in the log scale. Arrow: direction of CTCF binding site in the loop anchor. Source data are provided as a Source Data file.


a
Convergent CBS
Non-convergent CBS
No CBS
Loop strength (P2LL)
****
ns
ns
Control
RNase
b
Association of RNase emergent loops with:
Between-anchor HERV-H caRNA
**
Odds ratio
c
Association of CHRI loops with convergent CBS:
CHRI loops vs. Control loops crossing HERV-H caRNA
CHRI loops vs. Other control loops
CHRI loops vs. RNase emergent loops not crossing HERV-H caRNA
*
**
***
Odds ratio
d
Control
KO
63.80 Mb
chr11
63.99 Mb
63.74 Mb
63.93 Mb
Loop 1
Log(n)
16.94 Mb
chr1
17.13 Mb
16.86 Mb
17.05 Mb
Loop 2

Methods

Cell culture and treatments

Human embryonic stem cells (H1), hTert-immortalized human foreskin fibroblasts (HFF), and chronic myelogenous leukemia lymphoblasts (K562) were obtained from the 4D Nucleome (4DN) Cell Repository and cultured following the 4DN Consortium's approved culture protocol for each cell line (https://www.4dnucleome.org/cell-lines.html). The cell lines in the 4DN Cell Repository were established by the 4DN Consortium in collaboration with WiCell and ATCC for providing quality-controlled cells from the identical batch to minimize cell source and culture condition variations. The cell culture protocols were developed by the 4DN Cell Line Working Group and approved by the 4DN Steering Committee.

Ammonium acetate treatment

H1 cells were treated with 0.1 M NH4OAc in complete mTeSR medium for 10 min as described in a previous study (Saha & Hyman, 2017). Briefly, a crystalline NH4OAc (Sigma-Aldrich, Cat# A1542-500G) was dissolved in nuclease-free water and further diluted in cell medium. Aspirate medium in each well and H1 cells were treated with 0.1 M NH4OAc in medium for 10 min at RT.

Flavopiridol treatment

H1 cells were treated with 1 μM flavopiridol in complete mTeSR medium for 1 h in an incubator as described previously (Rahl et al., 2010). Specifically, a crystalline flavopiridol (hydrochloride) (Cayman Chemical, item# 10009197) was dissolved in DMSO to prepare 1 mM flavopiridol (FL) stock solution. 1 mM FL stock solution was further diluted with complete mTeSR medium. Aspirate cell medium in each well and H1 cells were either treated with 1 μM FL in medium or an equivalent amount of DMSO in the medium in an incubator at 37 °C for 1 h.

RNase A treatment

H1 cells were harvested from cell culture plate and aliquoted cell suspension to 10 million H1 cells per 1.5 mL tube. Wash the cells with 1 mL 1X PBS and centrifuge at 500 X g for 3 min at RT. Then, cells were gently permeabilized by resuspending cell pallets with 0.01% PBST (TritonX-100 in PBS) and treated for 5 min at RT. After permeabilization, cells were treated with 200 µg/mL RNase A as described previously (Barutcu et al., 2019) (Thermo Fisher Scientific, Cat# EN0531) on rotator for 10 min at RT. The treated cells were fixed with 4% formaldehyde (Thermo Fisher Scientific, Cat# 28906) for immunofluorescence imaging. For Hi-C and iMARGI library generation, the treated cells were fixed with 1 mL 1% formaldehyde on rotator for 10 min at RT. Then, the reactions were terminated with 250 µL 1 M glycine on rotator for 10 min at RT. The treated sample was centrifuged at 2000 X g for 5 min at 4 °C and washed with 1 mL cold 1X PBS.

dCas9-KRAB inducible cells

The doxycycline-inducible dCas9-KRAB H1 ES cell line is generated and karyotyped by the 4D Nucleome Consortium (Danwei Huangfu Laboratory) (https://4dnucleome.org), with TRE-dCas9-KRAB and CAGGS-M2rtTA targeted into the AAVS1 locus.

HERV-H deletion and insertion cells

The control H9 human ES cells (H9 MLC2v:H2B), HERV-H deletion cell line (H9 MLC2v:H2B HERV2-KO), and HERV-H insertion cell line (H9 MLC2v:H2B HERV2-ins-clone2) were generated by Bing Ren lab and described in (Zhang et al., 2019).

RPB1 The auxin-inducible degron 2 cells

The RPB1 auxin-inducible degron 2 cells (HCT116 RPB1-Dox-OsTIR1-mClover-mAID) were generated by Masato Kanemaki lab and described in (Yesbolatova et al., 2020).

Calling de novo TAD boundaries

TAD boundaries in WT (H9 MLC2v:H2B) and KI (H9 MLC2v:H2B HERV2-ins-clone2) were separately called based on their respective Hi-C data using The Arrowhead tool in the Juicer

Tools (Durand, Shamim, et al., 2016) with default parameters. A TAD boundary called in KI but not in WT is regarded as a de novo TAD boundary.

Immunofluorescence imaging

The cells on coverslip (Fisher Scientific, Cat# 12-541 A) were fixed with 4% formaldehyde at RT for 30 min. The fixed cells were washed with 1X PBS once and permeabilized with 0.1% TritonX-100 in PBS (PBST) at RT for 15 min on shaker. Afterwards, cells were blocked with 5% BSA (VWR, Cat# 97061-420) in PBST at RT for 30 min with gentle shaking. For SC35 staining, H1 cells were incubated with 1 mL diluted mouse monoclonal anti-SC35 primary antibody (1:250) (Abcam, Cat# ab11826) in 5% BSA at 37 °C for 1 h, and subsequently washed three times with PBST on shaker for 10 min. Cells were further incubated with 1 mL diluted goat anti-mouse secondary antibody with Alexa Fluor 568 (1:500 dilution) (Invitrogen, Cat# A-11004) in 5% BSA at 37 °C for 30 min. For SON staining, the cells were incubated with 1 mL diluted rabbit anti-SON primary antibody (1:2000 dilution) (Atlas Antibodies, HPA023535) in 5% BSA at 37 °C for 1 h, and subsequently washed three times with PBST on shaker for 10 min. The cells were incubated with 1 mL diluted goat anti-rabbit secondary antibody with Alexa Fluor 488 (1:500 dilution) (Invitrogen, Cat# A-11008) in 5% BSA at 37 in 5% BSA at 37 °C for 30 min. After staining, the cells were washed three times with PBST on shaker for 10 min. The cells on coverslips were mounted on slides (Fisher Scientific, Cat# 12-544-2) with 10 µL ProLong antifade glass mountant with NucBlue stain (Thermo Fisher Scientific, Cat# P36981), placed in dark room for air-dry overnight. Images in the size of 512 × 512 pixels were acquired on Applied Precision OMX Super Resolution Microscope using a 100X/1.518 oil objective (GE Healthcare Life Sciences) (pixel size = 0.079 µm). Z-stack images were acquired with a 0.3 µm sample thickness.

Identification of nuclear speckle foci

Nuclear speckle foci were identified by a previously described method (Rahl et al., 2010). Briefly, the nuclei were manually segmented and the mean fluorescence intensity in nuclei was measured with FIJI. The nuclear speckle foci were identified by FIJI 3D Object Counter plugin, with an appropriate intensity threshold of the mean fluorescence intensity in the cell nuclei and a size cut-off of more than 50 adjoining pixels (pixel size, 79 nm X 79 nm).

In situ Hi-C library generation and data processing

The Hi-C libraries were generated with the Arima-HiC kit (Arima Genomics, material# A510008, Document# A160134 v00) following the manufacturer's instructions and processed following 4DN consortium's in situ Hi-C processing protocol (https://www.4dnucleome.org/protocols.html). Next, the Hi-C data were processed using the 4D Nucleome (4DN)'s Hi-C data Processing Pipeline (v0.2.5) (https://data.4dnucleome.org/resources/data-analysis/hi_c-processing-pipeline), with MAPQ > 30 to filter out multiple mappings.

The output .pairs files were provided to Cooler (Abdennur & Mirny, 2020) (v0.8.10) and Juicer Tools (Durand, Shamim, et al., 2016) (v1.22.01) to generate .mcool and .hic files. The .mcool file was used in HiGlass (Kerpedjiev et al., 2018) for visualization. The .hic files were inputted in Juicer Tools for A/B compartment, TAD, and loop analyses. A/B compartments were called by Juicer's "Eigenvector" tool based on KR normalized observed/expected (O/E) contacts at 500 kb resolution. TADs were called by Juicer's "Arrowhead" tool based on KR-normalized contacts at 10 kb resolution. Loops were called by Juicer's "CPU HiCCUPS" tool based on KR-normalized contacts simultaneously at 5 kb and 10 kb resolutions. Except for the resolution parameter, all the other parameters were left as the default.

TAD boundaries were extracted as the genomic regions between TADs in each sample. TAD boundary insulation score was calculated according to the definition in (Crane et al., 2015).

Unique loops and overlapping loops were determined as follows. First, the Juicer called loops from each condition were merged into "unique loops" by taking the union. Then the unique loops in the union were reassigned to each condition by the following rule: a unique loop i (in the union) with anchor size s (either 5 or 10 kb) was re-assigned to a sample j if both anchors of loop i were within +/−s flanking regions of a loop in sample j. Aggregate Peak Analysis was performed using the Juicer's "APA" tool with default parameters. Metrics to define the loop strength such as Peak to Lower Left (P2LL), Z-score Lower Left (ZscoreLL), and Peak to Mean (P2M) were calculated as defined in Juicer's APA (Durand, Shamim, et al., 2016). The control loop straddling the AAVS1 locus was detected from H1-hESC Micro-C data (Krietenstein et al., 2020). To select RNase emergent loops that stride across HERV-H caRNA-attached genomic sequences in Control, i.e., the candidate HERV-H caRNA insulated loops (CHRI-loops), we used a threshold of at least 2 iMARGI read pairs with their RNA ends overlapping with HERV-H and their DNA ends mapped to the between-loop-anchor sequence. To check if deleting a copy of the HERV-H repeats led to increase of loop strengths of CHRI-loops, we used a threshold of 0.05 for the delta peak (KO peak - control peak), where peak is the normalized Hi-C read count at the loop's pixel, normalized by the total number of read pairs in each sample.

Production of gRNAs through in vitro transcription (IVT)

gRNA was synthesized by assembly PCR and in vitro transcription as previously described in ref (Wienert et al., 2018). Briefly, a T7 RNA polymerase substrate template was assembled by PCR from a variable 58–59 nt primer containing T7 promoter, variable gRNA guide sequence, the first 15 nt of the non-variable region of the gRNA (T7FwdVar primers, 10 nM), and an 83 nt primer containing the reverse complement of the invariant region of the gRNA (T7RevLong, 10 nM),

along with amplification primers (T7FwdAmp, T7RevAmp, 200 nM each). The two long primers anneal in the first cycle of PCR and are then amplified in subsequent cycles. Phusion high-fidelity DNA polymerase was used for assembly (New England Biolabs). Assembled template was used without purification as a substrate for in vitro transcription by T7 RNA polymerase, using the HiScribe T7 High Yield RNA Synthesis kit (New England Biolabs-E2040S) following the manufacturer's instructions. Resulting transcription reactions were treated with DNAse I (New England Biolabs-M0303S), and RNA was purified by Qiagen RNeasy mini spin column (Qiagen-74104) and eluted in RNAse-free water.

Phosphatase treatment of IVT gRNAs

gRNAs were treated with phosphatases as follows: rSAP (New England Biolabs-M0371S) were added per 20 µl in vitro transcription reaction, and samples were incubated at 37 °C for 3 h before proceeding to purification and DNAseI treatment. gRNA was purified using a Qiagen RNeasy Mini Kit (Qiagen-74104)). The detailed protocol and additional notes are available online (https://doi.org/10.17504/protocols.io.nghdbt6).

gRNA Transfection

dCas9-KRAB hPSCs were treated with doxycycline (2 µg/ml) for 24 h before and during transfection. For transfection, cells were dissociated using Accutase (Stem Cell Technologies), replated onto Matrigel (Corning)-coated plates and transfected in suspension with gRNAs using Lipofectamine RNAiMAX (Invitrogen-13778075) following manufacturer's instructions. Briefly, gRNA (mix of 3 guides against ZMYND8) were added at a 10 nM and 20 nM final concentrations respectively, unless otherwise indicated. gRNAs and Lipofectamine RNAiMAX were diluted separately in Opti-MEM (Gibco-31985062), mixed together, incubated for 5 min at RT, and added dropwise to cultured hPSCs. A second transfection was performed 24 hours later in some experiments. Cells were harvested after 2 days.

Generation of 3 C libraries

3 C libraries were generated as previously described in refs. (Dekker et al., 2002; Miele & Dekker, 2009; Miele et al., 2006). In brief, 10 million cells were crosslinked with 2% Formaldehyde (Thermo Scientific-28908). Chromatin was digested with DpnII (NEB- R0543T), ligated with T4 DNA ligase (NEB-M0202S) and reverse crosslinked by incubation with Proteinase K (Thermo Scientific-EO0491). DNA libraries were purified by phenol/chloroform (Sigma-77617) and subsequent ammonium acetate precipitation (Invitrogen-AM9740). 3 C libraries generated post CRISPRi are used as templates for downstream PCRs.

Bias correction on Hi-C data for loop visualization

H9 Control and HERV-H2 KO Hi-C data were subjected to HiCorr (Lu et al., 2020) with default parameters for bias correction and subsequently subjected to noise removal using the LoopDenoise function in DeepLoop (Shanshan Zhang et al., 2022). All data processing was done with Hg19 per HiCorr and DeepLoop software's requirements.

iMARGI library generation and data processing

iMARGI libraries were generated and processed as previously described in ref. (Wu et al., 2019). According to 4DN's approved iMARGI's data processing protocol (Wu et al., 2019), any iMARGI read pair in which the RNA end and the DNA end mapped to within 1000 bp of each other on the genome are removed from the data analysis. The RNA attachment level (RAL) of each genomic segment is the count of the DNA-ends mapped to this genomic segment (Sridhar et al., 2017). Only the inter-chromosomal and the intra-chromosomal iMARGI read pairs that are separated by at least 200 kb apart were used for calculating RAL in any of the correlation analyses. Repeats of hg38 were downloaded from RepeatMasker (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0). RAL of Alu-containing caRNA (Alu-caRNA) and LINE1-containing caRNA (L1-caRNA) were calculated as the count of the DNA ends mapped to each genomic

segment (500 kb size) whose RNA ends mapped to a repeat segment of the Alu or LINE1 family respectively.

RNA-defined domains

Each rectangular block on iMARGI's contact matrix was identified as a peak of the iMARGI's read pairs' RNA ends (the height of this RNA peak) and a corresponding DNA peak of the DNA ends (the width of this RNA peak). HOMER's findPeaks (Heinz et al., 2010) function was applied to the RNA ends of iMARGI' read pairs (peak size = 5000 bp, minimum peak interval = 12,000 bp) to identify the peaks on the RNA ends (RNA peak). For reach RNA peak, all the iMARGI's read pairs with their RNA ends inside this RNA peak were retrieved. The retrieved read pairs' DNA ends were subjected to HOMER's findPeaks (peak size = 25,000 bp, minimum peak interval = 50,000 bp) to identify the peaks on the DNA ends (DNA peaks). If multiple DNA peaks were reported, the DNA peak with the highest read number was designated as the corresponding DNA peak.

Acknowledgements

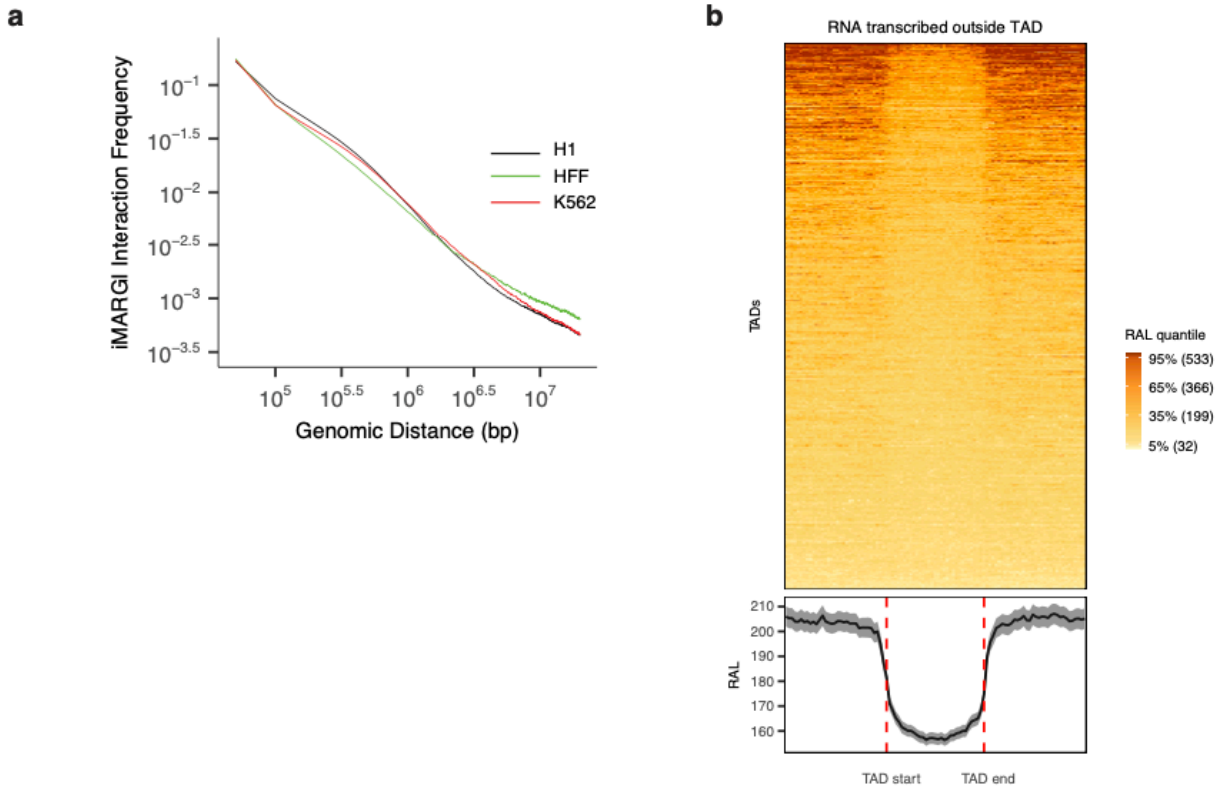
Chapter 1, in full, is a reprint of the material as it appears in Nature Communications 2023. Calandrelli*, R, Wen*, X, Richard, JL Charles, Luo, Z, Nguyen, TC, Chen, CJ, Qi, Z., Xue, S., Chen, W., Yan, Z., Wu, W., Zaleta-Rivera, K., Hu, R., Yu, M., Wang, Y., Li, W., Ma, J., Ren, B., & Zhong, S. Genome-wide analysis of the interplay between chromatin-associated RNA and 3D genome organization in human cells. Nature Communications, 14(1), 6519. The dissertation author was co-first author of this paper.

We thank 4DN Data Coordination and Integration Center for hosting the raw and the preprocessed datasets for public access. We thank the 4DN Cell Working Group, 4DN Omics Standards Working Group, and 4DN Steering Committee for providing guidelines on experimental protocols and data processing pipelines. We thank the 4DN Joint Analysis Working Group for

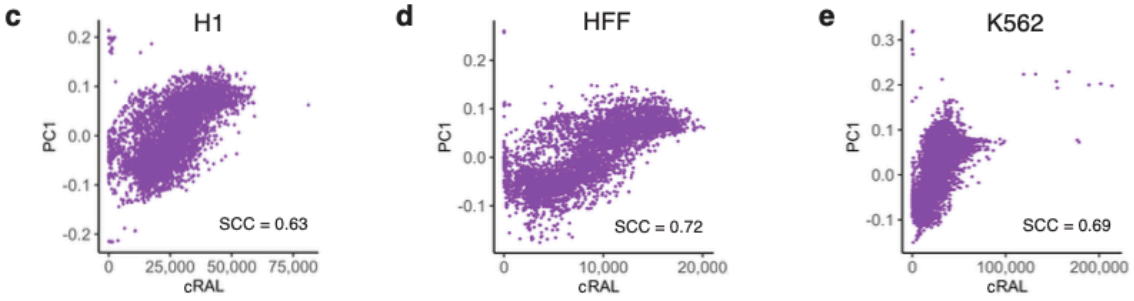
evaluation of all the data in this resource and cross-comparison with the other 4DN Consortium generated datasets. This work is funded by NIH grants DP1DK126138 (S.Z.), R01GM138852 (S.Z.), DP1HD087990 (S.Z.), R01HD107206 (S.Z.), R01GM136922 (W.L.), and NIH Common Fund 4D Nucleome grants UM1HG011593 (J.M.), UM1HG011585 (B.R.), U54DK107977 (B.R.), U01HL156059 (W.L.), U01CA200147 (S.Z.).

Supplementary information

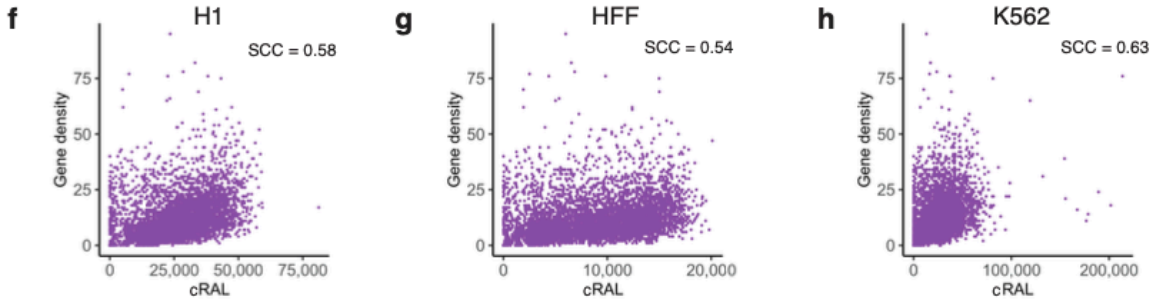
Supplementary Figure 1-1: Distribution of RNA attachment levels (RAL) on the genome.
(a) RNA-DNA contact frequency (y axis) vs. the genomic distance between the mapped RNA-end and the DNA-end (x axis) in H1 (blue), HFF (green), and K562 cells (red). (b) The RAL of every TAD (row) and its equal-length flanking regions, based on all the RNAs transcribed from any genomic sequences outside of this TAD (row). The TAD lengths are normalized (center, x axis). Curve at the bottom: average RALs of all TADs. (c-e) Scatterplot of Hi-C's first eigenvector (PC1, y axis) and cRAL (x axis) on every 500 kb genomic bin (dot) of the entire genome in H1 (c), HFF (d), and K562 (e). SCC: Spearman correlation coefficient. (f-h) Scatterplot of gene density (y axis) and cRAL (x axis) on every 500 kb genomic bin (dot) of the entire genome in H1 (f), HFF (g), and K562 (h). The correlation between cRNA and gene density is weaker than the correlation between caRNA and Hi-C's PC1.



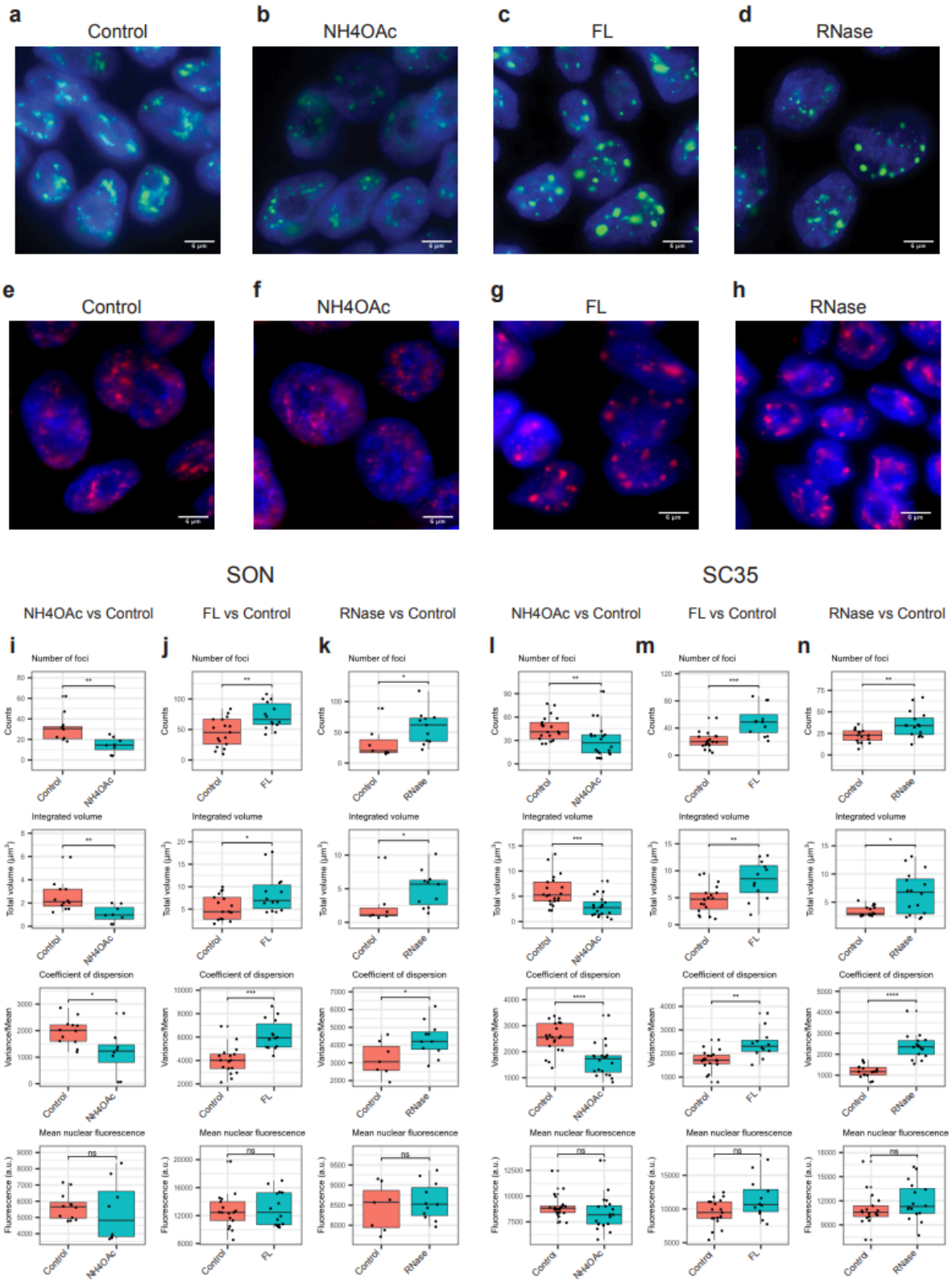
Genome-wide comparison of cRAL and A/B compartments

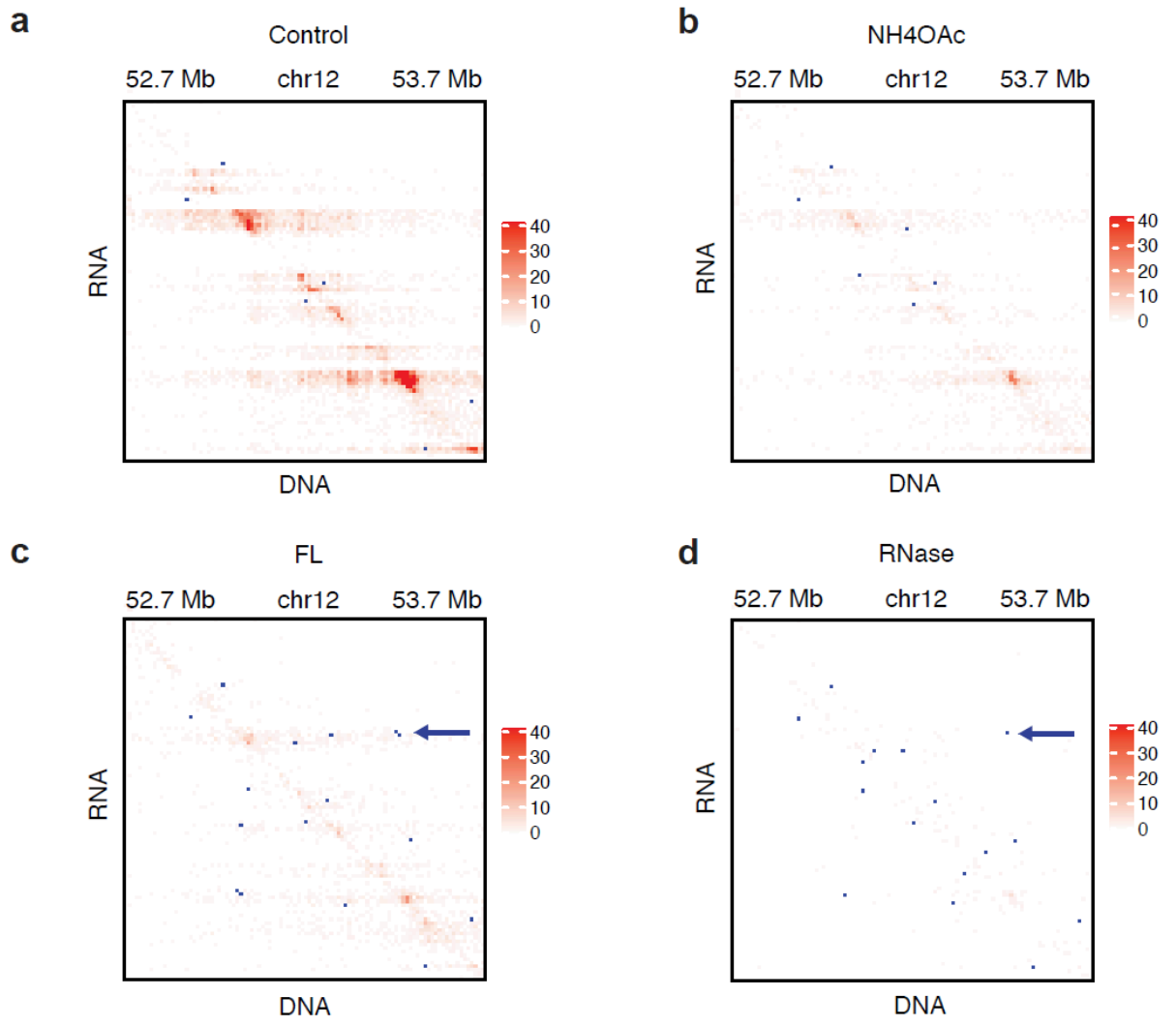


Genome-wide comparison of cRAL and gene density

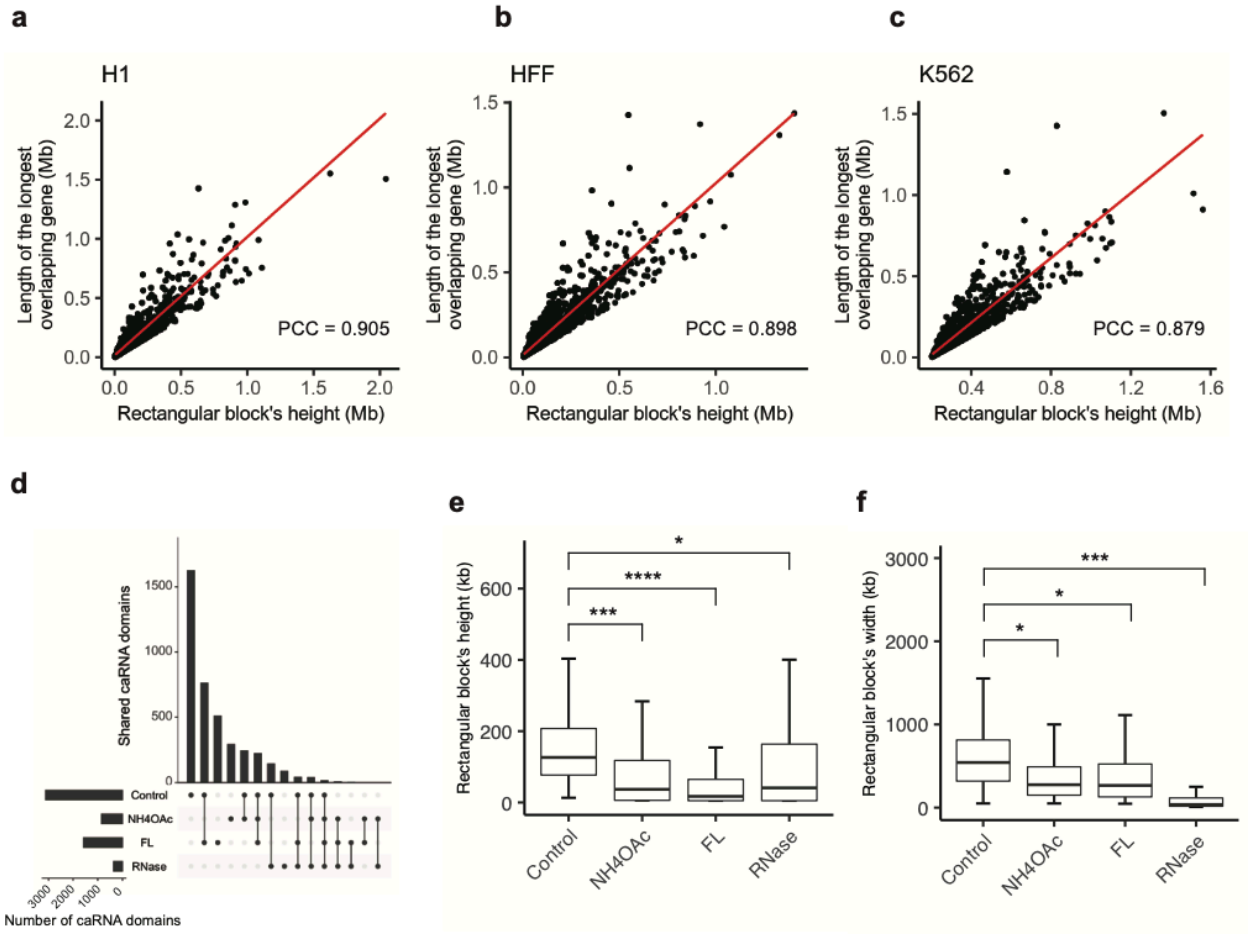


Supplementary Figure 1-2: Immunofluorescence analyses of SC35 and SON.
(a-h) Immunostaining of SON (ad) and SC35 (e-h) in control, NH4OAc, FL, and RNase-treated H1 cells. Scale bar = 6 μm. (i-k) Distribution of SON's average number of foci per nucleus in control and each treatment (first row). *: p-value < 0.05. **: p-value < 0.01, Wilcoxon test. In comparison, SON's mean background fluorescence (last row) does not change between control (pink) and each treatment (green). ns: not significant. (l-n) Distribution of SC35's average number of foci per nucleus in control and each treatment (first row). **: p-value < 0.01. ***: p-value < 1.0e-3, Wilcoxon test. In comparison, SC35's mean background fluorescence (last row) does not change between control (pink) and each treatment (green). ns: not significant. The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.


a
Control
b
NH4OAc
c
FL
d
RNase
e
Control
f
NH4OAc
g
FL
h
RNase
SON
SC35
NH4OAc vs Control
FL vs Control
RNase vs Control
NH4OAc vs Control
FL vs Control
RNase vs Control
i
j
k
l
m
n
Number of foci
Integrated volume
Coefficient of dispersion
Mean nuclear fluorescence
Counts
Total volume (μm³)
Variance/Mean
Fluorescence (a.u.)
Control
NH4OAc
FL
RNase
ns



Supplementary Figure 1-3: An example of loop change.
iMARGI contract matrix of control (a), NH4OAc (b), FL (c), and RNase (d) for the corresponding genomic region of Figure 1.3d. Blue dots: Hi-C derived loops that are superimposed on the iMARGI contact maps. Arrows: a shared loop in FL and RNase that is absent in control and NH4OAc.



Supplementary Figure 1-4: RNA-association domains.
(a-c) Scatter plots of each RNA-association domain's size (x axis) and the length of the longest gene in each RNA-association domain (y axis). Each dot represents an RNA-association domain, corresponding to a detected rectangular block from iMARGI's contact matrix. The width and the height of each rectangular block correspond to the size of an RNA-association domain and the length of the genomic sequence that produced the caRNA in this domain. The height of each rectangular block often matches the length of the longest gene overlapping with this RNA-association domain, suggesting that most RNA-association domains are decorated by the RNA of single genes. (d) Upset plot of the numbers of RNA-association domains in untreated H1 (Control) and H1 treated with NH4OAc, FL, and RNase. (e) Distributions of the heights of the detected blocks, *: p-value < 1e-25, ***: p-value < 1e-75, ****: p-value <1e-100, Wilcoxon test. (f) Distributions of the widths of the detected blocks. *: p-value < 1e-50, ***: p-value < 1e-180, Wilcoxon test. The centerline of the boxplots is the median. The lower and upper hinges correspond to the first and third quartiles (the 25th and 75th percentiles). The upper whisker extends from the hinge to the largest value no further than 1.5 * IQR from the hinge (where IQR is the inter-quartile range, or distance between the first and third quartiles). The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge.

Supplementary Table 1-1: Summary of iMARGI and Hi-C

a. iMARGI datasets				
Cell line	Treatment	Number of replicates	Total # of read pairs	Source
H1	None	4	2,642,778,166	This work
	NH4OAc	2	1,247,204,613	
	FL	2	1,438,312,761	
	RNase	2	1,670,230,715	
HFF	None	2	2,755,576,893	
K562	None	2	1,293,950,206	
b. Hi-C datasets				
Cell line	Treatment	Number of replicates	Total # of read pairs	Source
H1	None	2	616,625,628	This work
	NH4OAc	2	613,098,350	
	FL	2	604,503,572	
	RNase	2	654,798,738	
HFF	None	2	2,764,855,452	4DNESNMAAN97
K562	None	6	907,136,828	4DNESI7DEJTM
c. iMARGI datasets in engineered cells				
Cell line	Treatment	Number of replicates	Total # of read pairs	Source
H9 MLC2v:H2B	None	1	692,140,673	This work
H9 MLC2v:H2B HERV2-KO	None	1	566,783,631	
H9 MLC2v:H2B HERV2-ins-clone2	None	1	967,136,454	
d. Hi-C datasets in engineered cells				
Cell line	Treatment	Number of replicates	Total # of read pairs	Source
HCT116 RPB1-Dox-OsTIR1-mClover-mAID	Doxycycline (control)	1	632,233,849	This work
	Doxycycline and 5-Ph-IAA	1	716,607,191	
H9 MLC2v:H2B	None	2	552,532,207	GSE116862
H9 MLC2v:H2B HERV2-KO	None	2	635,909,624	

Chapter 2 Delineating the Functional RNA-Chromatin Interactome by RNase Treated iMARGI in Human Embryonic Stem Cells

Abstract

RNA-chromatin interactions play crucial roles in gene regulation and genome organization, but the interaction landscape remains poorly understood. In this study, we conducted an in-depth analysis of a previously published dataset on RNase-treated in situ mapping of the RNA–genome interactome in human embryonic stem cells (H1 cell line). This dataset globally profiles RNase-insensitive RNA-chromatin interactions. Our analysis revealed that RNase treatment selectively preserved long-range RNA-chromatin interactions while removing promiscuous interactions resulting from the local diffusion of nascent transcripts.RNase-insensitive chromatin-associated RNAs (RI-caRNAs) exhibited high sequence conservation and preferentially localized to functional genomic regions, including promoters, transcription factor binding sites, and regions with specific histone modifications. Interestingly, coding and non-coding RNA transcripts showed distinct sensitivities to RNase, with lncRNAs and disease-associated transcripts being enriched among RI-caRNAs. Furthermore, we identified specific caRNA classes associated with individual transcription factors and histone modifications. Altogether, our findings reveal a RNase-inaccessible regulatory RNA-chromatin interactome and provide a resource for understanding RNA-mediated chromatin regulation.

Introduction

The functions of nuclear RNAs are typically discerned through their interactions and localization patterns. Current methods for investigating the functions of nuclear RNAs can be categorized into two approaches: (1) Elucidating RNA function via its interactions with other molecules, such as RNA-protein interactions (Van Nostrand et al. 2016; Hafner et al. 2010) or RNA-RNA interactions (Lu et al. 2016; Wu et al. 2019). 2) understand RNA's function through

its localization pattern inside the nucleus by either probing the chromatin regions that RNAs are spatially close to using proximity ligation based methods (Li et al. 2017; Sridhar et al. 2017; Wu et al. 2019) or ligation free methods (Quinodoz et al. 2018), or alternatively investigating the nuclear bodies/compartments-associated RNAs by labeling the nuclear body specific proteins which will then diffusively add biotin to nearby RNA molecules (Chen et al. 2018; Fazal et al. 2019).

Although both methodologies have contributed novel insights, each approach has its advantages and limitations. The first category of methods, which focus on the RNA-centric interactome, can identify co-binding molecules and binding region sequences (motifs). However, they do not provide information on the genomic locations of these interactions, which is a crucial aspect with respect to gene regulation. Proximity-ligation-based methods enable the high-throughput genome-wide mapping of RNA footprints on chromatin, but they are also susceptible to capturing numerous non-specific interactions, such as those involving nascent transcripts.

Ribonucleases (RNases) are commonly utilized to digest RNAs and have been incorporated into various molecular experimental protocols for studying RNA structures and functions. For instance, PIP-seq is capable of accurately detecting RNA-binding protein (RBP)-RNA interaction sites by introducing RNase after protein-RNA crosslinking. RNase digestion will only leave behind the interaction motifs that are protected by RBPs, resulting in their enrichment (Silverman et al. 2014). Moreover, rChIP-seq is an RNA-dependent ChIP-seq protocol that enables the distinction between RNA-dependent and RNA-independent protein localization on the genome (Long et al. 2020). ChIP-seq is conducted with or without RNase treatment to determine whether a transcription factor (TF) is RNA-dependent or not.

iMARGI (in situ mapping of RNA–genome interactome) is a powerful genome-wide assay that reveals interactions between chromatin-associated RNAs (caRNAs) and genomic DNA sequences using proximity ligation (Wu et al. 2019; Yan et al. 2019). In this study, we reanalyzed a previously published iMARGI dataset (Calandrelli et al. 2023), , which incorporated RNase treatment before the standard iMARGI procedure to remove promiscuous, unprotected RNA transcripts and selectively enrich for RNA-inaccessible, potentially functional RNA-chromatin interactions. In human embryonic stem cells, we provide a comprehensive map of the RNA-inaccessible RNA-chromatin interactome and identify specific RNA classes associated with regulatory genomic regions. Our results show that RNA-inaccessible chromatin-associated RNAs (RI-caRNAs) exhibit high sequence conservation and preferential localization to functional genomic regions, including promoters, transcription factor binding sites, and histone modifications. This technology provides a valuable tool to study functionally relevant caRNAs, with potential implications for understanding embryonic development and disease genesis.

Results

Overview of RNase treated iMARGI libraries

Overview of RNase-treated iMARGI libraries In this study, we reanalyzed previously published RNase-treated iMARGI datasets to investigate the non-diffusive RNA-chromatin interaction profile genome-wide (Figure 2.1a). The original approach involved perturbing the cell nucleus with RNase A (an endoribonuclease that targets single-stranded RNA) before crosslinking, thereby eliminating unprotected RNA from the nucleoplasm and chromatin, and enriching for non-diffusive interactions. Compared to the standard iMARGI protocol, the RNase-treated method included a 10-minute RNase A treatment before crosslinking, followed by the standard iMARGI protocol (Methods). For this analysis, three biological replicates were generated for the RNase-treated samples, and five technical control replicates were prepared using the unmodified iMARGI

protocol (control libraries) in the human embryonic cell line (H1). In total, 1,589,053,749 and 2,485,334,610 mappable non-duplicated read pairs were generated for the RNase-treated and control libraries, respectively (Supplementary Table 1). We selected read pairs with both RNA and DNA ends uniquely mapped to the human genome and with a mapping score (MAPQ) greater than 30 for downstream analysis to ensure high confidence and quality; these pairs are referred to as valid pairs hereafter. These stringent criteria resulted in 192,077,439 usable read pairs in the three RNase-treated libraries combined and 341,486,278 in the five control libraries combined.

Library reproducibility was accessed using RNA ends or DNA ends from iMARGI read pairs individually. We first measured the RNA ends reproducibility by gene read counts using RNA ends only. The lowest correlation between any two pairs of RT-libraries is 0.908 (spearman correlation, pvalue<2.2e-16) and is 0.928 between any pairs of control libraries (Supplementary Figure 2.1a, b). For DNA ends, we calculated the genome-wide chromatin attachment levels using sliding windows of 100 kb and also observed high correlations between replicates, although slightly lower than those for RNA ends (Supplementary Figure 2-1c, d). These data suggest that the RT-iMARGI libraries are highly reproducible across replicates.

Long distance RNA-chromatin interactions are resistant to RNase treatment

An important aspect of RNA-chromatin interactions is the genomic distance between the RNA transcription site and its interacting chromatin regions. Different interaction distances are likely to be governed by distinct mechanisms. For example, extremely short-distance interactions might be attributed to local diffusion of nascent transcripts, while long-distance interactions are more likely to be mediated by RNA-binding proteins (X. Li & Fu, 2019a). We first investigated whether the distribution of genomic distances between RNA and DNA ends was altered after RNase A treatment. To this end, we calculated the genomic distance between the RNA and DNA ends for every valid read pair. We found that super-nascent interactions (RNA-DNA end distance

< 1 kb) were selectively preserved, increasing from 23.00% of total pairs in the control library to 84.53% in the RNase library (Supplementary Figure 2-2a, proportion test: p-value < 2.22e-16). However, for read pairs with RNA-DNA end distances longer than 1 kb, a larger proportion of distal pairs (>2Mbp or interchromosomal) were significantly enriched (Supplementary Figure 2-2b, c, proportion test: p-value < 2.22e-16). The selective preservation of long-distance RNA-chromatin interactions was independent of individual chromosomes. These results suggest that chromatin-associated RNAs (caRNAs) involved in short- to middle-range RNA-genome interactions are more susceptible to RNase treatment, likely due to their nature of local diffusion. The enrichment of extremely proximal RNA-genome interactions is RNase resistant, potentially due to protection by RNA polymerases. The preservation of distant interactions (RNA-DNA end distance >1 kb) in both RNase and control libraries might indicate functional roles for the protected caRNAs, which will be the focus of our subsequent studies.

RNase-inaccessible caRNA transcripts characteristics

We profiled RNase-inaccessible chromatin-associated RNA (RI-caRNA) features by examining RNA ends from iMARGI read pairs. 83.58% of RNA ends mapped to genes (GENECODE V36 (Frankish et al., 2019), strand specific) in the RNase-treated library, compared to 90.94% in the control (Figure 2.1b), suggesting genic caRNAs are more sensitive to RNase. 33.19% and 37.42% of caRNAs were derived from repeat elements in RNase and control respectively (Figure 2.1c). While overall repeat-derived caRNA proportions were similar, some classes like rRNA, satellite repeats, and low complexity RNAs were significantly enriched in the RNase library (FCs: 6.73, 5.17, 3.32, 3.24) (Figure 2.1d), suggesting a subset of intergenic and repeat caRNAs are RNase-inaccessible (Thakur and Henikoff, 2020).

Among all genic transcripts, we asked whether there are any genes whose transcripts are inaccessible after RNase A treatment. We called 624 RNase library specific genes and 493 control

library specific genes using differential analysis (Figure 2.1e, Methods). We first noticed that lncRNAs (n=257) are over-represented among these 624 RNase specific genes, accounting for 41.93% in RNase library compared to 19.27% in Control library specific genes (pvalue < 2.22e-16, Fisher exact test. Figure 2.1f). This result suggests that lncRNA transcripts are preferably protected from RNase A digestion. To associate the above identified RNase library specific genes with known functions such as disease genesis, we annotated all the RNase specific genes with DisGeNet registered disease associated genes (includes 20,158 genes associated with 21,775 disease) (Piñero et al., 2020) and all lncRNAs with LncRNADisease databases (including 19,485 genes associated with 498 diseases, Figure 2.1g) (G. Chen et al., 2013). The intersections between both libraries are significant (Supplementary Figure 2-2d, chi-square test, pvalue<2.2e-16). Disease ontology enrichment test also suggests all RNase library specific lncRNAs enriched with disease terms including several cancers, Alzheimer's disease, Huntington's disease and acquired immunodeficiency syndrome and others (hypergeometric test, FDR<0.05, Supplementary Figure 2-2e, Methods). In summary, these results suggest that RNase library specific genes are enriched with lncRNAs and significantly associated with diseases.

RNase-inaccessible caRNAs have high level of evolutionary conserveness

Previous studies using RNase-mediated protein footprint sequencing identified protein-protected RNA–protein interaction sites with conserved motif sequences (Ji et al., 2016). We hypothesized that the RNA ends inaccessible to RNase are protected by proteins, and thus are evolutionarily conserved. To test this, we calculated conservation scores of all 192 million RNA ends in the RNase-treated iMARGI library and 341 million RNA ends in the control library. We used phaseCons100 (multiple alignment of 100 vertebrate genomes) (Kent et al., 2002) as a measure of sequence conservation level. We categorized all pairs into three groups based on their RNA-DNA end distance: 1) nascent pairs (<1kbp), 2) cis pairs (>1kbp - 2Mb), 3) trans pairs (>2Mb

and interchromosomal). We calculated the conservation score at single base resolution, ranging from the center of each read to 500bp flanking regions. The conservation level was consistently higher for RNA ends in the RNase library compared to control in every group (Figure 2.2a, b: dashed lines vs. solid lines, t-test p-value < 2.22e-16). Trans pairs exhibited higher conservation than cis pairs, while nascent pairs had the lowest scores. These results suggest the RNase inaccessible caRNA is more conserved than the other caRNA.

To gain a clearer picture of what the remaining RNA ends after RNase digestion look like, we analyzed the reads coverage and sequence features of two non-coding genes from the top enriched RNase library specific RNAs: snRNA RNU5B-1 and snoRNA SNORA74A. Reads covering RNU5B-1 are piled up at the Sm domain at the tail of the gene (Figure 2.2d). Sm domain is a conserved region featured by the consensus sequence Purine-AU4-6G-Purine, and is recognized and bound by spliceosomal Sm proteins (Urlaub et al., 2001). The top one sequence motif identified from all RNA ends transcribed from RNU5B1 is consistent with Sm motif (Figure 2.2d). Reads transcribed from SNORA74 are overlapped with snoRNA featured H/ACA box motif (Figure 2.2c) (Urlaub et al., 2001). These examples illustrate that RNaseA treatment was less effective at removing transcripts within the functional regions of a gene whose parts are likely to be interacted with proteins.

Majority of RNase-inaccessible caRNAs targeting regions are functional genomic regions

To understand the potential functional roles of RNase-inaccessible caRNAs, we investigated their localization patterns on the genome, we first called peaks using DNA end coverage across the genome. In total, 99,784 peaks were identified in the RNase library, with a fragments in peaks ratio (FRiP) of 13.15%, indicating a high signal-to-noise ratio (ENCODE uses FRiP > 1% to filter high-quality transcription factor ChIP-seq experiments (Landt et al., 2012)).

After applying the peak max depth threshold of 15 reads, 25,333 peaks, referred to as RNA Attachment Hot zones (RAHs), were identified for further analysis.

Around 49.4% of RAHs were in intronic regions, followed by 23% in intergenic and 16.8% in promoter regions. To assess significance, we compared these proportions to 2,062,460 transcription factor binding sites (TFBS) from 71 TFs in H1 cells (R. Zheng et al., 2019) and randomized genomic regions with the same length distribution as RAHs. The 16.8% RAH overlap with promoters was significantly higher than 3.28% for randomized regions, but lower than the 25% TFBS promoter overlap (Figure 2.3d). While the 49.37% intronic RAH proportion was large, it was comparable to the 47.53% genomic background. The 22.97% intergenic RAH proportion was approximately half of the 40.19% background. These data suggest promoter regions are hotspots for caRNA.

We then annotated RAHs with chromatin features: 1) Accessibility (ATAC-seq and DNase I hypersensitive sites), 2) Histone modifications, 3) TFBS, 4) Nuclear periphery (Figure 2.3a). 35.88% of RAHs overlapped open chromatin regions marked by DNase I or ATAC-seq in H1 cells, six-fold higher than 5.8% of 2,169,280 control library peaks (Figure 2.3b). This RNA targeting preference mirrors transcription factors preferring open chromatin (ENCODE). Additionally, 53.85% of RAHs overlapped with one or more of the TFBS from the 71 TFs surveyed. Histone modifications further increased the overlap ratio with RAH to 63.71% (Figure 2.3b). In contrast, only 29% of control peaks were explained by open chromatin, TFBS, and histone modifications (Figure 2.3b). These data suggest RNase-inaccessible caRNA co-localize to the functional genomic regions.

Some individual transcription factors (TFs) extensively overlap with RAHs. For example, 29.50% of all RAHs overlapped with YY1 binding sites, 26.77% with TEAD4 binding sites, and

14.64% with CTCF. Reciprocally, RAHs also accounted for a sizable proportion of individual TF peaks. Ranking TFs based on the proportion of their TFBS overlapping with RAHs, we found EZH2 (18.60%), SUZ12 (17.53%), POU5F1 (14.42%), TAFII (14.05%), SP2 (13.91%), and others. All ratios were significantly larger than randomized genomic regions. Interestingly, top-ranked TFs with TFBS explained by RAHs have emerging evidence suggesting dual roles as transcription factors and RNA-binding proteins (see discussion). RAHs simultaneously overlapping multiple TFBS were highly enriched in promoter regions and CpG-rich regions (last two columns, Figure 2.3a). Notably, 7,551 RAHs (28.90%) overlapped with histone variant H2A.Z peaks (Figure 2.3a, column cluster 9), the largest factor explaining RAHs after DNase open chromatin regions (33.70%). These results might indicate potential interplay between RNA molecules and the H2A.Z variant. Hierarchical clustering of the RAH-feature overlap matrix highlighted functionally similar markers with similar RAH overlap profiles, e.g., cohesion/insulator markers CTCF, RAD21, and ZNF143 clustered in column cluster 6 (Figure 2.3a). Polycomb complex histone marker H3K27me3, EZH2, and SUZ12 clustered in column cluster 1 (Figure 2.3a).

For 8,400 RAHs not overlapping known TFBS or histone modifications, we asked whether they were enriched in inaccessible regions like heterochromatin and the nuclear lamina. We utilized Lamin B1 DamID signals, reflecting a genomic region's proximity to the nuclear lamina (Guelen et al., 2008), to annotate each RAH. RAHs without TF binding, chromatin accessibility, or histone modification annotations had significantly higher DamID-Lamin B1 signals, suggesting closer physical proximity to the nuclear lamina and heterochromatin (t-test, p-value < 2.2e-16, Figure 2.3e).

Finally, we asked whether RAHs overlapped with any repetitive elements (REs). Only 46.5% of RAHs overlapped with UCSC RepeatMasker-annotated REs, less than the 68.31% background from randomized regions with the same length distribution as RAHs (Figure 2.3f), suggesting RAHs are less enriched in genomic RE regions. Among RE-overlapping RAHs, SINE/MIR, LTR/ERV1, and LTR/ERVL were three RE classes significantly associated with RNase-insensitive caRNAs, with 1.74-, 2.85-, and 2.71-fold enrichment compared to the proportion of genome containing each repeat class annotated in RepeatMasker annotations (Figure 2.3g). Interestingly, previous studies suggested SINE/MIR (mammalian-wide interspersed repeats) are a conserved transposable element family with significant regulatory capabilities and sequence similarities to tRNA-related genomic insulators (Pol III binding), independent of the CCCTC-binding factor (CTCF) insulation mechanism (J. Wang et al., 2015). Similarly, the LTR/ERV1 and LTR/ERVL families include HERV-int repeats proposed to demarcate topologically associating domains in human pluripotent stem cells (Zhang et al., 2019). Together, these results suggest RI-caRNAs preferentially interact with genomic repeats functioning as organizational insulators.

In summary, these data show RNase-inaccessible caRNA tends to localize at functional genomic regions. Major caRNA localization hotspots include gene promoter regions, transcription factor binding sites, H2A.Z chromatic regions, as well as H3K27me3 modified and Polycomb associated chromatic regions.

RNase removes promiscuous RNAs and magnifies the stringent interactions between RNA-TFs and RNA-histone modifications.

Our previous results suggested that caRNA genome interaction hot zones (RAHs) often colocalize with transcription factor binding sites (TFBS) or specific histone marks. We then asked whether caRNA co-localize with the TFBS of a specific TF or a specific histone mark (referred to

as marker peak hereafter). The amount of caRNAs interacting at the marker peak versus surrounding regions was used to determine enrichment. To this end, we examined multiple markers in the H1 cell line and generated a series of heatmaps charting the caRNA coverage in both control and RNase libraries at each peak region (5kb extended upstream and downstream from the peak center) for transcription factors and histone modifications (Figure 2.4, Supplementary Figure 2-4). For each heatmap, rows represent all ChIP-seq peaks registered in the Cistrome or ENCODE database (ENCODE Project Consortium, 2012; T. Liu et al., 2011), and columns represent 10kb surrounding regions of a peak center (1000 bins). We summarized TFs and histone modifications into two classes based on their caRNA-peak abundance pattern in control and RNase libraries: Class 1) TFs or histone modifications with caRNA peak center versus background ratio increases in the RNase library compared to the Control. Class 2) TFs or histone modifications with caRNA peak center versus background ratio decreases in the RNase library compared to the Control.

For the first class, we observed that the caRNA are concentrated at the TFBS ChIP-seq peak centers. This observation applied to both the control and RNase libraries (Figure 2.4a, Supplementary Figure 2.3a, c). However, unlike the control library, the RNase-treated library maintained only the TFBS-associated RNAs, not the surrounding ones (Figure 2.4c). We then asked whether the remaining RNase peak signals were simply a result of the non-selective decrease of overall signal from the control to the RNase library. To answer this, we designed a concentration score, calculated as the ratio of caRNA coverage between the marker peak center surrounding 100 bins (upstream and downstream 500bp) and the surrounding 300 bins (upstream and downstream 1500bp). Higher values indicate stronger enrichment at the central peak. The median concentration score for CTCF was 0.405 for the control library, suggesting caRNAs specifically target CTCF peaks. However, the median CTCF concentration score was 0.445 in the RNase library, much

higher than control (Figure 2.4b). Paired t-tests suggest the concentration is significantly higher in the RNase than control, considering every CTCF peak (Figure 2.4b). The same trend was observed for many other Class 1 TFs (control:RNase): BCL11A (0.429:0.570), EP300 (0.374:0.523), SOX2 (0.379:0.492), RAD21 (0.408:0.462), CHD7 (0.441:0.559), YY1 (0.419:0.468) (Figure 2.4b, Supplementary Figure 2-4a, b). Paired t-tests for all above TFs showed significant enrichment (p-value < 2.2e-16). These results show that the effect of RNase in decreasing RNA abundance is not uniformly applied to every position. The RNA signals at the TF binding regions are selectively preserved. These data are in line with recent reports on RNA's roles in facilitating protein-chromatin binding (Hansen et al., 2019; Saldaña-Meyer et al., 2019; Sigova et al., 2015; Xiao et al., 2019).

Unlike the first class of markers where caRNAs are selectively preserved at their DNA binding sites after RNase treatment, the caRNA level at the second class shows significant depletion. Markers in this category include suppressive mark (control:RNase): H3K27me3 (0.337:0.305, p-value < 2.2e-16), gene body histone mark H3K36me3 (0.342:0.320, p-value < 2.2e-16), and the active gene histone mark H3K79me2 (0.340:0.310, p-value < 2.2e-16) (Figure 2.4d). The second category of histone modifications are known to be broad peaks (J. Wang et al., 2013) and require continuous deposition through protein complexes, such as PRC2 for H3K27me3 (Margueron & Reinberg, 2011) and RNA polymerase II for H3K36me3 during active transcription (Yoh et al., 2008), to maintain their states. Attenuated peaks in RNase suggest caRNA co-localizing with these histone marks are not protected by stable protein association. Decreased caRNA levels have also been observed for class II markers such as histone modifications H3K23me2 (0.320:0.298), H2BK12ac (0.411:0.373), TRIM28 (0.419:0.411), EZH2 (0.373:0.365), and SUZ12 (0.369:0.362).

We further surveyed all histone modification ChIP-seq and TFBS data from Cistrome and ENCODE to study the change in concentration score from the control library to the RNase A-treated library (Figure 2.4g). All markers broke down into two types: caRNA at marker peak enriched in the RNase-treated library or depleted. We tested different combinations of background bin and peak central bin numbers to derive robust concentration change rankings for all markers, reducing bias from varied peak sizes (we tested three combinations: TFBS central 1000bp vs. central 3000bp, central 500bp vs. central 2000bp, and central 800bp vs. 1500bp). In total, 69 markers showed consistent changes across different concentration calculation metrics, with 58 markers having enriched caRNA levels at their peaks after RNase A digestion and 11 markers decreasing (Figure 2.4h). By surveying published studies, we found that 22 out of the 58 iMARGI-identified Class 1 TFs are both RNA and DNA-binding proteins, including EP300, SOX2, JUN, CHD7, BRCA1, CHD2, GTF2F1, ATF2, EGR1, TAF7, MAX, SIN3A, GABPA, POLR2A, USF2, RAD21, CTCF, CEBPB, RBBP5, YY1, and others, suggesting a potential mechanism where caRNAs are protected from RNase digestion by protein binding.

Taken together, these data suggest that caRNA-TF association is pervasive and applies to many TFs. RT-iMARGI magnifies the stringent interaction between TFs and caRNAs that are constituents and robust to RNase perturbation. The attachment profile is highly specific to the intrinsic characters of the target.

Transcription factor specifically associated RNA classes

Previously, we have shown that for many transcription factors (TFs), chromatin-associated RNA molecules target their transcription factor binding sites (TFBS), and these associations are less sensitive to RNase treatment than the caRNA at the flanking genomic regions. We then asked whether, for each TF, its TFBS is associated with any specific RNA species.

To identify RNA species enriched in associating with specific TFs, we compared the relative abundance of TFBS-targeting caRNA and all iMARGI detected caRNA. Specifically, we tested the dependence between gene_i targeting TF_j and RNase treatment using a chi-square test. The null hypothesis is that there is no dependence between gene_i targeting TF_j TFBS and RNase treatment, such that the chance of seeing gene_i transcripts targeting all TFBS of TF_j compared to not targeting TF_j TFBS would be no different than without RNase treatment. The alternative hypothesis was a dependence such that RNase treatment selectively preserved some TF-associated caRNAs (Figure 2.5a). We tested TF-enriched genes for each TF that had increased center to flanking region caRNA attachment level in the RNase library from the previous section, examining 52 TFs. The chi-square p-value distribution was heavily skewed towards 0, indicating TF-associated genes significantly enriched in the RNase-treated library (Supplementary Figure 2-4a, c). We identified thousands of genes enriched for each TF under an adjusted p-value < 0.05 (Benjamini & Yekutieli 2001, no assumption on gene independence) and odds ratio > 1 (Figure 2.5d).

To examine whether iMARGI-identified TF-specific associated RNA species were expected, we compared the candidate list with two related datasets: 1) RedChIP data, measuring protein-involved chromatin-associated RNAs (Gavrilov et al., 2022), and 2) fRIP-seq data, measuring protein-bound RNA by formaldehyde RNA immunoprecipitation (G Hendrickson et al., 2016). The first dataset offers an orthogonal strategy for detecting caRNAs by pulling down target proteins and detecting associated DNA and RNA. Although the second dataset does not directly measure caRNAs, it aligns with our hypothesis that RNase treatment-insensitive caRNAs may be protected by protein binding. We found that iMARGI-identified CTCF TFBS-enriched genes significantly overlapped with both RedChIP-detected CTCF-associated caRNAs (RedChIP

data generated in K562 cells; odds ratio = 15.3, p-value < 2.2e-16) and fRIP-seq-detected CTCF-bound RNAs (fRIP-seq data generated in K562 cells; odds ratio = 6.26, p-value < 2.2e-16) (Figure 5b, c). The overlap with RedChIP was even more significant than fRIP-seq, consistent with the idea that both iMARGI and RedChIP detect chromatin-associated RNA components.

Similarly, TFs CHD7 and RBBP5 were surveyed in both iMARGI (in H1 cells) and fRIP-seq (in K562 cells). fRIP-seq nominated 994 CHD7 protein-bound RNAs, 361 of which overlapped with iMARGI-identified CHD7 binding site-enriched caRNAs (odds ratio = 6.46, p-value < 2.2e-16) (Supplementary Figure 2-4b). For RBBP5 (Retinoblastoma-binding protein 5), fRIP-seq identified 984 specific genes, and 107 overlapped with RBBP5 binding site-enriched caRNAs (odds ratio = 2.12, p-value = 5.4e-15) (Supplementary Figure 2-4d). Together, these data suggest RT-iMARGI-identified CTCF TFBS-enriched caRNAs can be rationalized using orthogonal technologies.

Since iMARGI-derived caRNA-TFBS association specificity does not restrict to any existing antibody of the target protein of interest, we could massively survey all TF-specifically associated caRNAs. We then asked whether any TFs share the same set of associated caRNA genes. We plotted all the caRNA that are enriched in the TFBS of any analyzed TF in a heatmap colored by the degree of association (odds ratio) (Figure 2.5d). TFs were clustered based on their specifically associated RNA profiles using hierarchical clustering based on Jaccard distance. We found that the RNA-mediated TF-TF similarities are highly correlated with TFBS similarities on the genome, measured by their co-localization (Figure 2.5e, rho = 0.74, p-value < 2.2e-16, Methods). Nevertheless, some TF pairs showed elevated correlation in RNA-mediated profiles compared to genomic footprints, such as BCL11A, SOX2, and EP300. Recent studies have

revealed that the pluripotency factor SOX2 is an RNA-binding protein (Z. E. Holmes et al., 2020) and has direct protein-protein interactions with the coactivator EP300 (B. R. Kim et al., 2017).

For each TF, we annotated the gene species composition (Figure 2.5d, bottom and right tracks). Surprisingly, protein-coding genes were the largest proportion of gene species for every examined TF. While many non-coding genes have been observed associated with transcription factors and histone marks, the role of protein-coding genes in epigenetic regulation is less studied. We then asked whether any gene species were more likely to associate with many markers. We found that snRNAs were the most promiscuous type, with the median number of associated markers for all snRNAs being 29, followed by snoRNAs (median = 23), mRNAs (median = 20), and lncRNAs (median = 6) (Figure 2.5f). We then asked whether the number of associated markers was attributable to high gene expression levels. The Spearman correlation between gene expression and the number of associated markers was 0.671 (Supplementary Figure 2-4e, p-value < 2.2e-16), suggesting transcripts with higher abundance tend to enrich at more TFBS. These data are consistent with the large proportion of mRNAs observed in both RedChIP and fRIP-seq data, suggesting chromatin-associated proteins (CAPs) bind promiscuous RNAs likely correlated with RNA abundance within the nucleus.

Taken together, these data suggest that different TFs are enriched with different caRNA species that differentiate them from their genomic footprints. Highly abundant gene transcripts, including protein-coding RNAs, are widely observed to be enriched at TFBS.

Histone modification coated caRNA species are sensitive to RNase treatment

The physical interaction of the PRC2 complex with RNAs has long been studied. RNA has been proposed to play a critical role in recruiting the PRC2 complex to promote complex assembly, Pol II pausing, and ultimately gene silencing. In the control library, Polycomb genomic regions, marked by the H3K27me3 histone modification, were shown to be coated with caRNAs, with

enrichment at the H3K27me3 peak center (Figure 2.5f). However, these H3K27me3 peak region-associated caRNAs were extremely vulnerable to RNase A treatment, with the aggregated peak nearly disappearing in the RNase-treated library (Figure 2.5f).

We first asked what caRNA species were significantly removed by RNase A. To this end, for each gene, we calculated the ratio between caRNA abundance associated with the H3K27me3 peak center 200bp and the whole library for the same gene, which we defined as the peak enrichment ratio. We calculated the peak enrichment ratio for all genes in both the RNase-treated and control libraries (Figure 2.6a). We expected that any Polycomb-associated RNAs sensitive to RNase treatment would have a higher peak enrichment ratio in the control library than in the RNase library. Our data suggest genes with the top peak enrichment ratios in the control library include: KCNQ1, FBRSL1, METRN, MACROD1, KCNQ1OT1, and APC2, all of which have low peak enrichment ratios in the RNase library, suggesting selective depletion by RNase A. Among these genes, KCNQ1OT1 is famously known for recruiting the PRC2 complex to silence genes in the KCNQ1 domain in a lineage-specific manner (Pandey et al., 2008). Although no studies have nominated the other genes as PRC2-associating RNAs, many of them were enriched with either GC-rich or (GA)n motifs in their RNase-insensitive H3K27me3-targeting transcripts (compared to the control library), which have previously been reported as PRC2 recognition motifs on RNA transcripts (Rosenberg et al., 2021) (Figure 2.6c, e).

We further defined Polycomb-associated caRNAs as the group of caRNAs with higher H3K27me3 peak enrichment ratios in the control library than in the RNase library. With this method, we identified 3,191 Polycomb-associated caRNAs. To test if these caRNAs were consistent with previously identified genes, we overlapped our candidates with RedChIP-identified PRC2 component EZH2-associated RNAs (n=3,021) in human embryonic cells (Figure

2.6b). Association tests suggest the overlap is significant (Odds Ratio: 60.1, p-value < 2.2e-16). Together, these results suggest that Polycomb-associated caRNAs can be identified using RT-iMARGI and untreated genome-wide RNA-chromatin interaction data. Polycomb-associated caRNAs are less protected from RNase digestion compared to some TF-associated caRNAs, possibly due to distinct RNA-protein-DNA interaction mechanisms. A conceptual model could be that caRNAs directly bound by proteins (complexes) are more resistant to RNase digestion (DNA-protein-RNA model), while RNAs are more sensitive to RNase if RNA acts as a bridge between DNA and proteins (DNA-RNA-protein model). This RNA bridge model has been validated in the PRC2 recruitment mechanism in earlier work (Long et al., 2020).

Applying the same idea, we identified transcripts enriched at H3K79me2 sites but sensitive to RNase, including genes BRD4, JARID2, ANKRD11, AKAP8, KANSL1, and others (Supplementary Figure 2-5a). RNAs enriched in these transcripts contained motifs such as spaced GC motifs for BRD4 or GA-rich motifs for JARID2 and ANKRD11 (Supplementary Figure 2-5b-d). Notice that RNase did not reduce RNA end abundance uniformly across transcripts, suggesting modularity of functions.

Altogether, these results suggest that RT-iMARGI, together with the control iMARGI data, could identify caRNAs specifically associated with certain histone modifications in normal conditions but sensitive to RNase treatment (including H3K27me3, H3K36me3, and H3K79me2). These sets of RNAs could potentially be involved in the deposition of these modifications through an RNA-mediated mechanism.

Discussion

The present an RNase-mediated RNA-chromatin interaction profiling approach, to globally investigate the protected RNA-chromatin interactome genome-wide. Our results reveal several key insights into the nature and function of chromatin-associated RNAs (caRNAs).

Historically, when chromatin-associated RNAs were first discovered, the prevailing belief was that they represented a major outcome of nascent transcription, with the newly transcribed RNAs attached close to their transcription start sites (Holmes et al., 1972). Recent studies also suggested that the localization pattern of caRNAs is often determined by their transcription start sites and expression levels (Limouse et al., 2023). However, in other cases, some RNAs can act in trans and play roles at distant genomic sites (Clemson et al., 2009; Tripathi et al., 2010). Our study shows that long-distance RNA-chromatin interactions are preferentially preserved after RNase treatment, while short-range interactions potentially arising from nascent transcript diffusion are depleted. This suggests that caRNAs involved in long-range targeting are more likely to be functionally relevant and protected, potentially by RNA-binding proteins. The enrichment of extremely proximal RNA-genome interactions likely reflects polymerase protection of nascent transcripts. Our study demonstrates that RNase treatment can remove the majority of caRNAs contributed by nascent transcripts, while the remaining RNase-inaccessible fraction usually involves long-range interactions.

Our study also demonstrates that RNase insensitive caRNAs exhibit distinct genomic features compared to the total caRNA population. They are enriched for intergenic transcripts, certain repeat classes (e.g. satellites, rRNA), and disease-associated lncRNAs. Furthermore, these RNase-inaccessible caRNAs display greater evolutionary sequence conservation, particularly at known functional RNA domains like Sm sites and H/ACA boxes. This evolutionary constraint implies functional importance for these caRNA molecules. RNase treated iMARGI pinpoints the localization preference of RNase-inaccessible caRNAs to regulatory genomic regions like promoters, transcription factor binding sites (TFBS), and histone variant (H2A.Z) sites. A subset also maps to heterochromatic lamina-associated domains. Integration with epigenomic data

suggests interplay between these caRNAs and chromatin regulatory machinery like transcription factors and Polycomb complexes.

RNase treated iMARGI magnifies the specific associations between caRNAs and regulatory factors like transcription factors. The caRNA binding signals at many transcription factor binding signals are depleted genome-wide after RNase treatment, except at cognate TFBS where caRNA enrichment persists. This concentration effect implies these TFBS-localized caRNAs are protected, potentially by direct binding to the transcription factors themselves. Supporting this, many transcription factors have been suggested to interact with caRNA, such as EP300 (Bose et al., 2017) , SOX2 (Z. E. Holmes et al., 2020), BRCA1 (Vohhodina et al., 2021) and many others. Notably, a recent landmark study (Oksuz et al., 2023) has revealed that the ability to bind RNA is a widespread property of transcription factors, with at least half of transcription factors exhibiting direct RNA-binding capabilities mediated by a previously unrecognized arginine-rich RNA-binding domain (ARM-domain). This suggests that the specific caRNA-transcription factor associations observed here likely represent a general phenomenon where transcription factors use RNA-binding domains to interact with chromatin-associated RNAs at their regulatory sites across the genome. The ability of RNase treated iMARGI to pinpoint these specific RNA-protein interactions provides a powerful means to delineate the functional RNA components involved in transcriptional regulation by diverse transcription factors.

In summary, RNase treated iMARGI enables high-resolution mapping of the functional RNA-chromatin interactome by removing diffusive RNA signals. The RNase-inaccessible interactions pinpoint regulatory RNAs, their genomic localizations, and mechanistic associations with chromatin factors. These findings nominate novel RNA components of nuclear regulation and provide a rich resource for investigating the multi-faceted roles of caRNAs in genome

organization and gene expression control. Future studies could leverage RNase treated iMARGI to probe RNA-chromatin interactions in diverse cellular contexts and link specific RNA molecules to regulatory phenomena. Additionally, integrating RNA perturbations or RNA-binding protein depletions could directly test the functional requirements for caRNAs in chromatin regulation. Overall, this work establishes a powerful approach for deciphering the RNA-centric nodes of nuclear regulation.

Figure

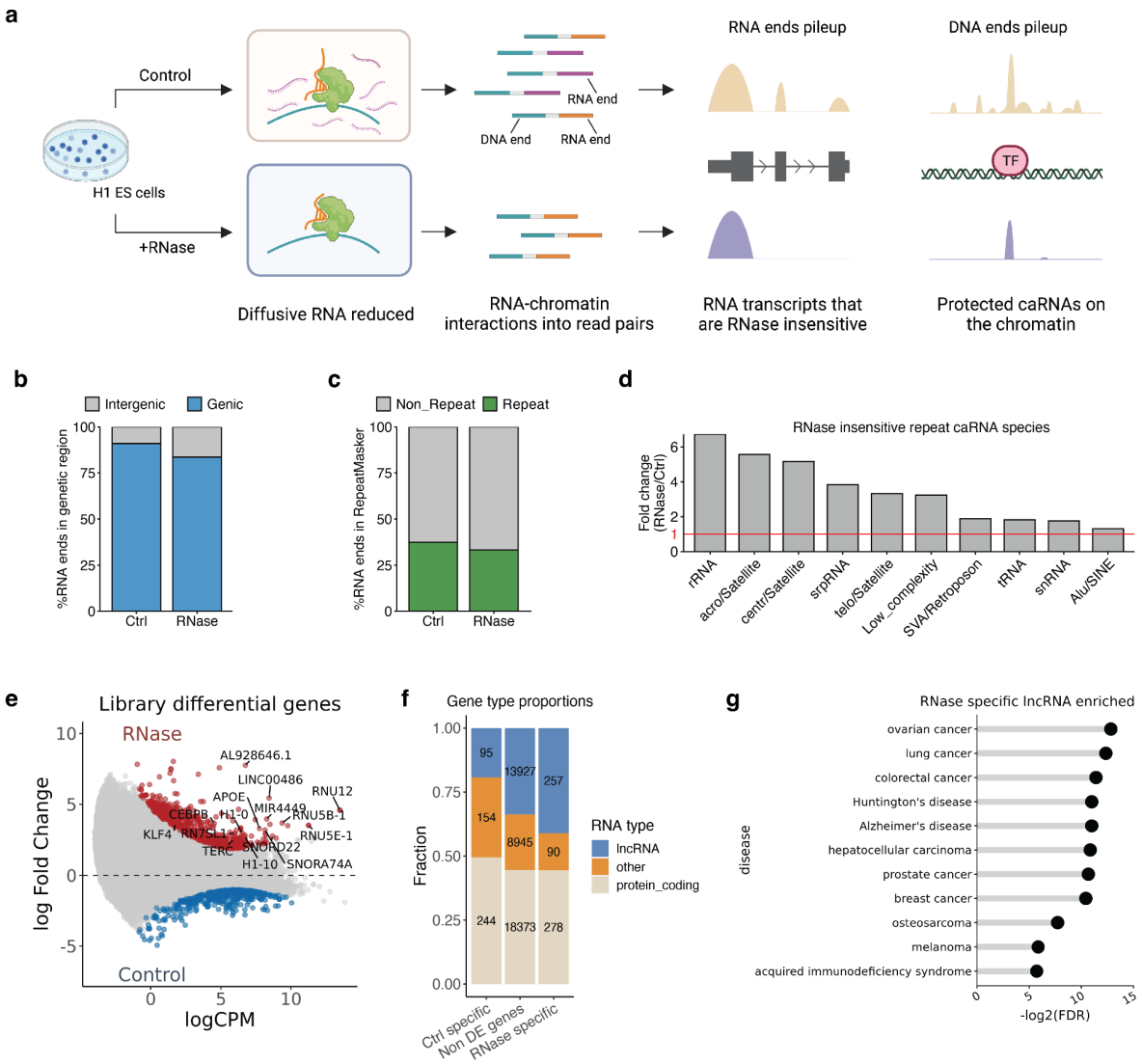


Figure 2.1: Overview of RNase treated iMARGI library.
a) Illustration of RNase treated iMARGI. RNase treatment removes diffusive RNA-chromatin interactions. b) Proportion of RNA reads landed in genic region (blue) and intergenic region (grey) for CTRL iMARGI, RNase treated iMARGI and RNAseq libraries. c) Proportion of RNA reads landed in repeat region (green) and non-repeat region (grey) for CTRL iMARGI, RNase treated iMARGI and RNAseq libraries. d) Enrichment of repeat RNA classes between RNase and CTRL library. e) Differential gene expression between CTRL and RNase treated iMARGI library. f) Proportion of lncRNA, protein coding RNA in CTRL, RNase treated iMARGI differentially expressed, and non-differentially expressed genes. g) RNase treated iMARGI specific lncRNA enriched diseases.

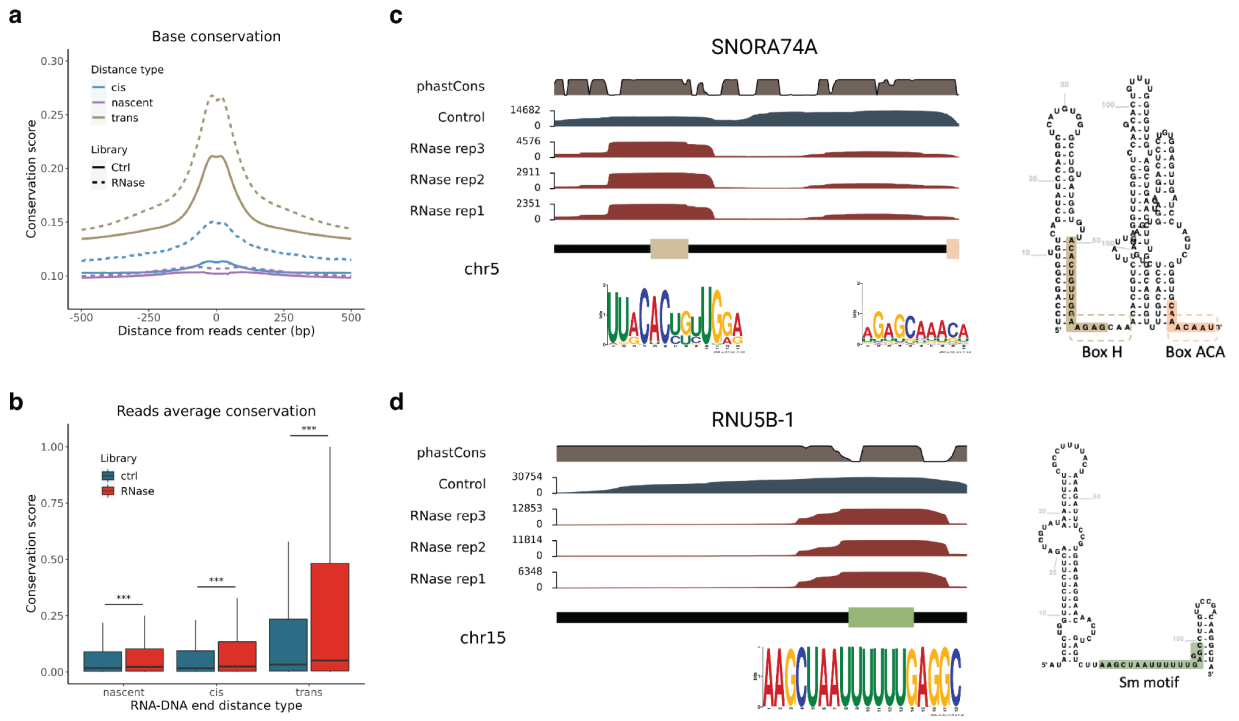


Figure 2.2: Sequence conservation of RNase inaccessible RNA ends.
a) Sequence conservation (PhaseCons 100 score) of each read's 1000bp flanking region. All pairs in the control and RNase library were divided into three groups based on their RNA and DNA end genomic distance. Nascent pairs (<1k bp), cis pairs (1k bp ~ 2Mb), trans pairs (>2Mb and interchromosomal). b) Average read conservation score in each library and each group. Differences of conservation score in each group were tested using t-test. c) RNA ends coverage of SNORA74A gene in three replicates of RNase libraries and one combined control library. Sequence conservation score phastCons100 were shown on the top track. Higher value indicates higher level of conserveness. Secondary structure of SNORA74A was computed by R2DT. RNA end reads enriched regions were highlighted on the secondary structure. Enriched binding sequence motif was called by MEME. d) Same information as C for gene RNU5B-1. Green box corresponds to the Sm motif.

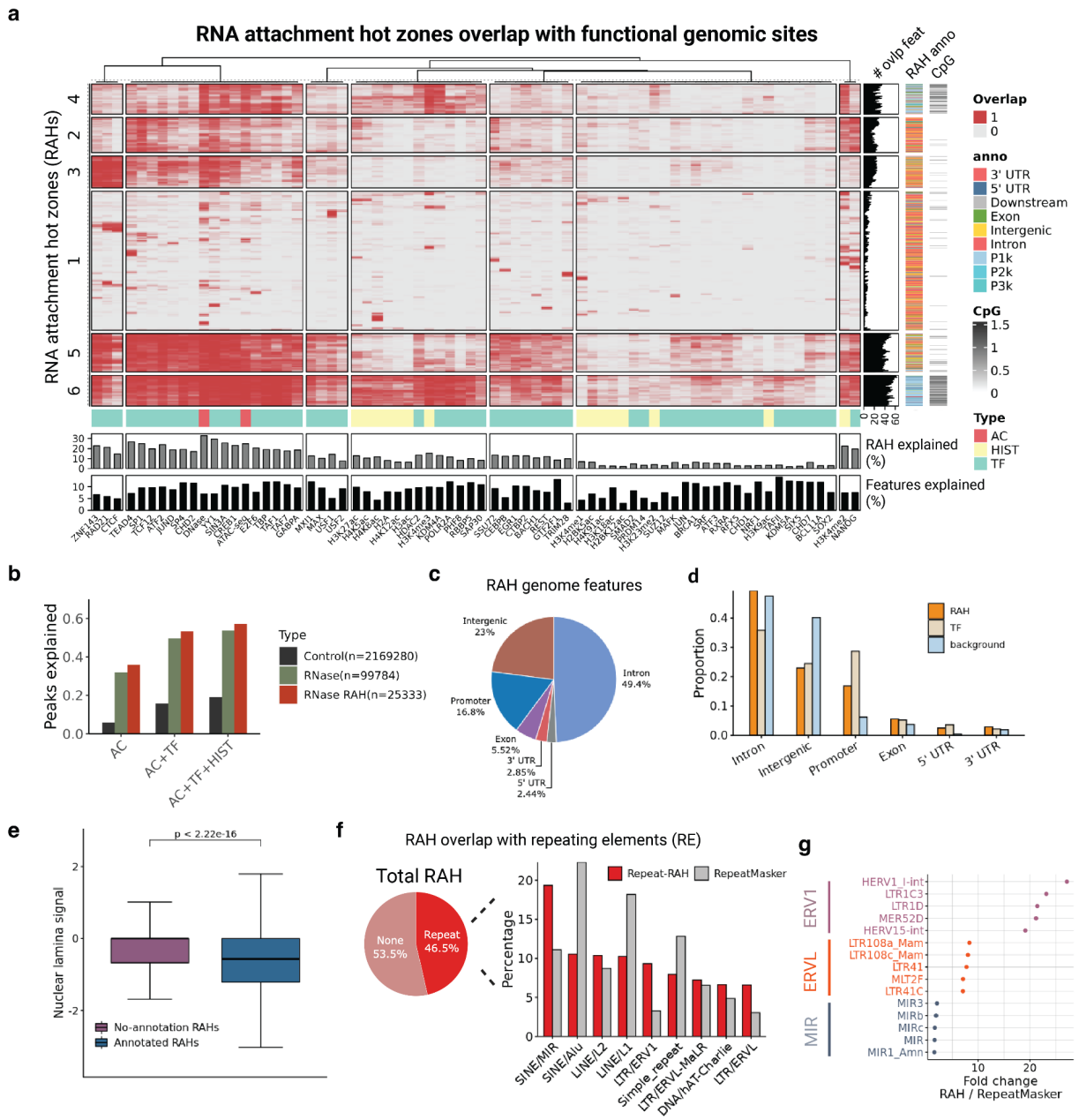


Figure 2.3: RNAs interacting with functional genomic regions genome wide.
a) Overlap of RNA attachment hot zones (RAHs) with transcription factor binding sites (TFBS), histone modifications, and chromatin accessibility data. Binary matrix shows overlap between 14,468 RAHs and 72 genomic features. Tracks below indicate feature type, proportion of peaks overlapping each feature, and proportion of each feature overlapping RAHs. Sidetracks show number of features per RAH, RAH genomic locations, and CpG content. b) Proportion of DNA end peaks explained by accessibility, TFBS, and histone modifications. c) Genomic feature composition of RAHs. d) Comparison of genomic features between RAHs, TFBS in H1 cells, and random background. e) Nuclear Dam-LaminB1 signals of RAHs grouped by overlap with genomic features. f) Proportion of RAHs overlapping RepeatMasker repetitive elements and categories. g) Top repetitive element classes enriched in RAHs compared to RepeatMasker.

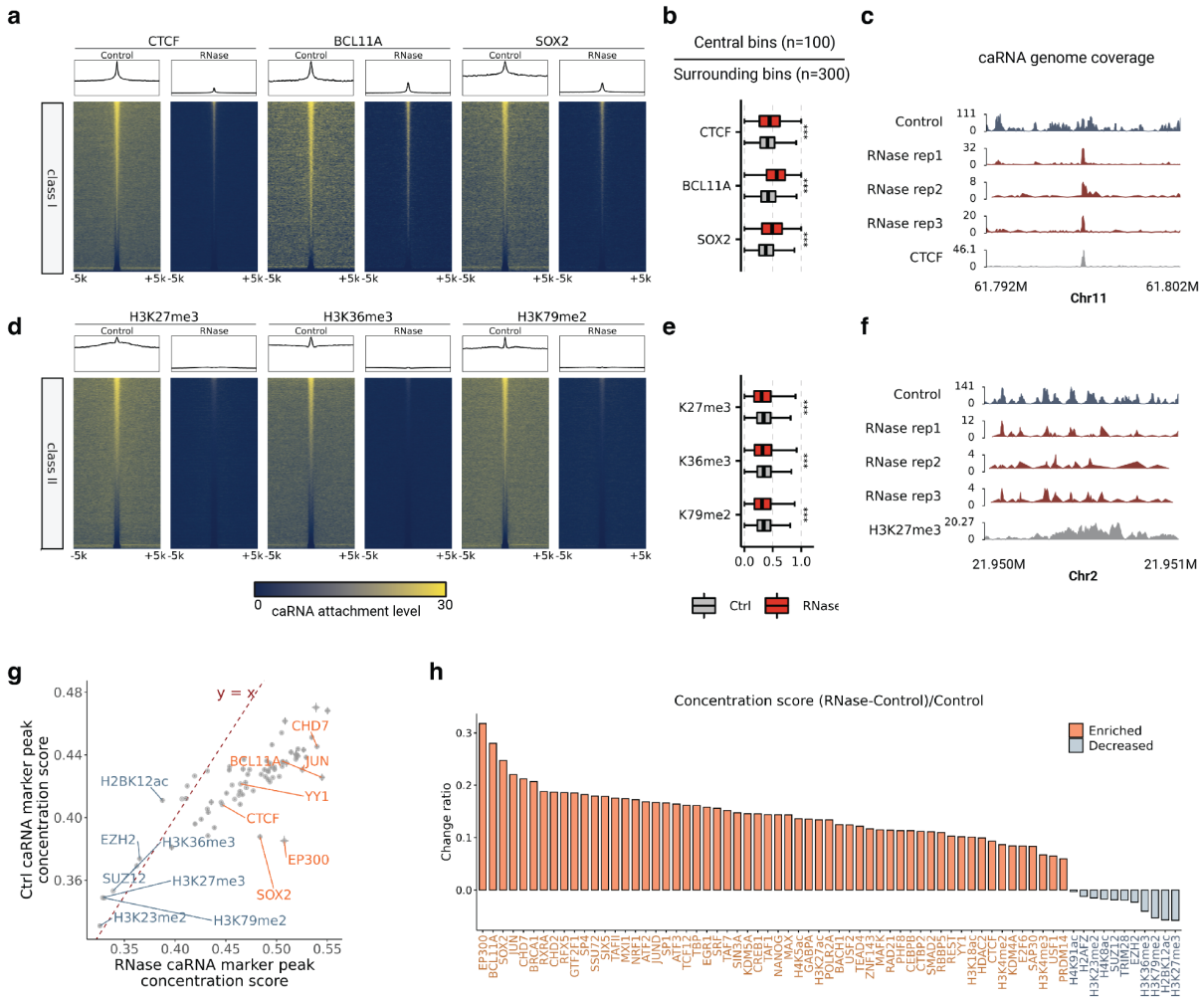


Figure 2.4: RNase treatment magnifies stringent interaction of caRNA-TF interactions.
a) Heatmap and summary plot for CTCF, BCL11A and SOX2. Rows represent all peaks of a TF. Chip-seq data annotated from Cistrome database H1 cell line. Columns include 1000bins represent 10kb surrounding regions from the center of a chip-seq peak. Color maps to the number of caRNAs interact with each bin. Summary plot shows the average coverage across all peaks for a transcription factor. Class 1 transcription factor have stronger caRNA – TF peak interaction after RNase treatment. b) Similar to A, showing peak coverage for histone modifications H3K27me3, H3K36me3 and H3K79me2. Class 3 histone modifications have diminished signals in RNase than in control library. c) Boxplot of concentration score in Control or RNase library. The score is defined as peak central 1000bp region caRNA attachment level divided by peak central 3000bp caRNA attachment level. d) Genome track view showing caRNA attachment level at TF peak or histone modification peak. e) Scatter plot of concentration in RNase or control library for 76 markers. f) Barplot showing concentration score changes (RNase-control)/control for 69 markers that are invariant of the definition of concentration score.

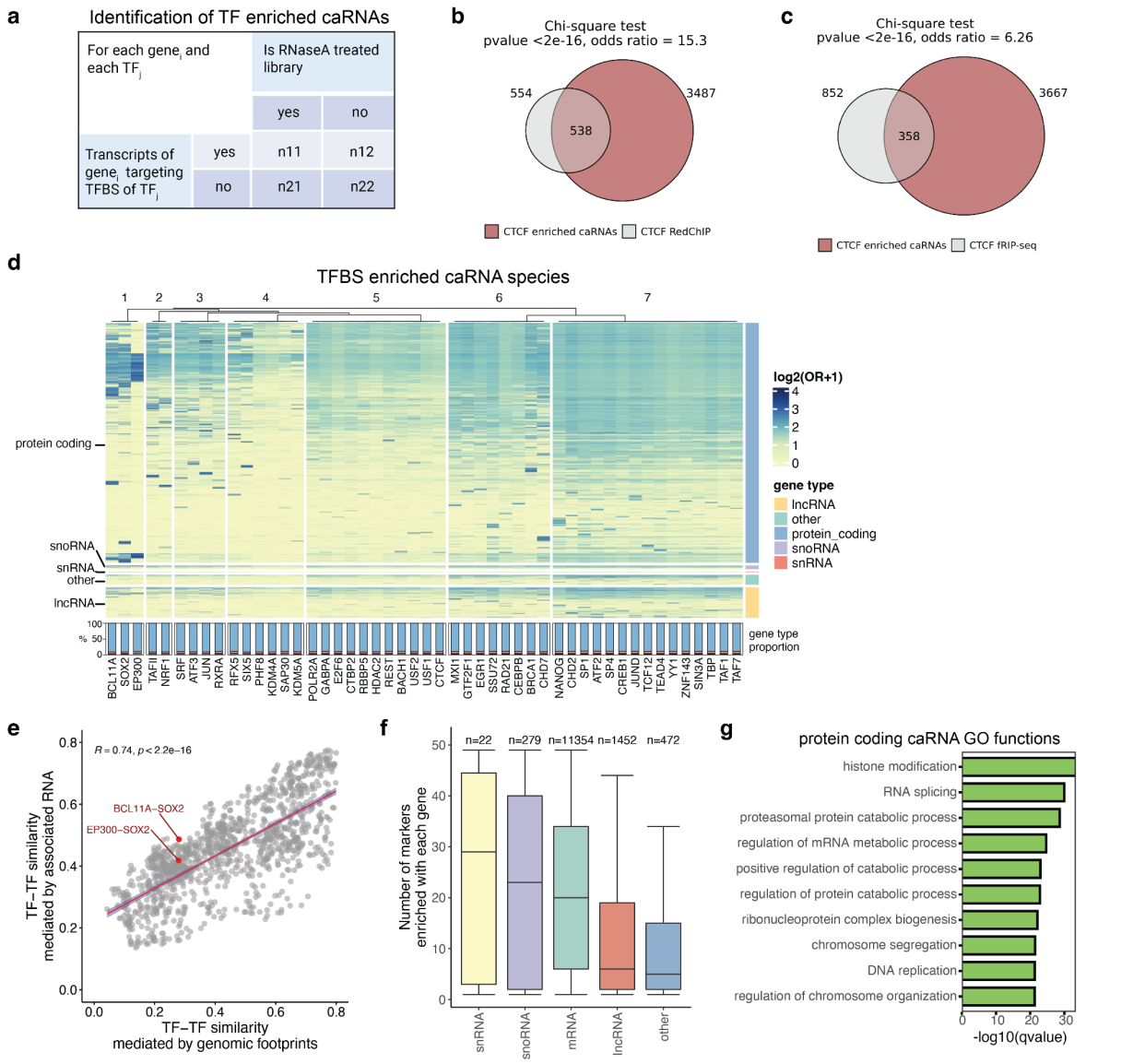


Figure 2.5: Transcription factor binding sites enriched caRNA species.
a) Contingency table of association test to identify TFBS enriched caRNA species. b) Venn plot showing the overlap between CTCF enriched caRNAs (in H1 cells) and RedChIP identified CTCF associated caRNAs (in K562 cells). c) Venn plot showing the overlap between CTCF enriched caRNAs (in H1 cells) and fRIP-seq identified CTCF associated caRNAs (in K562 cells). d) Heatmap showing each TF enriched caRNAs with color mapped to odds ratio from association test. Columns are clustered using hierarchical clustering based on Jaccard distance. Rows are genes split by gene type. Bottom track showing the proportion of each RNA species among all enriched caRNAs for each TF. e) Correlatioon between TF-TF similarities measured by TF genome footprints (number of TFBS in each non-overlapped 100kbp bin on the genome) or enriched RNA species. f) Boxplot showing the number of TFs each RNA species associated with. g) Top enriched GO terms of all mRNAs that enriched with any TF examined.

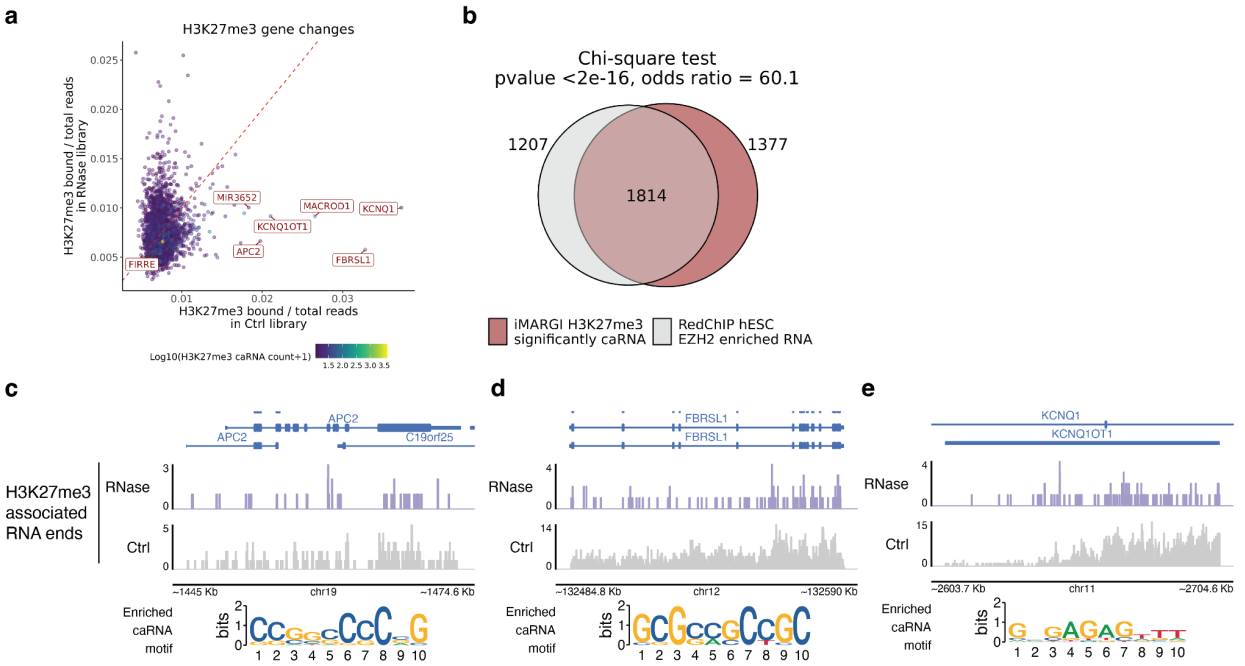


Figure 2.6: Characterizing caRNA species significantly decreased at H3K27me3 regions.
a) H3K27me3 associated caRNA abundance versus background abundance in the control (x axis) or RT-iMARGI (y-axis). b) Venn plot showing the overlap between H3K27me3 significantly associated RNA species in the control library and PRC2 complex component EZH2 associated caRNAs (identified using RedChIP) in hESC cells. c-e) RNA ends abundance transcribed from three genes (APC2, FBRSL1, KCNQ1OT1) in RT-iMARGI or control iMARGI libraries. Control library RNA ends enriched motifs (RNase sensitive motifs) were called using HOMER with control RNA ends as background.

Methods

RNaseA treatment

H1 cells were harvested from cell culture plate and aliquoted cell suspension to 10 million H1 cells per 1.5 mL tube. Wash the cells with 1 mL 1X PBS and centrifuge at 500 X g for 3 min at RT. Then, cells were gently permeabilized by resuspending cell pallets with 0.01% PBST (TritonX-100 in PBS) and treated for 5 min at RT. After permeabilization, cells were treated with 200 µg/mL RNase A (Thermo Fisher Scientific, Cat# EN0531) on rotator for 10 min at RT. The treated cells were fixed with 4% formaldehyde (Thermo Fisher Scientific, Cat# 28906) for immunofluorescence imaging. For Hi-C and iMARGI library generation, the treated cells were fixed with 1 mL 1% formaldehyde on the rotator for 10 min at RT. Then, the reactions were terminated with 250 µL 1M glycine on the rotator for 10 min at RT. The treated sample was centrifuged at 2000Xg for 5 min at 4°C and washed with 1 mL cold 1X PBS.

Library specific genes identification

We first summarized the number of RNA ends mapped to each gene in control and RNase A treated libraries respectively. We filtered out genes that have zero counts across all replicates. We then applied edgeR [PMID: 19910308] to identify the differentially expressed genes between all replicates under the null hypothesis that the fold change of gene abundance between the two conditions is zero. For RNase library the threshold is adjusted p-value cutoff at 10-6, a log fold change increase at 5. For the control library the adjusted p-value cutoff at 0.05, a log fold change decreases at 5. The different thresholds chosen in the two libraries due to the unbalanced expression profile for genes, namely, many genes only have transcripts in the control library.

Gene set disease ontology term enrichment analysis

We derived a disease ontology reference by assigning each disease term with their associated genes annotated from the LncRNADisease database. For each of these DE genes related disease-gene sets with size larger than 10 (including more than 10 genes associated with each

disease term), we applied the hypergeometric test to see if any disease ontology term is enriched with RNase specific DE genes. Enriched disease terms were selected using FDR < 0.05.

RNA attachment hotzones identification

To systematically identify genomic regions enriched with RNA transcripts in the RNase library, we employed the following approach. First, peaks were called using DNA end coverage across the genome to reduce the influence of weak signals. To avoid peaks formed due to nascent transcription, only read pairs with an RNA-DNA end distance greater than 1kb were considered. Peak calling was performed using MACS2 on the combined replicates of the RNase library. The fragments in peaks ratio (FRiP) was calculated to assess the signal-to-noise ratio. To focus on peaks with high read coverage, a peak max depth threshold of 15 was applied, and peaks meeting this criterion were designated as RNA Attachment Hot zones (RAHs).

Transcription factor similarities measured by colocalization similarities.

TF-TF similarity based on ChIP-seq data localization was calculated as follows: ChIP-seq peak files for each TF were imported as genomic ranges. The human genome was tiled into 100kb bins and for each TF, the number of overlapping peaks per bin was counted, generating a bin x TF matrix. Pairwise TF-TF similarity was then computed as the correlation between the respective rows of this matrix using the simil function from the proxy R package.

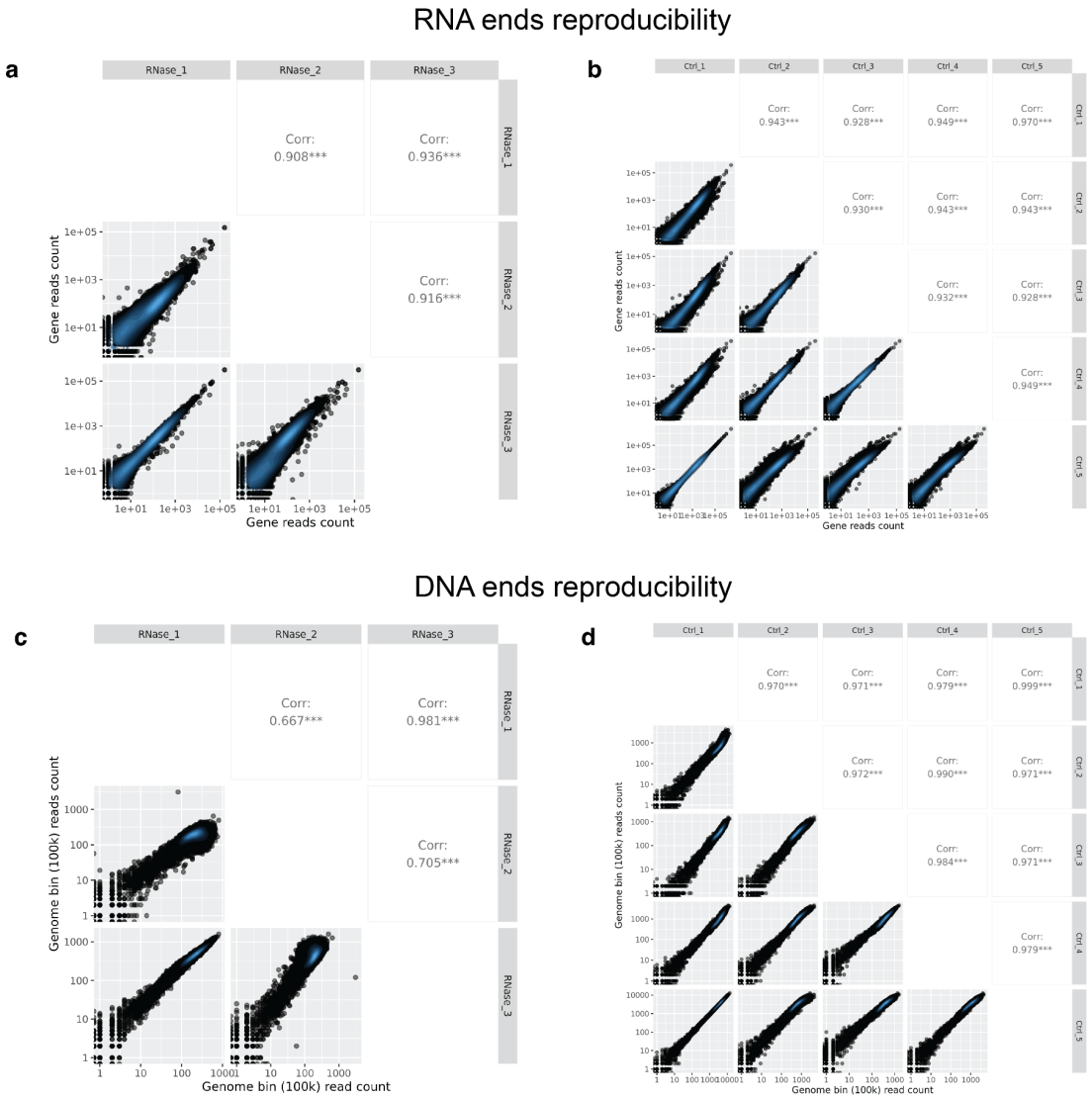
RIPseq data process and enriched target selection methods

Without replication we use the chi-square test to select target antibodies enriched over IgG. The contingency table is as follows: is geneA versus is target antibody. We chose fdr < 0.01 and odds ratio >1 as enriched genes.

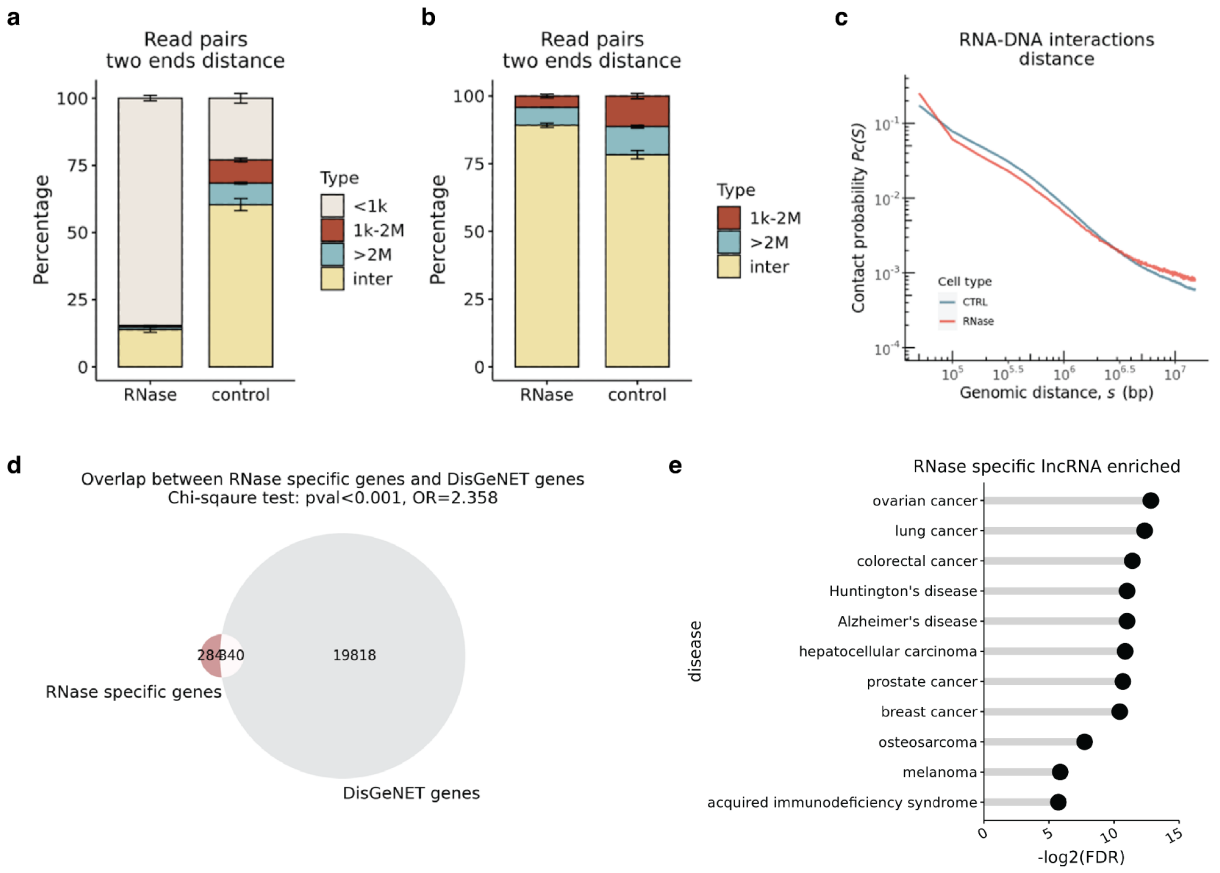
Acknowledgements

Chapter 2, in full, is currently being prepared for submission for publication of the material. Xingzhao, Wen; Sheng, Zhong. The dissertation author was the primary investigator and author of this paper.

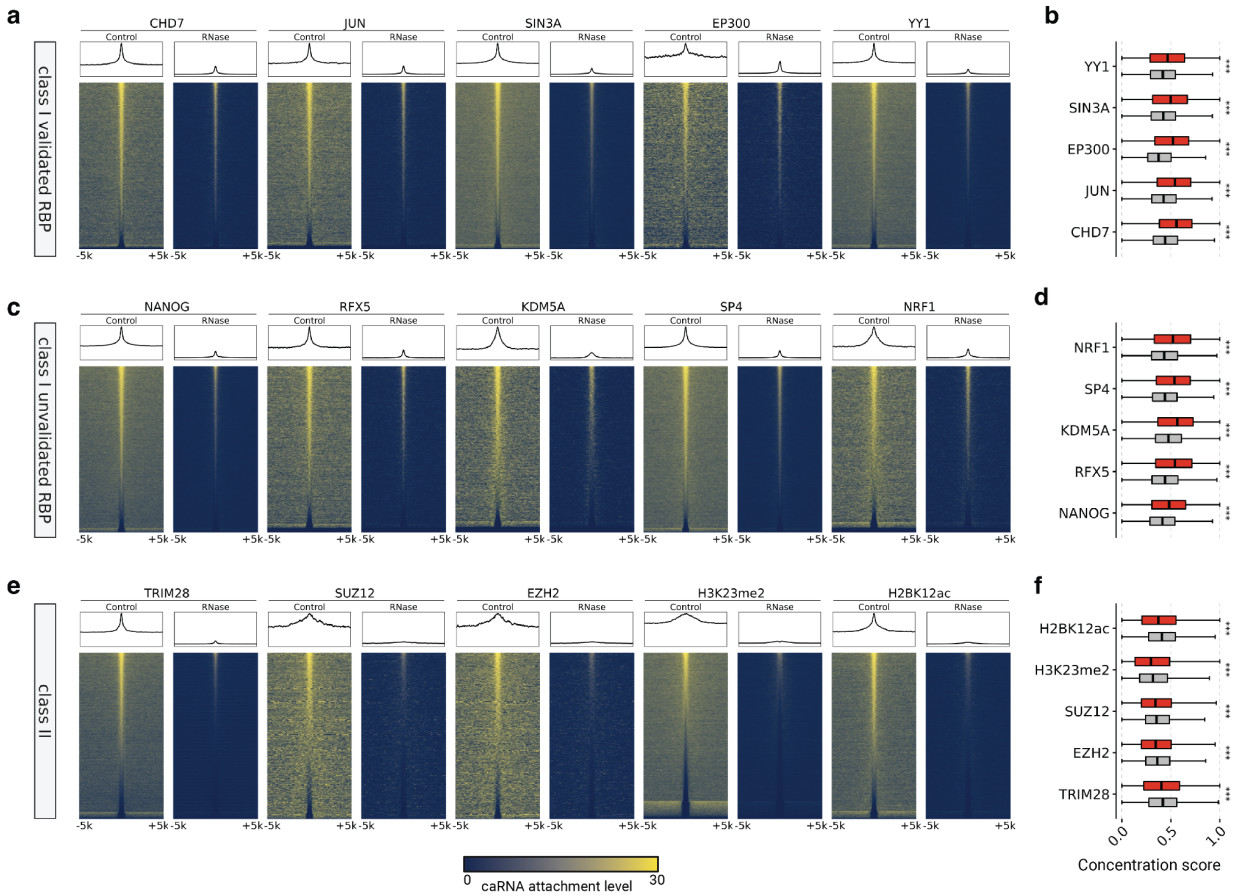
Supplementary information



Supplementary Figure 2-1: Library reproducibility.
iMARGI pairs with RNA and DNA end distance larger than 1000bp used. a) Reproducibility of iMARGI RNA ends read counts for each gene between three RNase replicates. iMARGI RNA ends are counted in a strand specific manner. Spearman correlation was calculated with p-value. b) Similar to A, measured between control replicates. c) Reproducibility between iMARGI DNA ends mapped to genomic bin. 100kbp bins are derived by sliding the whole genome without overlap. Spearman correlation was calculated with p-value. d) Similar to c), measured between control libraries.

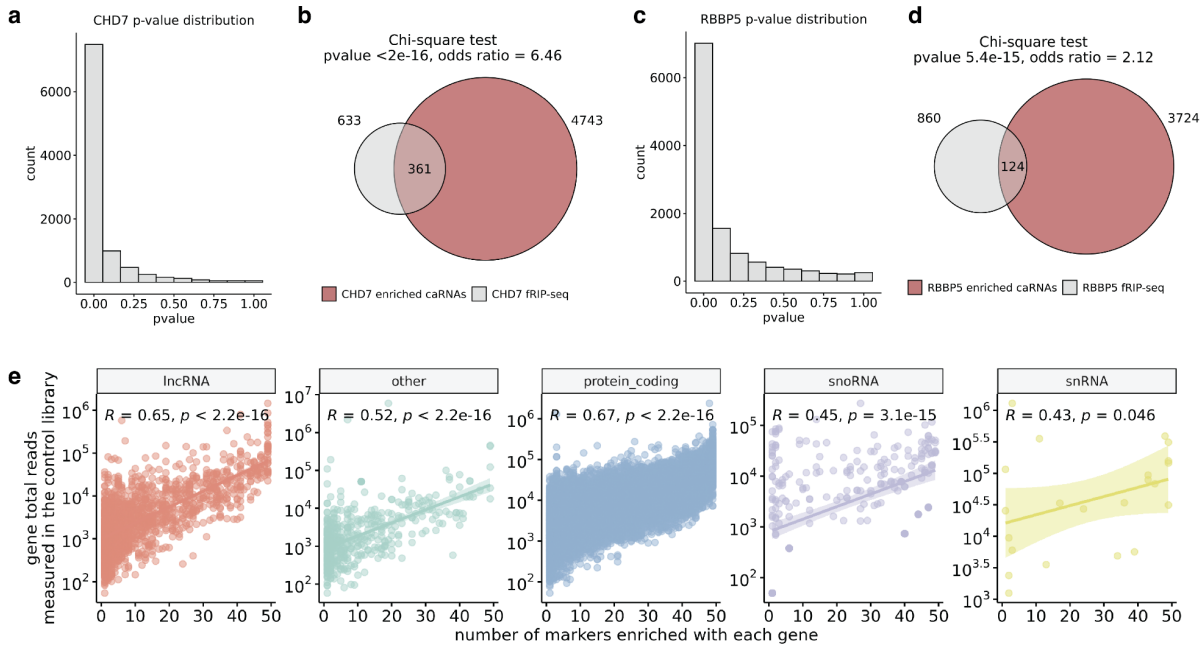


Supplementary Figure 2-2: iMARGI RNA, DNA ends distance distribution.
a) Barplot showing the proportion of read pairs that with RNA, DNA end distance belongs to <1kbp, 1kbp~2Mbp, >2Mbp but within one chromosome and finally interchromosomal read pairs. Error bar of the proportion was derived from library replicates. b) Similar to A, but without <1kbp category. c) Genomic distance versus contact frequency for all RNA, DNA read pairs with two ends distance longer than 1kbp.

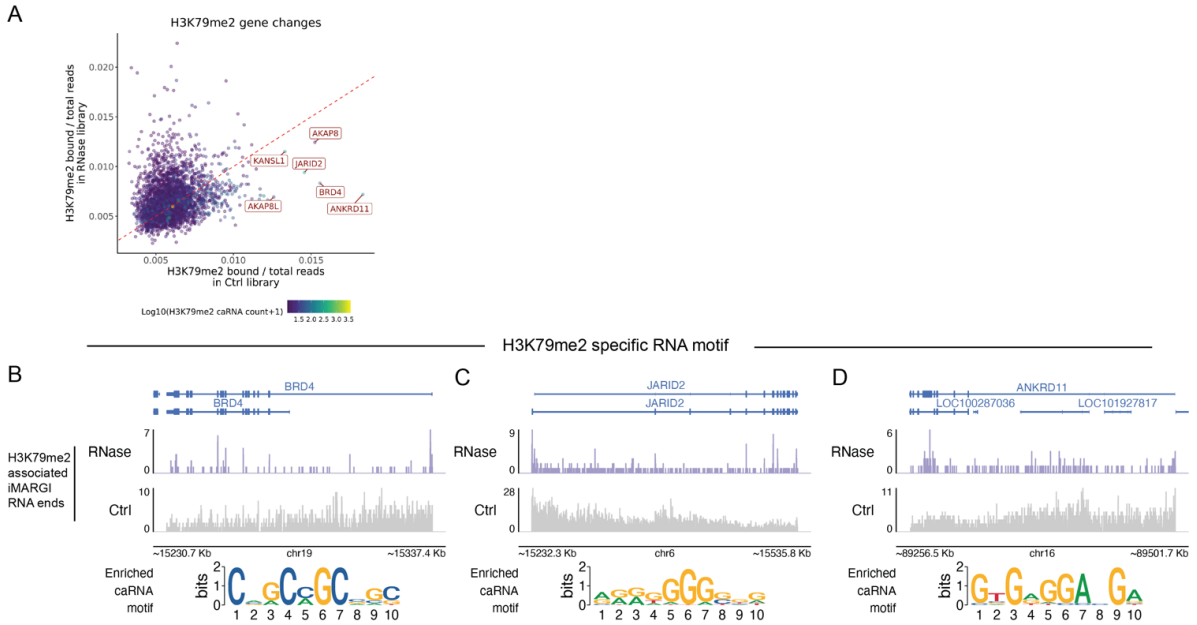


Supplementary Figure 2-3: RNase treatment magnifies stringent interaction of caRNA-TF interactions.

a) Heatmap and summary plot for five class I TFs (stronger caRNA enrichment at TFBS after RNase digestion) that have been suggested to have RNA binding properties: CHD7, JUN, SIN3A, EP300, YY1. Rows represent all peaks of a TF. Chip-seq data annotated from Cistrome database H1 cell line. Columns include 1000bins represent 10kb surrounding regions from the center of a chip-seq peak. Color maps to the number of caRNAs interact with each bin. Summary plot shows the average coverage across all peaks for a transcription factor. b, d, f) Boxplot of concentration score in Control or RNase library. The score is defined as peak central 1000bp region caRNA attachment level divided by peak central 3000bp caRNA attachment level. c) Similar to a), showing five TFs have enriched caRNA binding after RNase digestion but have not been reported has RNA binding property. e) Similar to a), showing five markers including TF TRIM28 or histone modifications whose DNA sites showing depleted caRNA level after RNase digestion.



Supplementary Figure 2-4: RI-caRNA gene features.
a) Histogram of p-value distribution from association test. b) Overlap between CHD7 TFBS enriched caRNA species and CHD7 fRIP-seq identified CHD7 protein targeting RNAs. c) Similar to b) but for RBBP5. e) Relationship between number of caRNA associated transcription factors and gene total transcribed read counts in untreated control library.



Supplementary Figure 2-5: RNase sensitive caRNAs at H3K79me2 modifications.
a) H3K79me2 peak center 200bp associated caRNA abundance over the whole library for each gene in RT-iMARGI and the control library. Red dashed line indicates y=x, below which we select genes that are H3K79me2 enriched in normal condition but sensitive to RNase treatment. Point color maps to the total reads count of the gene in RT-iMARGI library. b-d) RNA ends abundance transcribed from three genes (BRD4, JARID2, ANKRD11) in RT-iMARGI or control iMARGI libraries. Control library RNA ends enriched motifs (RNase sensitive motifs) were called using HOMER with control RNA ends as background.

Supplementary Table 2-1: iMARGI libraries overview.

Library	**Total Read pairs(de-multiplexed)**	**Total dedupped mapped reads (both ends mapped)**	**MAPQ>=30 and unique mapping**
RNase 1	719,424,571	409,443,906	48,926,803
RNase 2	950,806,144	344,012,782	50,633,451
RNase 3	1,519,702,193	835,597,061	92,517,185
Control 1	1,009,524,877	783,821,122	134,185,427
Control 2	297,811,325	184,538,199	25,330,555
Control 3	330,531,513	202,604,067	16,224,875
Control 4	1,004,910,451	587,454,453	42,362,018
Control 5	922,845,725	726,916,769	123,383,403

Chapter 3 Single-cell multiplex chromatin and RNA interactions in ageing human brain

Abstract

The dynamically organized chromatin complexes often involve multiplex chromatin interactions and sometimes chromatin-associated RNA (caRNA) (Calandrelli et al., 2023; Li and Fu, 2019; Quinodoz et al., 2021). Chromatin complex compositions change during cellular differentiation and aging, and are expected to be highly heterogeneous among terminally differentiated single cells (Flyamer et al., 2017; Nagano et al., 2017; Ramani et al., 2017; Stevens et al., 2017). Here we introduce the Multi-Nucleic Acid Interaction Mapping in Single Cell (MUSIC) technique for concurrent profiling of multiplex chromatin interactions, gene expression, and RNA-chromatin associations within individual nuclei. Applied to 14 human frontal cortex samples from elderly donors, MUSIC delineates diverse cortical cell types and states. We observed the nuclei exhibiting fewer short-range chromatin interactions are correlated with an "older" transcriptomic signature and with Alzheimer's pathology. Furthermore, the cell type exhibiting chromatin contacts between cis expression quantitative trait loci (cis eQTLs) and a promoter tends to be the cell type where these cis eQTLs specifically affect their target gene's expression. Additionally, the female cortical cells exhibit highly heterogeneous interactions between the XIST non-coding RNA and Chromosome X, along with diverse spatial organizations of the X chromosomes. MUSIC presents a potent tool for exploring chromatin architecture and transcription at cellular resolution in complex tissues.

Introduction

The three-dimensional (3D) folding of the genome is known to exhibit dynamic changes during cellular differentiation processes and demonstrates heterogeneity among terminally differentiated single cells (Flyamer et al., 2017; Nagano et al., 2017; Ramani et al., 2017; Stevens et al., 2017). While its regulatory role in the expression of specific genes has been well-established

(Dekker et al., 2017; Lupiáñez et al., 2015; Sun et al., 2018), the extent to which the 3D genome structure impacts the expression of most genes remains a topic of debate (Rao et al., 2017). Given the pronounced heterogeneity observed in chromatin structure and gene expression across individual cells (Arrastia et al., 2022; Chen et al., 2015; Takei et al., 2021a), a comprehensive understanding of the relationship between 3D genome structure and gene expression at the single-cell resolution is necessary. Therefore, developing single-cell multimodal technologies capable of simultaneously profiling chromatin conformation and gene expression is instrumental for elucidating these intricate relationships.

Single-cell multi-omic technologies enabled joint analysis of chromatin conformation and gene expression (Chen et al., 2019; Z. Liu et al., 2023; Ma et al., 2020; Plongthongkum et al., 2021; Zhu et al., 2020) (Supplementary Table 1, Supplemental Note 1). Despite these technical advancements, the simultaneous profiling of multiplex chromatin interactions (co-complexed DNA sequences), gene expression, and RNA-chromatin associations from a single cell remain challenging. To fill this gap, we developed the Multi-Nucleic Acid Interaction Mapping in Single Cell (MUSIC) technique, which enables the simultaneous profiling of gene expression and co-complexed DNA sequences with or without co-complexed RNA at the single-cell level.

The architecture of chromatin can encompass both pairwise and multiplex chromatin interactions, highlighting the intricate nature of chromatin complexes (Arrastia et al., 2022a; Deshpande et al., 2022; Quinodoz et al., 2018; M. Zheng et al., 2019). ChIA-Drop has facilitated the mapping of multiplex chromatin interactions at single-complex resolution from bulk cells, revealing that multiplex chromatin interactions are prevalent in Drosophila (M. Zheng et al., 2019). The MUSIC technique expands the capability to evaluate the composition of pairwise and multiplex chromatin interactions in individual human cells at single-cell resolution.

In addition to DNA, chromatin complexes can also encompass RNA molecules, introducing another layer of complexity to chromatin architecture (Calandrelli et al., 2023; Quinodoz et al., 2021; Thakur and Henikoff, 2020). Chromatin-associated RNA (caRNA) has been shown to contribute to gene expression regulation (Lee, 2011; Quinodoz et al., 2021; Rinn and Chang, 2012). For instance, the accumulation of the XIST long noncoding RNA (lncRNA) on the X chromosome (XIST-chrX association) is crucial for the silencing of one of the two X chromosomes in female cells, a process known as X-chromosome inactivation (XCI). Various human tissues exhibit both shared and tissue-specific incomplete XCI genes, which are expressed from the silenced X chromosome (Tukiainen et al., 2017). Genes with incomplete XCI can display higher expression levels in females, potentially contributing to sex differences in disease susceptibility (Y. Yan et al., 2022). Recent advancements have enabled genome-wide mapping of RNA-chromatin associations in bulk cells (Bell et al., 2018; Bonetti et al., 2020; Li et al., 2017b; Quinodoz et al., 2021; Wu et al., 2019; Yan et al., 2019). With the application of MUSIC, we can now obtain RNA-chromatin association maps at the single-cell level. Utilizing MUSIC, we uncover cellular heterogeneity in XIST-chrX association levels (XAL) within the female cortex and explore the co-variation between XAL and chromatin interactions among the female cortical cells.

Results

Design and workflow

The development of the MUSIC technology is guided by three specific design goals, which collectively enable the joint profiling of gene expression, co-complexed DNA sequences, and RNA-chromatin associations from the same nucleus (Figure 3.1a). The first goal is to construct RNA and fragmented DNA into a single sequencing library and identify which RNA and DNA sequences originated from the same nucleus. This goal is achieved by labeling all the RNA and

the fragmented DNA in the same nucleus with a unique Cell Barcode. This Cell Barcode enables the identification and matching of RNA and DNA sequences originating from the same nuclei.

The second goal is distinguishing RNA inserts from DNA inserts in the sequencing library. To achieve this, distinct nucleotide sequences are used for the RNA Linker and the DNA Linker, which are ligated to the RNA and DNA molecules, respectively. These linkers are sequenced alongside the RNA and DNA inserts, enabling the differentiation of RNA and DNA molecules within the sequencing data. The third goal is to capture and identify DNA-DNA and RNA-DNA associations, including multi-way contacts. To achieve this, each molecular complex is labeled with a unique Complex Barcode. A molecular complex can encompass various combinations of DNA and RNA, including (1) an isolated RNA molecule, (2) an isolated DNA fragment, (3) multiple DNA fragments, (4) multiple RNA molecules, and (5) at least one RNA molecule and at least one DNA fragment. The Complex Barcodes together with the Cell Barcodes allow for the identification of co-complexed DNA and/or RNA in each cell.

The MUSIC workflow contains two major steps (Supplemental Note 2). The first step ligates the RNA Linker to the RNA molecules and the DNA Linker to the fragmented DNA and adds the Cell Barcodes (Supplementary Figure 3-1a-e). The second step adds a Complex Barcode to any RNA or DNA within the same molecular complex. The Complex Barcode consists of a 10X Barcode and a I7 Barcode (Supplementary Figure 3-1f, g). The final sequencing library is sequenced with a 28bp Read1 sequence, an 8bp index sequence that is the I7 Barcode, and a 150bp Read2 sequence (Extended Data Fig 1h). The 28bp Read1 corresponds to the 10X Barcode, which consists of a 16bp 10X GEM Barcode and a 12bp 10X UMI. Read2 contains the 3rd, 2nd, and 1st Cell Barcodes, the RNA Linker sequence or the DNA-Linker(bt) sequence, and the RNA insert or the DNA insert. It should be noted that each read pair is designed to capture only one insert, either

an RNA or a DNA insert, as the Read1 is dedicated to reading the 10X Barcode. This design differs from several ligation-based methods such as Hi-C (Lieberman-Aiden et al., 2009) and iMARGI (Wu et al., 2019; Yan et al., 2019) where each read pair represents two inserts.

RNA and chromatin interactions in ESC

We applied MUSIC to analyze a mixed population of H1 human and E14 mouse embryonic stem cells (ESC). The resulting mixed-species MUSIC library was sequenced on a NovaSeq platform, generating 3,067,956,666 read pairs. These read pairs resolved 533,233,368 uniquely mapped, non-duplicate, and barcode-complete (containing Cell Barcode, 10X Barcode, I7 Barcode, and either DNA Linker or RNA Linker) (UMNDBC) read pairs (Supplementary Table 3). Because as per the experimental design, each UMNDBC read pair contains only one DNA or RNA insert, we will refer to a UMNDBC read pair as a DNA read or an RNA read. This mix-species dataset revealed small species-mixing rates at the cell level and at the chromatin-complex level, supporting the ability of MUSIC to generate data at single-cell and single-complex resolutions (Supplementary Figure 3-3a, b, Supplemental Note 3).

We identified 2,546 human H1 cells from this dataset. Each H1 cell contained an average of 144,049 UMNDBC DNA reads, 11,384 UMNDBC RNA reads, corresponding to 7,036 DNA-only clusters (DD clusters), 232 RNA-only clusters (RR clusters), and 1,170 RNA-DNA clusters (RD clusters) (Figure 3.1b, c), which based on an established procedure (Arrastia et al., 2022a) can be resolved into 2,639,302,084 co-complexed DNA-DNA pairs, 7,089,720 co-complexed RNA-RNA pairs, 250,525,581 co-complexed RNA-DNA pairs (Figure 3.1d). In total, there were 18,144,410 non-singleton DD clusters, accounting for 55,670,578 DNA reads; 324,121 RR clusters, accounting for 835,184 RNA reads, and 2,401,392 RD clusters, accounting for 13,151,716 RNA reads and 216,515,595 DNA reads (Supplementary Figure 3-3c). Among the non-singleton DD clusters, 13,111,228 (72.26%) contained two DNA reads that correspond to

pairwise interactions, and 5,033,182 (27.74%) contained three or more DNA reads that respond to multiplex interactions (Supplementary Figure 3-3d, Supplementary Figure 3-4). In the RD clusters, 1,009,706 (42.05%) contained two reads, i.e., one DNA read and one RNA read, 783,709 (32.64%) contained 3-10 reads, and 607,977 (25.32%) contained more than 10 reads.

We compared MUSIC's ensemble DNA reads with Micro-C data that were generated from the same human H1 cell line and cultured under the same standard operating protocol as recommended by the 4D Nucleome Consortium. The contact map of MUSIC DD clusters reproduced the structures observed in the Micro-C derived contact map (4DN data portal: 4DNFI2TK7L2F (Krietenstein et al., 2020)) (Figure 3.2a), resulting in a similar distribution of the compartment scores across the genome (Supplementary Figure 3-5a). We compared the different-sized MUSIC DD clusters (lower triangles, Figure 3.2b-d) keeping the same Micro-C dataset as a reference (upper triangles, Figure 3.2b-d). At the Topologically Associating Domain (TAD) level, MUSIC's small DD clusters (2-10 DNA reads per cluster) primarily contained contacts within TADs (50 Kb resolution) (Figure 3.2b). MUSIC's middle-sized (11-50 reads per cluster) and large clusters (51-100 reads per cluster) recapitulate the TADs and the nested TAD structure (Figure 3.2c, d). Large clusters revealed more contacts between the nested TADs within a larger TAD (Arrows, Figure 3.2d).

To visualize the clusters, we plotted each cluster in a row, with every DNA read of this cluster aligned to their respective genomic coordinates. We ordered the clusters by the genomic coordinates of their leftmost DNA reads. This way, we created a stacked map of the clusters (Figure 3.2g, h). By comparing the 2-D contact map based on ensemble MUSIC data (Figure 3.2f) to the stacked maps, we observed that pairwise interactions (clusters with two DNA reads) alone poorly reflected the TAD structure (Figure 3.2g), while multiplex interactions (clusters with three

or more DNA reads) recapitulated the TAD structure (Figure 3.2h). An analysis involving downsampling the reads suggested that this difference was not due to the different read numbers in pairwise and multiplex interactions (Supplemental Note 4, Supplementary Figure 3-5b). This analysis corroborates the difference in the contact maps of different cluster sizes (Figure 3.2b-d) to suggest that the TAD chromatin structure predominantly consists of multiplex chromatin interactions. Additionally, compared to the pairwise contacts, the multiplex interactions displayed higher contact frequencies at submegabase to several megabase genomic distances, indicating enrichment of long-range chromatin interactions in the multiplex complexes (Figure 3.2e, Supplemental Note 5).

We compared MUSIC's ensemble RNA reads (RNA ensemble) from the 2,546 H1 cells to RNA measurements obtained from two bulk assays in H1. Using all 60,719 genes defined in GENCODE v36, we quantified the RNA level of each gene in terms of reads per kilobase (RPK). The RPKs of MUSIC's RNA ensemble correlated with those of bulk RNA-seq (ENCSR000COU (ENCODE Project, 2012)) (Figure 3.2i, rho = 0.8, p-value = 2.2e-16). Furthermore, iMARGI is a bulk assay of RNA-chromatin interactions, where the collection of the RNA reads from iMARGI (iMARGI's RNA reads) measures the transcriptome in the nuclei (Calandrelli et al., 2023). MUSIC's RNA ensemble also correlated with those of iMARGI's RNA reads (Figure 3.2j, rho = 0.9, p-value = 2.2e-16). This indicates that the gene expression levels quantified by MUSIC's ensemble RNA reads are consistent with those obtained from bulk RNA assays. Moreover, MUSIC detects various types of RNA species (Supplementary Figure 3-5 c,d), and is strand specific (Figure 3.2k, Supplementary Figure 3-5e). Additionally, MUSIC recapitulated the known chromatin association patterns of pre-mRNAs and nuclear-speckle associated RNAs (nsaRNA) (Figure 3.2l, m, Supplemental Note 6).

A MUSIC map of human frontal cortex

We generated a MUSIC dataset on 14 postmortem samples of the human frontal cortex (FC) from tissue donors aged 59 and above (Beach et al., 2015) (Supplementary Table 4). This dataset, hereafter referred to as MUSIC FC, resolved 9,087 single nuclei, 755,123,054 UMNDBC DNA reads and 29,319,780 UMNDBC RNA reads (hg38) (Supplementary Figure 3-6a, e). MUSIC FC resolved comparable numbers of single-nuclear RNA reads and DNA-DNA contacts with other methods (Supplemental Note 7, Supplementary Figure 3-6b-d).

A clustering analysis based on MUSIC's single-nucleus RNA reads identified seven cell types (Figure 3.3a, Supplemental Note 8). The microglia cluster consists of two subclusters, marked by low and high expression levels of Membrane-spanning 4A (MS4A) genes (Supplementary Figure 3-7m-o), which may reflect the microglial sub-populations at the chemokine state to the interferon state (Deming et al., 2019). Additionally, a joint analysis of MUSIC FC with a single-nucleus RNA-seq (snRNA-seq) dataset of human frontal cortex (Mathys et al., 2019) revealed highly consistent clustering structures and clustering-based cell type assignments between the two datasets (Supplemental Note 9).

The stratification analysis by sex (Supplementary Figure 3-7a, e) and individual cortex sample (Supplementary Figure 3-7d) did not significantly impact the proportions of cells in the clusters or subclusters, except for a higher number of oligodendrocytes in males compared to females (Supplementary Figure 3-7b, c, e). Our data indicate a sex difference in the number of cortical oligodendrocytes in the elderly people (>=59 years of age), which aligns with previous studies showing "the lifespan of oligodendrocytes is shorter in females than in males" in mice (Cerghet et al., 2006). In summary, MUSIC FC data formed clear clusters and these clusters correspond with known cortical cell types and cellular states.

Heterogeneity in chromatin interactions

Bulk analyses of chromatin conformation revealed that chromatin interaction frequency (Pc) decreases as the genomic distance (s) increases, forming an approximately linear relationship on the log-log scale (Lieberman-Aiden et al., 2009). This trend is observed in MUSIC FC as well, where the aggregate chromatin interaction frequency (Pc) in the ensemble of single cells decreases with increasing genomic distance (s) (Figure 3.3b). Hereafter, we will refer to this trend as the "aggregate Pc-s relationship".

At the single-cell level, most single cells exhibited a reverse correlation between Pc and s as well, whereas a minority of single cells exhibited the largest Pc not necessarily at the smallest s, a deviation from the aggregate Pc-s relationship (Figure 3.3c, d). This observation is reminiscent of the recently reported increase in ultra-long-range intrachromosomal interactions during aging in cerebellar granule cells (Tan et al., 2023) (Supplementary Figure 3-8e). To test if this observed cellular heterogeneity is compatible with the aggregate Pc-s relationship, we binned the genomic distances and counted the proportion of single cells that exhibit the largest Pc in each genomic distance bin (Figure 3.3d). The proportion of single cells is smaller in the bins of longer genomic distances, conforming to a reverse correlation that is approximately linear in the log-log scale (Figure 3.3c). Thus, despite the high degree of cellular heterogeneity, the population summary of the single cells reproduces the previously reported aggregate relationship.

While the different cell types exhibited similar aggregate Pc(s) curves, these Pc(s) curves are not identical (Figure 3.3b). These differences indicate cell-type variations in chromatin conformation. In particular, excitatory neurons exhibited more frequent chromatin interactions within the sub-Mb range of genomic distances than other cell types (Figure 3.3b). Consistent with this aggregate behavior, excitatory neurons had a larger proportion of single cells exhibiting the most frequent chromatin interactions at the sub-Mb range compared to the other cell types (Figure

3.3c, Supplementary Figure 3-7f). Together, these observations highlight the influence of cellular composition in each cell type on the cell-type variation in chromatin conformation.

We compared the cellular heterogeneities in Pc(s) and transcriptomic profile. For convenience, we will call the cells with large Pc at small s as 'Local chromatin structure preserved' (LCS-preserved) cells and those with large Pc at large s as “Local chromatin structure eroded” (LCS-eroded) cells. We simplified the Pc(s) for each cell into a singular score, that is the peak genomic distance of Pc(s), which we call “LCS-erosion score” for this cell. A larger LCS-erosion score reflects a greater loss in the local chromatin structure. In every cell type, we identified the genes with their single-cell expression levels correlated with the single-cell LCS-erosion scores. Pathway enrichment analysis revealed that these genes are enriched in the expected functions of the corresponding cell type (Supplementary Figure 3-8a). For example, in excitatory neurons, LCS-eroded cells exhibited reduced expressions of genes associated with Axon guidance, ErbB signaling, Glutamatergic synapse; whereas in inhibitory neurons, LCS-eroded cells exhibited reduced expression of genes in GABA synthesis. These data suggest that the cell-type specific functions are impaired in the LCS-eroded cells.

Next, we analyzed all the cortical cells together, and identified the genes with their single-cell expression levels correlated with the single-cell LCS-erosion scores. A group of small nuclear RNA (snRNA) genes (RNU4ATAC, RNU4-2, RNU5A-1, RNU12) emerged as the group highly correlated with LCS-erosion scores, with high expression in LCS-preserved cells in every cell type (Figure 3.3d). snRNAs are integral components of the spliceosome, and reduced spliceosome fidelity has emerged as a characteristic of cellular senescence and aging (Bhadra et al., 2020; Huang et al., 2022) . These data and the cell-type specific analyses let us speculate that the single-cell’s Pc(s) indicates the cell’s “age”. To test this idea, we used the recently published SCALE

model (Mao et al., 2023) to compute each single cell's transcriptomic age, which is a model-based weighted average of the age-related genes' expression in this cell. LCS-eroded cells exhibited higher transcriptomic ages than the LCS-preserved cells in every cell type, suggesting LCS-eroded cells are older in transcriptomic ages (Figure 3.3e). In comparison, the chronological age of a sample (age at death) only exhibited weak correlations with LCS-erosion scores (Supplementary Figure 3-8b), suggesting the limited ability of the donor's chronological age to explain the cellular heterogeneity of the LCS-erosion scores (or Pc(s)). Taken together, the cells with reduced chromatin contact frequencies at small genomic distances tend to exhibit older transcriptomic ages.

Seven of the 14 cortex tissues exhibited high pathology of Alzheimer's disease (AD) (Braak score >= 4), whereas the other seven exhibited low pathology (Braak score <=3). The single cells of the low pathology samples exhibited smaller LCS-erosion scores than those of the high pathology samples in excitatory neurons, inhibitory neurons, astrocytes, oligodendrocytes, and microglia (Supplementary Figure 3-8c), revealing a correlation between the loss of local chromatin structure and high pathology. These data are reminiscent of the recent report on the association of a global epigenome dysregulation and AD (Xiong et al., 2023). Additionally, a regression analysis reveals the factors with the largest correlations to LCS erosion score include transcriptomic age, cell type, AD pathology, and sex (Supplemental Note 10).

Correlation of eQTL and chromatin contacts

Cis expression quantitative trait loci (cis eQTLs) represent genomic regions where individual sequence variations contribute to the expression variability of nearby genes (Bryois et al., 2022). A recent study reported 481,888 cis eQTLs in the human brain (Bryois et al., 2022). Notably, the majority of these cis eQTLs exert their influence on the expression of specific target genes within particular cell types, indicating a strong cell-type specificity in eQTL-target pairings

(Figure 3.3f). The underlying factor responsible for the cell-type specificity observed in eQTL-target pairings remains elusive. We compared the cell-type variations in chromatin contacts (Figure 3.3b) with eQTL-target pairings in an association test. This analysis involved all eQTL-target pairs and all the chromatin contacts overlapping with an eQTL and its target gene promoter (supporting DDs), irrespective of their appearance in the same cell type. This test identified a significant enrichment of supporting DDs from the same cell type as the eQTL-target pair (p-value < 2.2e-16, Chi-square test, degrees of freedom = 25).

Further analysis focusing on eQTL-target pairs exclusive to individual cell types (cell type-specific (CTS) eQTL-target pairs) reiterated this observation, demonstrating a similar enrichment of supporting DDs within corresponding cell types (p-value < 2.2e-16, Chi-square test, degrees of freedom = 25) (Figure 3.3g). For example, DD contacts linking NELL1's cis eQTLs to its promoter were exclusively observed in excitatory neurons, where these eQTLs singularly influence NELL1 expression variability, while similar cell-specific connections were identified for PREX2 in oligodendrocytes, where the eQTLs exclusively impact PREX2 expression variability (Figure 3.3h). Taken together, the cell types exhibiting chromatin contact between a cis eQTL and its target gene promoter tend to coincide with those where the cis eQTL influences the expression of the target gene.

Cellular variation in XIST-chrX contacts

The XIST lncRNA is detected in female cortical cells but not in any male cells (Figure 3.4a), which is consistent with its expected presence in female somatic tissues and absence in male tissues (Weakley et al., 2011). In the ensemble of female cells, the XIST lncRNA exhibited a strong association with the entire X chromosome (Figure 3.4b, c), consistent with its known ability to spread across one of the X chromosomes (the Xi chromosome) (Disteche, 2012; J. T. Lee, 2012).

At the single-cell level, the female cortical cells exhibited heterogeneous XIST association level (XAL), as measured by the number of RD clusters involving XIST lncRNA and X chromosomal DNA in a nucleus (Supplementary Figure 3-9c). Recognizing the potential false negatives, we exercised caution in data analysis and interpretation (Supplemental Note 11, Supplementary Figure 3-9a). Filtering the female cells based on a threshold of total RNA reads per cell (> 5000) did not eliminate cellular heterogeneity, indicating that the observed heterogeneity cannot be solely attributed to the limited sensitivity of the technique. As expected, the XIST RNA read count in a cell correlated with XAL among the female cells (XIST_ct column vs. Heatmap, Supplementary Figure 3-9e), whereas the number of chromosome X DNA reads remained relatively invariant, confirming that the total DNA read count of the X chromosome is independent of XAL (log2 chrX DNA column, Supplementary Figure 3-9e). Furthermore, the loss of XIST-chrX association in single female cells correlates with larger sex-difference in the gene expression in the human frontal cortex (Supplemental Note 12).

We compared the X-chromosomal clusters associated with XIST lncRNA (XIST+) and those not associated with XIST lncRNA (XIST-). XIST+ clusters included RD clusters with at least one XIST RNA read, while XIST- clusters were RD and DD clusters that did not contain any XIST RNA read. At the chromosomal scale, most DNA-DNA contacts in XIST- clusters were concentrated near the diagonal line in the chromatin contact map (Figure 3.4d), similar to the contact maps of autosomes. However, XIST+ clusters exhibited not only near-diagonal contacts but also a significant number of contacts spanning distances of 10 Mb or more (Figure 3.4d).

Consistent with the chromatin contact maps, the frequency of chromatin contacts (Pc) of XIST- clusters is greater than that of the XIST+ clusters when the genomic distance (s) is smaller than ~10Mb (Figure 3.4e). To determine if this separation in Pc(s) curves could be attributed to

limited sensitivity in detecting XAL at the single-cell level, a stratification analysis was performed. Female cells were stratified into four groups based on zero, low, medium, and high XALs. Of note, the zero XAL group (Group 1) only contains XIST- clusters, and Groups 2, 3, and 4 contain the same number of cells. Pc(s) of the XIST- clusters (XIST- Pc(s)) is above XIST+ Pc(s) when s is smaller than ~10Mb in Groups 2, 3, and 4 (Figure 3.4f). Importantly, the difference between XIST- and XIST+ Pc(s) curves increased from Group 2 to Group 4, indicating that higher XALs in a cell group led to a more pronounced chromatin conformation difference between XIST- and XIST+ clusters. As a control, the Pc(s) curves of X chromosomal clusters associated with or without any RNA (Any_RNA+ Pc(s) or Any_RNA- Pc(s)) in Group 1 (zero XAL group) were nearly indistinguishable (Figure 3.4f). These data suggest that the active X chromosome (Xa) has a higher contact frequency than the inactive X chromosome (Xi) in the sub-10 Mb range of genomic distances in the female human cortex.

Different cell types exhibited different proportions of XAL-positive (XAL+) cells (Supplementary Figure 3-9b), with excitatory neurons having the highest proportion of XAL+ cells (Chi-square, p-value < 10e-16) (Figure 3.4g). Considering the larger separation between XIST+ and XIST- Pc(s) curves (ΔPc(s)) in XAL-high cells, we anticipated seeing a cell-type difference in ΔPc(s), particularly with excitatory neurons to exhibit a larger ΔPc(s) compared to other cell types. To test this idea, we compared the three cell types with at least 45% of cells exhibiting non-zero XAL, namely excitatory neurons (ExN), inhibitory neurons (InN), and astrocytes (Ast) (Supplementary Figure 3-9b). As expected, ExN displayed a larger ΔPc(s) and more long-range interactions than InN and Ast (Figure 3.4h, i). Specifically, ExN's XIST+ Ps(c) curve is beneath the XIST- Ps(c) curve when the genomic distance is less than ~10Mb and traverses above the XIST- Ps(c) curve at ~10Mb. Although this transversion was consistently observed in InN and

Ast, the gaps between the Ps(c) curves were narrower in InN and Ast. Downsampling ExN, InN, and Ast to the same number of cells did not change this observation (Supplementary Figure 3-9g). Consistently, seqFISH+ analysis manifested a larger conformational difference between Xa and Xi in ExN compared to InN and Ast in female mice (Takei et al., 2021b) (Supplemental Note 13). Taken together, these data suggest a conserved cell-type variation in the spatial organization of the two X chromosomes in the female cortex in mice and humans.

Discussion

Human cortical cells exhibited a heterogeneity in the distribution of genomic-distance-dependent chromatin contact frequency (Pc(s)). To capture this diversity, we introduced a metric called the LCS-erosion score, designed to condense Pc(s) into a singular value. A higher LCS-erosion score signifies a more pronounced decline in local chromatin contacts. Notably, cells with elevated LCS-erosion scores (termed LCS-erosion cells) demonstrated a tendency toward exhibiting transcriptomic profiles indicative of an 'older' cellular age in comparison to their counterparts, termed LCS-preserved cells. This distinction sheds light on the correlation between a cell's chromatin conformation and its transcriptomic age, thus extending our understanding from prior observations associating chromatin structural decline with aging (Yu et al., 2020) to the single-cell level. Consequently, we introduce the concept of a cell's 'chromatin conformational age', indicated by the LCS-erosion score.

Compared to the males, female cortexes exhibit less “old” neurons and more “old” oligodendrocytes in chromatin conformation age (Figure 3.3d, Supplementary Figure 3-8d), i.e., a larger oligodendrocyte/neuron ratio in the “old” cells (odds ratio = 6.85, p-value < 2.2e-16, Chi-square test). Reminiscent to this observation, compared to males, female mice exhibit more age-related cell deaths in oligodendrocytes but not in neurons (Cerghet et al., 2006). Healthy oligodendrocytes protect the normal neuronal activities, and a balance between neurons and

oligodendrocytes is required for maintaining their bidirectional communications which ensure the necessary protectivity (Depp et al., 2023; Mazuir et al., 2021). We speculate that the disproportionately more “old” oligodendrocytes in females contribute to explain the increased risks in neurodegenerative and mental disorders in women.

The genomic sequence at each expression quantitative trait locus (eQTL) usually remains constant across various cell types within an individual, raising questions about why most eQTLs exert their influence on gene expression in specific cell types. Our findings reveal a connection between cell-type-specific pairing of eQTLs with their target genes and the variability in chromatin contacts between the eQTL and the target gene's promoter. This underscores the significance of cell-type differences in chromatin contacts for future investigations aimed at unraveling this puzzle.

The widespread impact of 3D genome organization on the expression of numerous genes (Model 1) and the reciprocal influence of gene expression on chromatin conformation (Model 2) continue to be subjects of debate (Calandrelli et al., 2023; Rao et al., 2017). Recent findings suggest that alterations in chromatin conformation precede changes in gene expression during development, lending support to Model 1 while challenging Model 2 (Z. Liu et al., 2023). However, experiments involving the degradation of cohesin, a pivotal regulator of chromatin conformation, demonstrated limited impact on the expression levels of most genes in a human cell line, casting doubt on Model 1 (Rao et al., 2017). Our data potentially reconcile the limited impact of cohesin degradation with Model 1. The correlation observed between variations in cell-type-specific chromatin contacts and eQTL-target pairing highlights the potential significance of incorporating cell-type and interindividual variations to comprehend a more comprehensive impact.

Following initial debates (Holmes et al., 1972), caRNA has been increasingly acknowledged as a structural component of chromatin (Thakur and Henikoff, 2020). Work in Drosophila and Gallus gallus suggested 2%–5% of total chromatin-associated nucleic acids are RNA (Rodríguez-Campos and Azorín, 2007). In the human MUSIC data, RNA reads account for approximately 4.6% of all the chromatin-derived reads, including all the DD and RD clusters, indicating a relatively consistent proportion of RNA in chromatin-associated nucleic acids across species. Interestingly, approximately 11.7% of the non-singleton chromatin clusters (DD or RD) contain RNA reads. The chromatin clusters containing RNA more frequently demonstrate multiplex DNA-DNA contacts than those devoid of RNA, as shown in Supplementary Figure 3-4c and d. This observation aligns with the hypothesis that RNA contributes to spatial genome compartmentalization (Calandrelli et al., 2023; Quinodoz et al., 2021).

Women exhibit numerous differences in neurodegenerative diseases and mental disorders from men, including that there are twice as many women with late-onset Alzheimer's disease (LOAD) as men, and that women have a significantly higher frequency of adulthood depression and anxiety compared to men. Notably, many X-linked genes are expressed in the brain and have a role in cognitive functions (Nguyen & Disteche, 2006). MUSIC data revealed that in female cortical cells, the diminishing association between XIST and chromosome X correlates with reduced structural differences between the active and inactive X chromosomes, and this is associated with greater differences in chrX gene expression between sexes. These multimodal single-cell data provide a critical resource for future investigations on sex differences in health and disease. In summary, MUSIC provides a unique tool to jointly analyze gene expression, multiplex chromatin interactions, and RNA-chromatin associations with single-cell resolution from complex tissue.

Figure

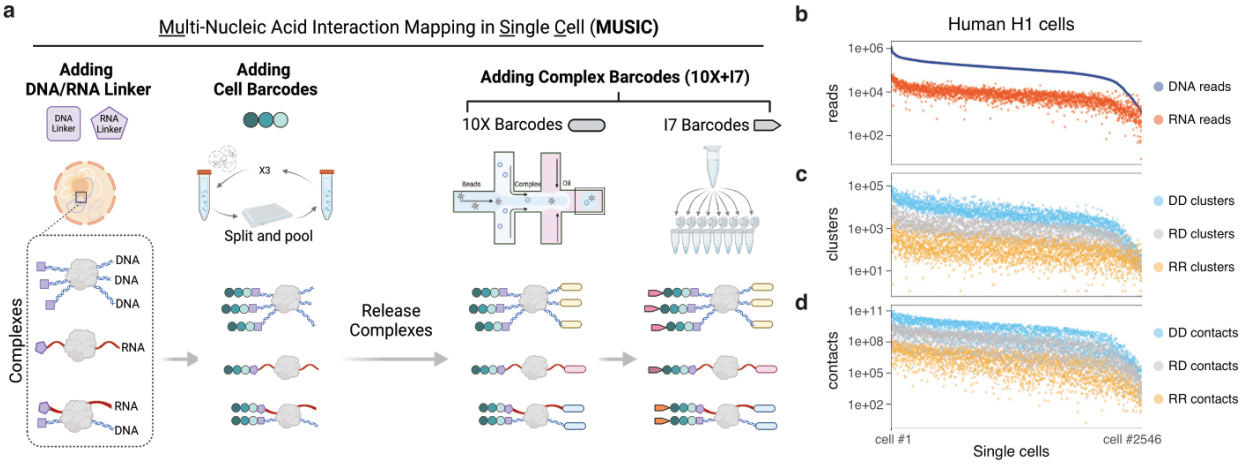


Figure 3.1: MUSIC workflow and statistics.
(a) Schematic view of MUSIC experimental pipeline. (b-d) Summary of MUSIC data in H1 cells. The numbers of uniquely mapped non-duplicate reads (b), clusters (c), and pairwise contacts (d) in every H1 cell (column, n=2,546). DNA-DNA (DD, blue), RNA-DNA (RD, gray), and RNA-RNA (RR, yellow) clusters are separately counted. Multiplex interactions are projected to pairwise interactions and the numbers of pairwise contacts are reported in (d).

Figure 3.2: H1 cells MUSIC data.
(a) Comparison of Micro-C (upper triangle) and ensemble MUSIC (lower triangle) derived chromatin contact maps on Chromosome 1 at 1Mb resolution. (b-d) Chromatin contact maps based on the ensemble of small (b), middle-sized (c), and large (d) clusters. Resolution: 50Kb. Arrows: Contacts between nested TADs. (e) Pc(s) curves showing the frequency of chromatin contacts (Pc, y axis) vs. genomic distance (s, x axis) for MUSIC DD clusters with different sizes (shades of blue) and Micro-C (black). DNA cluster size: the number of DNA reads in a DD cluster. (f-h) Comparison of 2-D contact map from ensemble MUSIC data (f) with stacked maps of distinct DNA-DNA clusters. Each row represents a cluster, ordered by the smallest genomic coordinate of any DNA read. Yellow dots denote genomic locations of DNA reads within a cluster. Clusters with 2 DNA reads (pairwise interactions) are shown in (g), while clusters with 3 or more DNA reads (multiplex interactions) are shown in (h). (i-j) Scatterplots of RNA levels measured by Reads Per Kilobase (RPK) for every gene (dot) in ensemble MUSIC (x axis) vs. RNA-seq (y axis) (i) and iMARGI (y axis) (j). R: Spearman correlation. (k) RNA reads from RNA-seq, iMARGI, and MUSIC mapped to both strands (+ and -) in human H1 cells. Ensemble MUSIC and individual MUSIC data from three single cells are presented. (l) Distribution of chromatin-associated pre-mRNA and nsaRNA on Chromosome 1 as measured by ensemble MUSIC in H1 cells. Micro-C derived A/B compartments are colored in red/blue. The “Speckle compartmentalization” derived from the SPIN model is denoted in the SPIN-Speckle track. (m) Scatter plot of normalized counts of pre-mRNA reads (y axis) and nsaRNA reads (x axis) in every 1 Mb genomic bin (dot) across the entire genome, based on Ensemble MUSIC data in H1 cells.

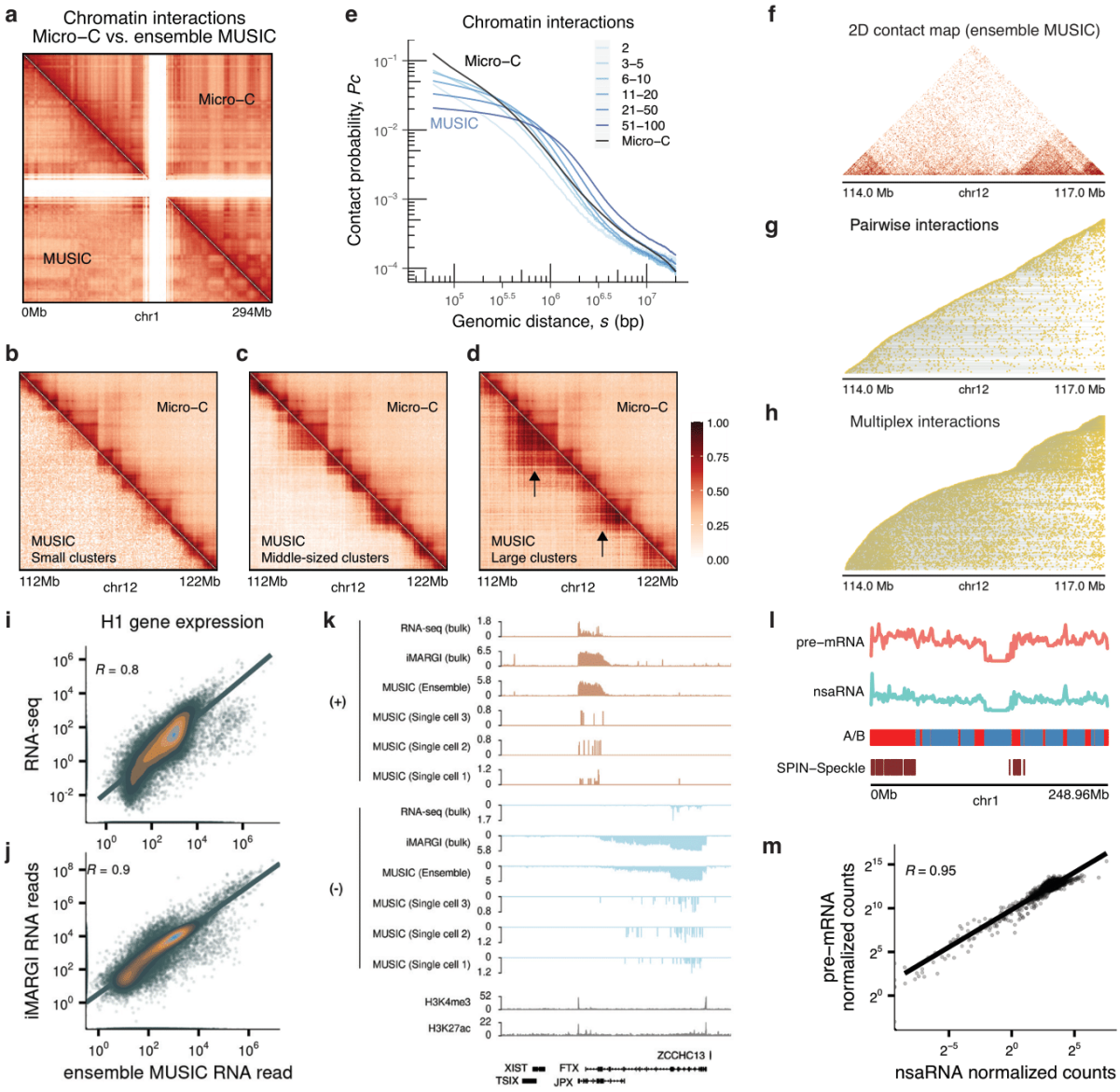

a
Chromatin interactions
Micro–C vs. ensemble MUSIC
Micro–C
MUSIC
0Mb
chr1
294Mb
e
Chromatin interactions
Micro–C
MUSIC
2
3–5
6–10
11–20
21–50
51–100
Micro–C
Contact probability, Pc
Genomic distance, s (bp)
f
2D contact map (ensemble MUSIC)
114.0 Mb
chr12
117.0 Mb
g
Pairwise interactions
114.0 Mb
chr12
117.0 Mb
h
Multiplex interactions
114.0 Mb
chr12
117.0 Mb
b
Micro–C
MUSIC
Small clusters
112Mb
chr12
122Mb
c
Micro–C
MUSIC
Middle-sized clusters
112Mb
chr12
122Mb
d
Micro–C
MUSIC
Large clusters
112Mb
chr12
122Mb
1.00
0.75
0.50
0.25
0.00
i
H1 gene expression
R = 0.8
RNA-seq
j
R = 0.9
iMARGI RNA reads
ensemble MUSIC RNA read
k
RNA-seq (bulk)
iMARGI (bulk)
MUSIC (Ensemble)
MUSIC (Single cell 3)
MUSIC (Single cell 2)
MUSIC (Single cell 1)
(+)
(-)
H3K4me3
H3K27ac
XIST
TSIX
FTX
JPX
ZCCHC13
l
pre–mRNA
nsaRNA
A/B
SPIN–Speckle
0Mb
chr1
248.96Mb
m
R = 0.95
pre–mRNA normalized counts
nsaRNA normalized counts

Figure 3.3: A single-cell map of transcriptome and chromatin conformation in the human frontal cortex.
(a) UMAP representation of individual cortical cells based on MUSIC RNA reads. (b) Chromatin contact frequency (Pc, y axis) versus genomic distance (s, x axis) for each cell type (color). (c) Histogram of the proportions of single excitatory neurons (ExN) (rows) with their most frequent chromatin interactions in each genomic bin (x axis) is aligned with the normalized contact frequency (color intensity) vs. genomic distance (x axis) plot for every ExN (row). (d) Chromatin contact frequency vs. genomic distance (y axis) in individual cortical cells (columns). Rows: genomic bins with an exponential size increase. Color scale: chromatin contacts frequency normalized by bin size. Bottom tracks showing the Pc(s) group, Chronological age, Transcriptomic age, AD pathology status, Sex, and the expression levels of several genes. (e) Single cell transcriptomic age (x axis) in each cell type (y axis) colored by chromatin conformation age. Wilcoxon test, one-sided, p-value on the right. Left numbers: sample size. Boxplot: left, center and right edges of box represent the 25th, 50th and 75th quantile. Whiskers extend to 1.5 times the Interquartile Range from the box edges. Data points beyond the whiskers are outliers. (f) Upset plot of the eQTL-target pairs in every cell type. Box indicates the cell-type specific eQTL-target pairs. (g) Association tests between the DD contacts and CTS eQTL-target pairs in every cell type. Center dot: odds ratio. Whiskers: 95% confidence interval of the odds ratio. One-sided Chi-square test, N = 101,785 DD contacts. (h) Examples of CTS eQTL-target pairs supported by cell-type specific DD contacts. Top panel: check mark indicates the cell type (column) where the cis eQTLs of a gene affect this gene's expression. Bottom panel: genome track view of supporting DDs in every cell type (row), which are the DD contacts (curves) overlapping with the cis eQTLs (blue arrows) and the target gene's promoter (red arrow).

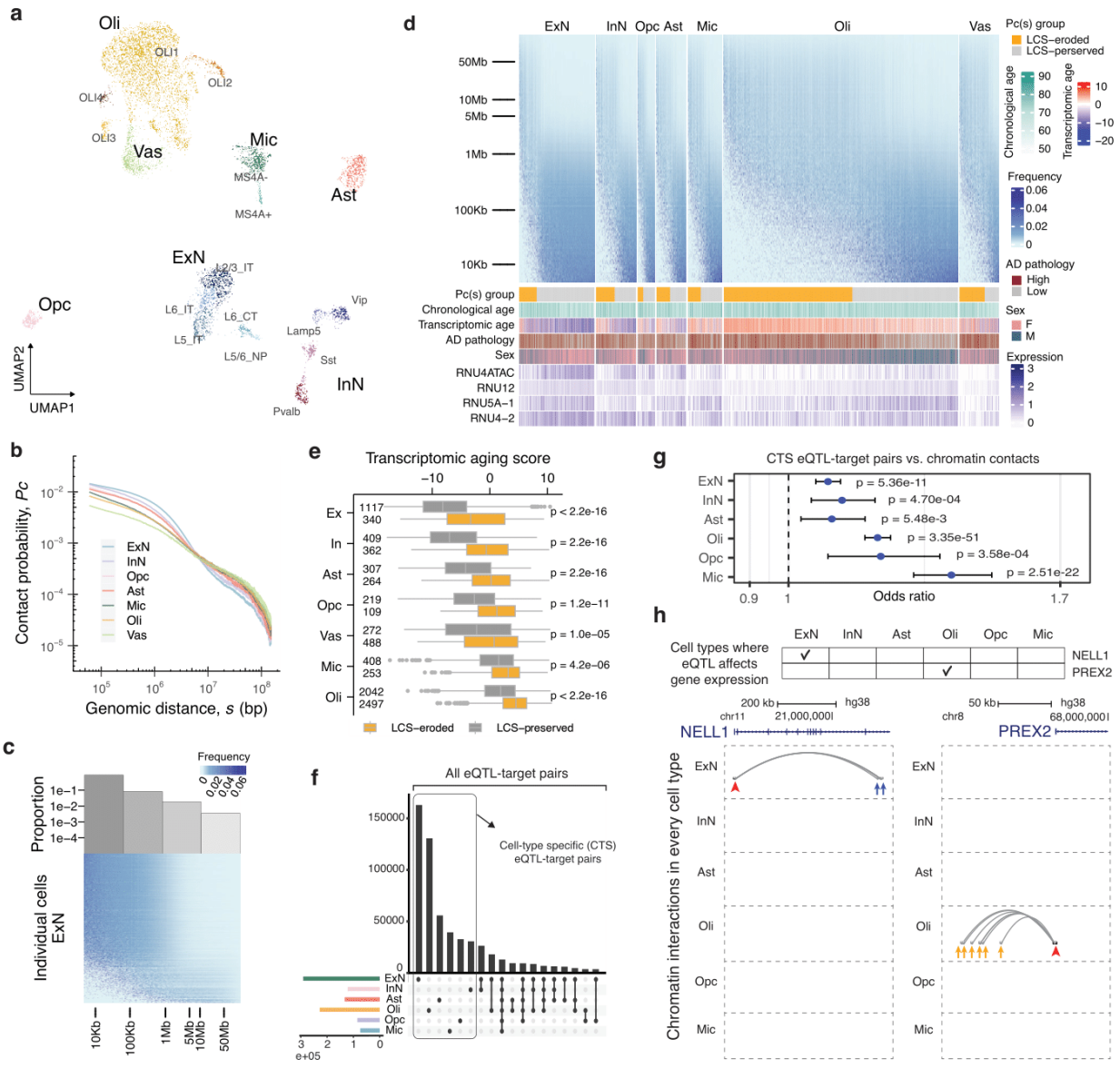
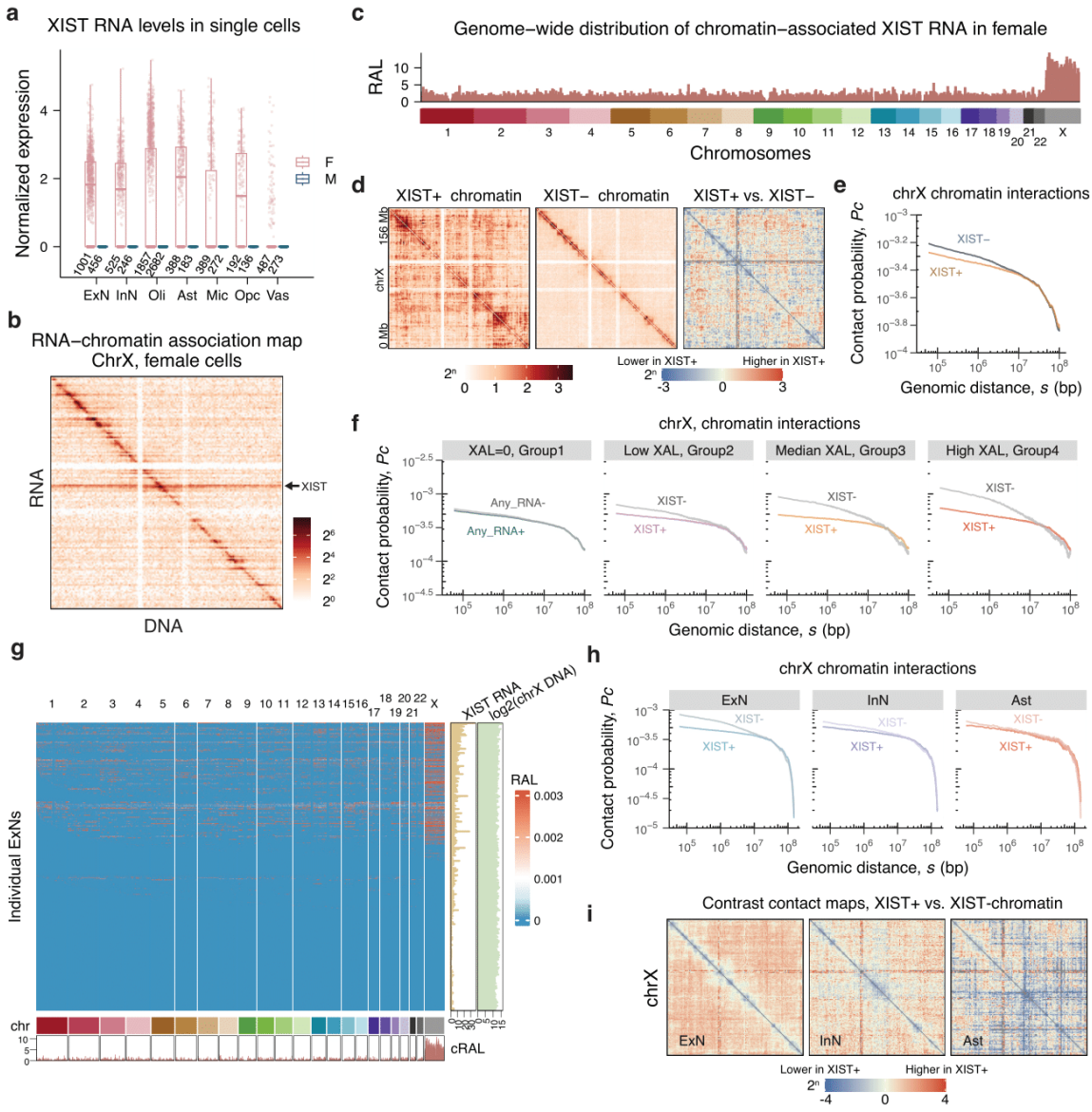

a
Oli
OLI1
OLI2
OLI4
OLI3
Vas
Mic
MS4A-
MS4A+
Ast
ExN
L2/3_IT
L6_IT
L6_CT
L5_IT
L5/6_NP
Opc
Vip
Lamp5
Sst
InN
Pvalb
UMAP2
UMAP1
d
ExN
InN
Opc
Ast
Mic
Oli
Vas
50Mb
10Mb
5Mb
1Mb
100Kb
10Kb
Pc(s) group
Chronological age
Transcriptomic age
AD pathology
Sex
RNU4ATAC
RNU12
RNU5A−1
RNU4−2
Pc(s) group
LCS−eroded
LCS−perserved
Chronological age
Transcriptomic age
Frequency
AD pathology
High
Low
Sex
F
M
Expression
b
Contact probability, Pc
ExN
InN
Opc
Ast
Mic
Oli
Vas
Genomic distance, s (bp)
e
Transcriptomic aging score
Ex
In
Ast
Opc
Vas
Mic
Oli
p < 2.2e-16
p = 2.2e-16
p = 2.2e-16
p = 1.2e−11
p = 1.0e−05
p = 4.2e−06
p < 2.2e-16
LCS−eroded
LCS−preserved
g
CTS eQTL-target pairs vs. chromatin contacts
p = 5.36e-11
p = 4.70e-04
p = 5.48e-3
p = 3.35e-51
p = 3.58e-04
p = 2.51e-22
Odds ratio
h
Cell types where eQTL affects gene expression
NELL1
PREX2
Chromatin interactions in every cell type
c
Frequency
Proportion
Individual cells
ExN
10Kb
100Kb
1Mb
5Mb
10Mb
50Mb
f
All eQTL-target pairs
Cell-type specific (CTS) eQTL-target pairs

Figure 3.4: Cellular heterogeneity of XIST-chromatin interactions in female frontal cortex.
(a) XIST expression level (y axis) of every single cell (dot) in each cell type for the female and male. Boxplot: bottom, center and top edges of box represent the 25th, 50th and 75th quantile. Whiskers extend to 1.5 times the Interquartile Range from the box edges. Data points beyond the whiskers are outliers. (b) RNA-DNA contact map for Chromosome X based on the ensemble female cells. Each pixel represents the amount of RNA that is transcribed from the genomic bin (row) and is associated with the genomic bin (column). Resolution: 1Mb. (c) Distribution of chromatin-associated XIST on every chromosome in the ensemble of female cells. Resolution: 1Mb. Y axis: the RNA attachment level (RAL) in a genomic bin among the female cells. (d) X chromosome contact maps for XIST-associated (XIST+), not-associated (XIST-) chromatin in female cells, and their contrast. (e) Pc(s) curves showing the frequency of chromatin contacts (Pc, y axis) vs. genomic distance (s, x axis) for XIST+ and XIST- female chrX chromatin complexes. (f) ChrX Pc(s) curves in four female cell groups, with zero, low, medium, and high XAL (left to right). The Pc(s) curves for chromatin with (Any_RNA+) and without any associated RNA (Any_RNA-) do not exhibit a notable difference. The difference between XIST+ and XIST- Pc(s) curves increases as XAL increases. (g) Genome-wide distribution of XIST-chromatin association in ExNs. Each row corresponds to a single ExN, displaying XIST lncRNA attachment levels (red color intensity) across genomic regions on various chromosomes. The bottom track illustrates cumulative XIST RNA attachment levels. Right tracks show XIST RNA read count and X chromosomal DNA read counts in log2 scale. (h) X chromosome Pc(s) curves in female excitatory neurons (ExN), inhibitory neurons (InN), and astrocytes (Ast). The difference between XIST+ (darker color) and XIST- Pc(s) (lighter color) curves is most pronounced in ExN. (i) Contrast contact maps between XIST+ and XIST- chromatins on chromosome X in female ExN, InN, and Ast.


a
XIST RNA levels in single cells
Normalized expression
F
M
ExN InN Oli Ast Mic Opc Vas
b
RNA–chromatin association map
ChrX, female cells
RNA
DNA
XIST
c
Genome–wide distribution of chromatin–associated XIST RNA in female
RAL
Chromosomes
d
XIST+ chromatin
XIST− chromatin
XIST+ vs. XIST−
Lower in XIST+
Higher in XIST+
e
chrX chromatin interactions
Contact probability, Pc
XIST−
XIST+
Genomic distance, s (bp)
f
chrX, chromatin interactions
XAL=0, Group1
Low XAL, Group2
Median XAL, Group3
High XAL, Group4
Any_RNA-
Any_RNA+
g
Individual ExNs
XIST RNA log2(chrX DNA)
RAL
cRAL
h
chrX chromatin interactions
ExN
InN
Ast
i
Contrast contact maps, XIST+ vs. XIST-chromatin
chrX

Methods

Critical Reagents

The RNA Linker

The RNA Linker is a single stranded chimeric oligonucleotide with 17 DNA nucleotides at its 5' end (ssDNA, Supplementary Figure 3-2e) and 10 RNA nucleotides at its 3' end (ssRNA, Supplementary Figure 3-2e), denoted as 5' /5OH/CGAGGAGCGCTTNNNNNrArUrArGrCrArUrUrGrC/3OH/ 3', where A, C, G, T, and N denote DNA nucleotides, rA, rC, rG, rT denote RNA nucleotides, NNNNN denotes 5 randomized DNA nucleotides that serves as a unique molecular identifier (UMI). The RNA Linker was synthesized by IDT.

The RNA Linker is designed for (1) efficient ligation with RNA through the RNA Linker's ssRNA, (2) efficient ligation with the 1st set of Cell Barcodes through the DNA Linker's ssDNA, which complements with the 7 nt overhang in the 1st set of Cell Barcodes.

The DNA Linker

The DNA Linker is a hybridized product of two DNA strands with the top strand being 5' /5Phos/CTAGACACTGTGCGTATCTNBAAAAAAAAAAAAAAAAAAAAAAAAAAAA /3OH/ 3', where N denotes a random base and B denotes any base except A, and the bottom strand being 5' /5OH/CGAGGAGNNNNNACAACGCACAGTGTCTAGT/3OH/ 3', where NNNNN denotes 5 randomized DNA bases that serve as a UMI. After hybridization, the DNA Linker contains 15bp double-stranded DNA (referred to as dsDNA) and 36 nt (top-ssDNA) and 15 nt unhybridized single stranded DNA (bottom-ssDNA), and a single base (T) overhang at the bottom strand (Supplementary Figure 3-2g). The 36 nt top-ssDNA is reverse complementary to the 10X Barcodes in the Chromium Next GEM Single Cell 3' Reagent kit (PN-1000268). The two strands of the DNA Linker are synthesized by IDT.

The DNA Linker is designed for (1) efficient ligation with fragmented chromosomal DNA through the DNA Linker's 1 nt overhang by a sticky-end ligation, (2) efficient ligation with the Cell Barcodes, through the DNA Linker's 15 nt bottom-ssDNA, and (3) efficient hybridation with the 10X Barcodes through the polyA sequence within the top-ssDNA.

The Cell Barcodes

The Cell Barcodes contain three sets of barcodes, referred to as the 1st, 2nd, and 3rd sets of Cell Barcodes. Every set of Cell Barcodes has three components, namely a 7 nt top-strand overhang, a 14 bp dsDNA region, and a bottom-strand overhang (7 nt for the 1st and 2nd sets, 11 nt for the 3rd set). The 14 bp dsDNA region contains a unique sequence to every Cell Barcode (double-stranded N14, Supplementary Figure 3-2d, i). Every set of Cell Barcodes contains 96 unique barcodes, where each barcode is unique in this 14 bp dsDNA (Supplementary Table 2).

In the current version of MUSIC (v1.0), each set of Cell Barcodes contains 96 unique barcodes based on their dsDNA regions, resulting in a total of 884,736 unique sequence combinations. The three sets of Cell Barcodes are designed to maximize the ligation efficiency for sequentially ligating the 1st set of Cell Barcodes with the RNA Linker and the DNA Linker, the 2nd set with the 1st set, and the 3rd set with the 2nd set of Cell Barcodes. The optimal ligation efficiency is achieved by the complementarity of the overhang sequences (Supplementary Figure 3-2d, i). Out-of-order ligations such as a ligation of the 3rd set with the 1st set of Cell Barcodes are minimized because the overhang of the 3rd set does not complement with the overhang of the 1st set of Cell Barcodes. Additionally, the 3rd set of Cell Barcodes is also designed to complement the 22 nt sequence at the 3' side of the Index Adaptors. The 1st and 2nd sets of Cell Barcodes are synthesized by Sigma-Aldrich. The 3rd set of Cell Barcodes are synthesized by IDT.

10X Barcodes

The 10X Barcodes are included in the Chromium Next GEM Single Cell 3' Reagent kit PN-1000268 from 10X Genomics. Each 10X Barcode is an 82 nt oligonucleotide with a partial (22 nt) Illumina Read 1 sequence (Read 1), a 16 nt unique barcode sequence (N16, 10X GEM Barcode), a 12 nt unique molecular identifier (N12, 10X UMI), a 30 nt polyT sequence, a V (A or C or G) and an N (any base) (10X Barcode, Supplementary Figure 3-2j, k). The 10X GEM Barcode is shared among the barcodes of the same GEM. The 10X UMI is unique to every 10X Barcode.

The Index Adaptors

The Index Adaptors contain three segments, namely the 24 nt Illumina P7 sequence, a 8 nt unique identifier sequence called I7, and the 34 nt Illumina Read2 sequence (Index Adaptor, Supplementary Table 2). In this release, MUSIC v1.0 uses eight distinct I7 Barcodes, providing for a total of approximately 83.5 million Complex Barcodes (8 (I7 Barcodes) × ~3.5 million (10X Barcodes)). Eight Index Adaptors are used for every library construction. Each of the eight Index Adaptors has a unique I7 sequence. We call these eight Index Adaptors a set of Index Adaptors. Each Index Adaptor can hybridize with the complementary Read2 sequence in the 3rd set of Cell Barcodes to start a PCR reaction.

Meanwhile, they serve as sample barcodes. We designed a total of three sets of Index Adaptors to allow for constructing three libraries from three input samples and sequencing them together (Supplementary Table 2). These three sets of Index Adaptors share P7 and Read2 sequences and differ by their I7 sequences. These three sets of Index Adaptors serve as a sample index to differentiate three samples. The Index Adaptors are synthesized by IDT.

The Universal Adaptor

The Universal Adaptor contains an Illumina P5 sequence and an Illumina Read1 sequence (Universal Adaptor, Supplementary Table 2). The Universal Adaptor can hybridize with the 10X

Barcodes through their complementary Read1 sequence to start a PCR reaction. The Universal Adaptor is synthesized by IDT.

Cell Culture

H1 human embryonic stem cells and E14 mouse embryonic stem cells were obtained from the 4D Nucleome Consortium and cultured according to the 4D Nucleome Consortium approved protocols (https://www.4dnucleome.org/). Briefly H1 cells were grown at 37°C under 5% CO_2 on matrigel (Corning, 354277) coated dishes. Cells were maintained in complete mTeSR medium prepared from basal medium (Corning, 85851) with 5X supplement (Corning, 85852). Medium was replaced daily. Cell passage number was kept under P10. E14 cells were cultured on plates coated with 0.1% gelatin (EMD, SF008) in serum free 2i/LIF medium. The serum free 2i/LIF medium was made from the base medium (1:1 mixture of NeuroBasal medium (Gibco, 21103-049) and DMEM/F12 medium (Gibco, 11320-033) supplemented with 0.5X N2 supplement (Gibco, 17502-048), 0.5X B27 supplement (Gibco, 17504-044), and 0.05% BSA fraction V (Gibco, 15260-037)) supplemented with 1 μM PD0325901 (Reprocel,l 04-0006-02C), 3 μM CHIR99021 (Reprocell, 04-0004-02C), 0.15 mM Monothioglycerol (Sigma, M6145-25ML), 1000 U/mL LIF (Cell Guidance Systems, GFM200). The medium was replaced daily. Cell passage number was kept under P10.

Crosslinking and nuclei isolation for cell lines

After cells become confluent in a 10 cm dish, the media was removed and washed once with PBS. 1 mL of Accutase (EMD, SF006) was added and incubated for 3 min at 37°C to dissociate the cells. 10 mL of PBS was used to generate single cell suspension by pipetting. Cell pellets were formed by centrifugation at 330 x g for 3 min. 10 mL of 2 mM disuccinimidyl glutarate (DSG) dissolved in 1X PBS was added to crosslink and resuspend the cells in a LoBind tube and incubate at room temperature for 45 min with gentle rotation. After incubation, cells were collected

by centrifugation at 1,000 x g for 4 min to remove DSG solution and washed once with 1X PBS then centrifuged again at 1,000 x g for 4 min to remove the supernatant. After washing, cells were thoroughly resuspended in 10 mL of 1X PBS containing 3% formaldehyde and incubated for 10 min with gentle rotation. The crosslinking reaction was stopped by adding 3 mL of 2.5 M glycine per 10 mL of 3% formaldehyde and incubating for 5 min with rotation. Cells were then centrifuged at 1,000 x g for 4 min to remove the supernatant. Next, cells were washed twice with ice-cold 1X PBS containing 0.5% BSA (wt/vol) and centrifuged at 1,000 x g for 4 min. After the wash, cells were resuspended in 1X PBS with 0.5% BSA (wt/vol), and each cell aliquot contained 5 million cells in a 1.5 mL tube. The cell pellets were obtained by centrifugation at 1,000 x g for 5 min, snap-frozen in liquid nitrogen, and stored at -80°C.

Frozen cells were thawed on ice and resuspended in 1.4 mL of cell lysis buffer A for every 5 million cells as previously described (Arrastia et al., 2022a) (50 mM HEPES pH 7.4, 1 mM EDTA pH 8.0, 1 mM EGTA pH 8.0, 140 mM NaCl, 0.25% Triton X-100, 0.5% IGEPAL CA-630, 10% glycerol, 1X proteinase inhibitor cocktail (PIC)). For the mixed species experiment, equal amounts of human H1 (2.5 million) and mouse E14 (2.5 million) cells were resuspended together. After a 10-min incubation on ice, cell pellets were collected by centrifugation at 900 x g for 4 min at 4°C. Cell pellets were resuspended in 1.4 mL of cell lysis buffer B (10 mM Tris-HCl pH 8, 1.5 mM EDTA, 1.5 mM EGTA, 200 mM NaCl, 1X PIC) and kept on ice for 10 min. The isolated nuclei were collected at 900 x g for 5 min at 4°C. 200 µL of 1X rCutSmart buffer (NEB, B7204S) containing 0.25% SDS was used to thoroughly resuspend and permeabilize the nuclei and incubated at 62°C for 10 min using Eppendorf Thermomixer C (Eppendorf). After the incubation, 60 µL of 1X rCutSmart buffer containing 10% Triton X-100 (wt/vol) was mixed with the SDS solution, and the reaction was incubated at 37°C for 15 min while shaking at 800 rpm. The treated

nuclei were centrifuged at 900 x g for 2 min at 4°C to remove supernatant and washed once with 1X rCutSmart buffer.

Crosslinking and nuclei isolation for post-mortem brain

Each 50 mg of post-mortem human brain frontal cortex sample was kept on ice in a 1.5ml LoBind tube and chopped into smaller pieces by pestle. The brains samples were transferred into a 15 mL LoBind tube and incubated at room temperature for 45 min with a gentle rotation in 10 mL of 2 mM DSG that was dissolved in 1X PBS. After incubation, the tissue sample was centrifuged at 1,000 x g for 4 min and washed once with 1X PBS to remove DSG solution. After washing, the tissue sample was thoroughly resuspended in 10 mL of 1X PBS containing 3% formaldehyde and incubated for 10 min with a gentle rotation. The crosslinking reaction was stopped by the addition of 4 mL of 1.25 M glycine followed by an incubation of 5 min with a rotation. The tissue sample was then centrifuged at 1,000 x g for 4 min and washed twice with ice-cold 1X PBS containing 0.3% BSA (wt/vol).

Chromium Nuclei Isolation kit (10X genomics, 1000494) was used to isolate nuclei from crosslinked cortex samples according to the section for "single cell gene expression & chromium fixed RNA profiling" (page 25 to page 30 from sample user guide). Specifically, 50mg frozen tissue was placed into a pre-chilled sample dissociation tube. 400 µL of the provided lysis buffer was added to the tube, and tissues were dissociated until homogeneous with plastic pestle. 600 µL lysis buffer was further added into the tube and mixed 10 times by pipetting. After a 10 min incubation on ice, the solution was evenly loaded into two nuclei isolation columns and centrifuged at 16,000 x g for 20 s at 4°C. The flowthrough in the collection tube which contains nuclei was vortexed 10 s at 3,200 rpm to resuspend nuclei. The collection tube was centrifuged for 3 min at 500 x g at 4°C to pellet nuclei, and the supernatant was removed. The nuclei were resuspended in a 500 µL debris removal buffer provided by the kit by pipetting 15 times. The nuclei were

centrifuged at 700 x g for 10 min at 4°C, and the supernatant was removed. The nuclei were resuspended in 1 mL of wash and resuspension buffer twice. The supernatant was removed after centrifugation at 500 x g for 5 min at 4°C, which left a purified pellet of isolated nuclei. All the following steps are the same for either cell line or human cortex samples.

Ligation of the RNA Linker with RNA

The nuclei were resuspended in 250 µL of 5' phosphorylation master mix (1X T4 PNK buffer, 500 U/mL of T4PNK, 1 mM ATP, 1 U/µL of RNAse inhibitor (Roche, 3335399001)) and incubated at 37°C while rotating at 800 rpm for 1 hour to phosphorylate the 5' ends of RNA. Nuclei were washed once with PBS Wash Buffer-1 (1X PBS, 1 mM EDTA, 1 mM EGTA and 0.1% Triton X-100) and twice with PBS Wash Buffer-2 (1X PBS, 0.5% BSA (wt/vol) and 0.1% Triton X-100). The RNA Linker is a single stranded chimeric oligo with DNA's 5' hydroxyl group end and RNA' 3' hydroxyl group (5'-OH-CGAGGAGCGCTTNNNNNrArUrArGrCrArUrUrGrC-OH-3'). An RNA Ligation Mix is made with 4 µM of the RNA linker, 1X T4 RNA ligation buffer, 400 U/mL of T4 RNA ligase 1, 15% PEG 8000, 1 mM of ATP, and 1 U/µL of RNAse inhibitor. The isolated nuclei were thoroughly mixed with 250 µL of the RNA Ligation Mix to ligate the RNA Linker with nuclear RNA. The mixture was incubated at 25ºC for 2 hours then 16ºC overnight with an intermittent mixing at 800 rpm (30 seconds on and 270 off). After ligation, the nuclei were washed once with PBS Wash Buffer 1 and three times with PBS Wash Buffer 2.

Chromatin Digestion

All the washed nuclei were resuspended in a digestion master mix (300 µL of 1X rCutSmart buffer containing 30 µL of 5,000 U/mL of HpyCH4V with 1 U/µL of RNAse inhibitor). The digestion master mix was kept for 3 hours at 37°C while rotating at 800 rpm. The nuclei were collected at 900 x g for 2 min with supernatant removed. The nuclei were further washed once

with 900 µL of PBS Wash Buffer 1 (1X PBS, 1 mM EDTA, 1 mM EGTA and 0.1% Triton X-100) and three times with 900 µL of PBS Wash Buffer 2 (1X PBS, 0.3% BSA (wt/vol) and 0.1% Triton X-100).

Ligation of the DNA Linker with DNA

To create sticky end for DNA linker ligation, the nuclei was suspended in 250 µL of dA-tailing reaction master mix (1X NEBNext dA-tailing reaction buffer, 200 U/mL of Klenow fragment, 1 U/µL of RNAse inhibitor) and incubated at 37 ºC while rotating at 800 rpm for 1.5 hours. Next, the nuclei were washed once with PBS Wash Buffer-1 and three times with PBS Wash Buffer-2. The DNA Linker is a hybridized product of two DNA strands with the top strand being 5'-Phos-CTAGACACTGTGCGTATCT NBAAAAAAAAAAAAAAAAAAAAAAAAAAAA-OH-3', and the bottom strand being 5'-OH-CGAGGAGNNNNNACAACGCACAGTGTCTAGT-OH-3'. The DNA Linker contains 14 bp double-stranded DNA (referred to as dsDNA) and 36 nt and 15 nt unhybridized single stranded region which is referred to as the (top-ssDNA and bottom-ssDNA) (Supplementary Figure 3-2g, h). A DNA Ligation Master Mix is made of 4.5 µM of the DNA Linker, 0.2X of 2X Instant Sticky-end Ligase Master Mix (NEB, M0370), 0.8X of 5X Quick Ligase Buffer (NEB, B6058S), 6% (vol/vol) of 1,2-propanediol (Sigma-Aldrich, 398039), and 1 U/µL of RNAse inhibitor). To ligate the DNA Linker with the sticky end DNA, all the nuclei were thoroughly mixed with 250 µL of the DNA Ligation Master Mix. The ligation reaction was kept at 20ºC for 6 hours with an intermittent mixing at 1600 rpm (30 seconds on and 270 seconds off).

Ligation of Cell Barcodes

To phosphorylate the 5' end of the Linker, the nuclei were resuspended in 250 µL of 5' phosphorylation master mix and incubated at 37°C while rotating at 800 rpm for 1 hour. Afterwards, the nuclei were washed once with PBS Wash Buffer 1 and three times with PBS Wash

Buffer 2. The nuclei were resuspended in 900 µL of PBS Wash Buffer 2 with 0.2 U/µL of RNase Inhibitor and filtered through a 10 µM cell strainer (pluriStrainer, 43-10010-50). 6 µL of the nuclei suspension was stained with 6 µL of Ethidium homodimer-1, and the number of nuclei was counted by Countess II Automated Cell Counter (ThermoFisher). A total of 288 barcodes were taken from Hawkins et al[2] and split to three barcode sets, named Barcode Set 1, 2, and 3. Each barcode takes the form of 7 nt_overhang-dsDNA-7 nt_overhang (Supplementary Figure 3-2d, i, Supplementary Table 2). These 288 barcodes are collected called the Cell Barcodes. Three Ligation Master Mixes are prepared, with each Ligation Master Mix containing one set of barcodes (5.4 µM) and 0.2X of 2X Instant Sticky-end Ligase Master Mix, 0.8X of 5X Quick Ligase Buffer, 6% (vol/vol) of 1,2-propanediol, 0.8 U/µL of RNAse inhibitor. The three Ligation Master Mixes are named Ligation Master Mix 1, 2, and 3, corresponding to Barcode Set 1, 2, and 3, respectively.

The first round of split-pool. Up to 100,000 nuclei were collected for split-pool to ensure the majority of nuclei were labeled with unique cell barcode combinations. The nuclei suspension was filled to 1144 µL with PBS Wash Buffer 2 and 24 µL of RNAse inhibitor and subsequently split into 96 wells. To ligate Barcode Set 1 with the RNA Linker and the DNA linker, the nuclei in each well are incubated in the Ligation Master Mix 1 at 20ºC overnight with an intermittent mixing at 1600 rpm (30 seconds on and 270 off). After overnight incubation, the reaction was quenched by an addition of 60 µL of quenching buffer (1X PBS, 50 mM EDTA, 50 mM EGTA, 0.1% Triton X-100) and incubated for 10 min at 20ºC. The nuclei solutions from the 96 wells were pooled together into a 15 mL LoBind tube. 95 µL of quenching buffer was added to each well to rinse and collect any remaining nuclei and pooled into the same 15 mL tube. The nuclei were centrifuged at 900 x g for 4 min, and the nuclei were transferred into a 1.5 mL tube with 0.5 mL of remaining supernatant. 500 µL of PBS Wash Buffer 2 was used to rinse the 15 mL tube and

collect the residual nuclei into the same 1.5 mL tube. The nuclei were washed three times with 900 µL of PBS wash buffer 2 by the centrifugation at 900 x g for 2 min.

The 2nd and 3rd round of split-pool. The pooled nuclei were subjected to the same split-pool procedure as the first round, except that the Ligation Master Mixes 2 and 3 were used for the 2nd and the 3rd rounds, replacing the Ligation Master Mix 1.

Addition of Complex Barcodes. We use a combination of two sets of the barcodes to jointly differentiate individual molecular complexes. This combination is hereafter referred to as the Complex Barcodes. The first set of barcodes are those 3.5 million oligos provided in the Chromium Next GEM Single Cell 3' Reagent kit PN-1000268 (10X Barcodes). Each oligo is an 82-base oligonucleotide with 16 nt barcode and 12 nt unique molecular identifier (10X BC+UMI, Supplementary Figure 3-2j, k). The second set of barcodes is composed of 8 barcodes (Index Barcodes). Each barcode is 8 nt (i7 in Supplementary Figure 3-1h, Supplementary Table 2).

The nuclei were resuspended in 250 µL of 3' Dephosphorylation Buffer (1X PNK buffer, 0.5 U/µL T4PNK, 1 U/µL of RNAse inhibitor) and incubated at 37ºC for 1 hour with rotating at 800 rpm to convert any 2', 3' cyclic phosphate on RNA to 3'-OH. The nuclei were washed once with PBS Wash Buffer 1 and three times with PBS Wash Buffer 2 and centrifuged at 900 x g for 2 min. To add polyA sequences to all the RNA molecules, the nuclei were resuspended in PolyA Tailing Buffer (1X E. coli Poly(A) Polymerase Reaction Buffer, 0.08 U/µL of E. coli Poly(A) Polymerase, 1 mM of ATP, RNAse inhibitor 1U/µL). The mixture was kept at 37ºC for 10 min while rotating at 800 rpm. After adding polyA tails, the nuclei were thoroughly resuspended in 1X PBS with 0.04% BSA (wt/vol) and filtered through a 10 µM cell strainer (pluriStrainer, 43-10010-50) into 1.5 mL tube to obtain isolated nuclei. 6 µL of the nuclei-containing solution was stained with 6 µL of Ethidium homodimer-1 and counted by Countess II Automated Cell Counter. 5,000 single nuclei were transferred to a Covaris microtube-15 and filled to 15 µL with 1X PBS with 0.04% BSA (wt/vol). The nuclei were sonicated using Covaris M220 Focused-ultrasonicator with water temperature 6ºC, incident power 50 W, duty factor 5 for 5 min to release chromatin complexes.

To add the 10X barcodes to the polyadenylated RNA and the top-ssEND of the DNA Linker, the sonicated nuclei complexes were transferred into a 1.5mL LoBind tube and mixed with 25 µL of water, 18.8 µL of RT reagent B, 2 µL of Reducing Agent B, 8.7 µL of RT enzyme C. The mixture was transferred to one well in ChIP G, which is loaded to the 10X Chromium controller according to Steps 1.1 to 1.5 in the protocol of Chromium Next GEM Single Cell 3' Reagent kit. The retrieved droplets were transferred into a PCR tube for cDNA synthesis according to the 10X protocol. The droplets were broken, and the aqueous phase was obtained according to Step 2.1 in the 10X protocol.

The aqueous phase containing nucleic acids were filled to 200 µL with nuclease free water and split into eight aliquots in LoBind 1.5 mL tubes. 25 µL of 2X reverse crosslinking buffer (400 mM NaCl, 0.4% SDS, 50 mM EDTA, 50 mM EGTA, 0.04 U/µL Proteinase K) was added into each tube, and reverse crosslinking reaction was incubated at 50ºC for 2 h and 55ºC overnight with shaking at 800 rpm. In each aliquot, the reverse crosslinked nucleic acids were purified with Monarch RNA purification kit (NEB, 76307-460) and eluted into 21 µL of nuclease free water. The eluted DNA and RNA molecules were incubated at 55ºC for 15 min with 1X isothermal amplification buffer II, 0.32 U/µL of Bst 3.0 DNA polymerase, additional 6 mM MgSO4, 1.4 mM dNTP Mix, and 0.5 U/µL of RNasin. The product was purified with 1.8X RNA clean Ampure beads (Beckman Coulter life science, A63881) and eluted into 20 µL of nuclease free water. PCR is carried out for each aliquot with 2.5 µL of 10 µM shared Universal Adaptor (P5 and Read1 in Supplementary Figure 3-1h) and 2.5 µL of 10 µM aliquot-specific primer in 25 µL of NEBNext Ultra II Q5 Master Mix. The aliquot-specific primers are the eight Index Adaptors (see Reagents created by this work). PCR is carried out at 13 to 14 cycles. The amplified DNA was purified with 1.2X Ampure beads and eluted into 12.5 µL of nuclease free water. The purified DNA solutions from the eight aliquots were combined and loaded into 5 lanes of 4% E-gel (Invitrogen, G401004). The DNA bands with the size between 300 bp and 1.2 Kb were excised. The DNA was extracted with NEB Monarch gel purification kit (NEB, T1020S) using two columns and eluted in 30 µL of the elution buffer.

Sequencing

The molarity of the sequencing library was measured by Qubit 4.0 Fluorometer (Invitrogen, Q33238) using Qubit dsDNA HS assay kit (Invitrogen, Q33231). Fragment size distribution was assessed by Agilent bioanalyzer with High sensitivity DNA ChIP. The library was sequenced by UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000. The sequencer is set to read a 28 bp sequence next to the Universal Adapter as the Read1, a 8 bp index sequence from the I7 region inside the Index Adapter, and a 150 bp sequence next to the Index Adapter as the Read2.

The MUSIC-docker data processing pipeline

We developed MUSIC-docker to process MUSIC sequencing data using Docker to encapsulate a Snakemake (Köster and Rahmann, 2012) pipeline, ensuring cross-platform execution. It handles i7 index-splited paired-end fastq files, processes them into RNA and DNA sequences separately, adds cell and complex barcodes, and maps to the genome, removing PCR duplicates and deriving processed files. Detailed documentation is available at: http://sysbiocomp.ucsd.edu/public/wenxingzhao/MUSIC_docker/intro.html, and Supplemental Note 14.

The raw sequencing output (.bcl) is converted to FASTQ files with bcl2fastq, producing eight FASTQ files with 28bp Read1 and 150bp Read2. Read1 includes the 10X Barcode and 10X UMI, while Read2 contains Cell Barcodes, RNA/DNA Linker, and RNA/DNA insert sequences.

Demultiplexing step extracts cell and complex barcodes information and fragment sequence information, creating separate FASTQ files for RNA and DNA reads individually where the read sequence will be the insert, and the read name will be the fragment identity information.

To address potential artificial sequences introduced by sequencing errors and experimental design, we remove consecutive As or Gs from the 3’ of the DNA inserts if their length is longer

than 20bp. For RNA reads, we will detect the ssDNA region of the RNA Linker sequence "CGAGGAGCGCTT" and remove any sequence following it. We use cutadaptor (v2.8) (Martin, 2011) with parameter `-q 15 -m 20`. DNA and RNA inserts are mapped to the genome using bowtie2 (v5.4.0) (Langmead & Salzberg, 2012) with parameters "bowtie2 -p 10 -t --phred33 -x " and bwa (v0.7.17) (Korsunsky et al., 2019) mem with parameters "-SP5M", respectively. Uniquely mapped reads are selected for downstream analysis.

After reads mapping, PCR duplicates, identified by shared 10X UMI and mapped coordinates are removed using customized script. The script sorts the BAM file, scans it once, and flags duplicates if they meet specific criteria: 1) it maps to a location within 8 base pairs of the previous read, 2) it shares the same cell barcode and molecular barcode with the previous read, and 3) its UMI exhibits a Levenshtein distance of less than 2 base pairs from the UMI of the previous read. All identified PCR duplicates are subsequently removed to ensure the integrity of downstream analyses.

Finally, deduplicated BAM files from each I7 index library are merged into a comprehensive, sorted BAM file, capturing essential information for both DNA and RNA, including cell barcode, molecular barcode (10X and I7 index), and insert mapping location.

Mix species experiment

Based on the published method (Arrastia et al., 2022a), we assigned a cell to a species if 95% of all its DNA reads could map to a single species. Cells with uniquely mapped non-duplicated DNA reads less than 1000 are classified as ambient cells. To calculate the single cluster mix species rate, we assigned a cluster to a single species if more than 99% of its uniquely mapped non-duplicated DNA reads come from that species.

Parsing multiplex clusters into pairwise interactions with normalization

Each cluster is a collection of reads that share the same Cell Barcodes, 10X Barcode and I7 Barcode (CB+GEM+I7). A cluster of two reads (cluster size = 2) corresponds to a pairwise interaction. A cluster of three or more reads corresponds to a multiplex cluster. Each multiplex cluster is decomposed into pairwise interactions with a normalization procedure that adjusts for the total number of combinations of pairwise interactions from a multiplex cluster as previously described (Arrastia et al., 2022a; Quinodoz et al., 2018). For homotopical clusters, which are clusters containing exclusively DNA reads or RNA reads, each homological cluster of size N is first decomposed into all non-overlapping pairs, and then each pair is normalized by a factor of 1/N. For heterotypic clusters, which are clusters containing both DNA and RNA reads, each heterotypic cluster is first decomposed into all DNA-RNA pairs, and then each pair is normalized by a factor of 1/(M+N) where M and N are the numbers of DNA and RNA reads. This normalization removes the difference in the number of pairwise decompositions from different-sized clusters, thus ensuring the larger clusters do not get an inflation in the number of decomposed pairwise interactions (Arrastia et al., 2022a; Quinodoz et al., 2018).

To generate a two-dimensional contact map of DNA-DNA interactions, we first define the size of each unit, typically represented as genomic bins, for rows and columns. The weight assigned to bin [i, j] is determined by the total number of clusters that contain DNA reads mapped to both the ith and jth bins, and it is normalized by the cluster size as previously described. To compare DNA-DNA contact differences within a specific genomic region, cluster sizes are calculated using only DNA reads mapped to that region, considering small, medium, and large clusters. If the DNA-DNA contact map is generated for multiple cells, the weighted sum of all clusters is calculated. Similar methods are applied to derive RNA-DNA 2D contact maps, with the calculation of the weight of each interaction adjusted accordingly.

To determine the RNA attachment level (RAL) at specific genomic bins, representing 1D RNA-chromatin contacts for a particular RNA of interest, we calculate the total weighted RNA-DNA interactions

involving the RNA of interest and DNA ends mapped to the desired genomic bins. For ensemble maps, RAL values from individual cells are summed.

Finally, for visualizing the two-dimensional contact maps of DNA-DNA or RNA-DNA interactions, the raw contacts, obtained following the previously mentioned procedures, are scaled up using a linear factor, typically 100 or 1000. This amplification step prevents the application of logarithmic transformation to decimal values. After amplification, a logarithmic transformation is performed to enhance visualization of the contact maps.

Comparison between MUSIC DNA-DNA contacts with Micro-C data

In order to achieve consistency in the total contacts scale between MUSIC and Micro-C, contact maps obtained from both methods underwent a standardized transformation. Initially, the contact maps derived from Micro-C or MUSIC were logarithmically scaled. Subsequently, the scaled contact maps were normalized to their respective maximum values, resulting in all contacts being constrained within the range of [0, 1]. The Micro-C data used in this study was obtained from the 4DN portal (Reiff et al., 2022) (4DNFI2TK7L2F) (Krietenstein et al., 2020). To extract the raw contacts from the .hic file, the straw (Durand, Robinson, et al., 2016) tool was employed.

To calculate the compartment score (PC1 score) for both Micro-C and MUSIC DNA-DNA contact matrices, we first computed the Expected Contact Matrix. Subsequently, we determined the PC1 score from the correlation matrix of the Observed/Expected ratio matrices, where the Expected Matrix was derived by calculating average contact frequencies as a function of genomic distances. For the comparison of PC1 correlations between Micro-C and MUSIC, the calculations were performed individually on each chromosome. The reported correlation represents the median of these correlations across all chromosomes.

nsaRNA and pre-mRNA RAL

Pre-mRNAs are RNA reads that exhibit an overlap of at least 15 base pairs with gene introns and are classified as protein-coding RNAs. The calculation of the RD (RNA-DNA) cluster weight involves the inclusion of all nsaRNAs and pre-mRNAs, along with their associated DNA reads. RD clusters with a cluster size not exceeding 1000 were selected for the computation of the

RNA attachment level (RAL). The H1 A/B compartment data used in this study was obtained from the 4DN data portal (4DNFID162B9J) (Calandrelli et al., 2023).

Calculate genomic distance versus contact frequency curve from MUSIC data

The relationship between contact frequency and genomic distances within chromosomal arms was systematically examined. Initially, the genomic distance ranges from 10 base pairs to 150 megabases was divided into 500,000 equally sized bins. Subsequently, for DNA reads originating from DD or RD clusters, the genomic distances of all intrachromosomal pairwise interactions were determined, and the frequencies for each genomic distance bin were computed. These frequencies were then normalized based on the weight assigned to each interaction. The normalized frequencies were calculated for each genomic bin in every cluster and single cell. Finally, the genomic distances versus contact frequencies were aggregated across all clusters from all single cells.

We used cooltools (0.5.4) (Open2C et al., 2022) for generating the genomic distance versus contact frequency curve in Micro-C data. We downloaded the .mcool file for H1 Micro-C from the 4DN data portal (4DNFI9GMP2J8) (Krietenstein et al., 2020).

Brain single cells preprocessing and filtering

For each brain sample we applied standard MUSIC_tool docker pipelines to get valid RNA and DNA reads information for each single cell. To select high quality single cells for robust analysis and interpretation of our data, we removed cells that have less than 100 RNA reads, or less than 5000 DNA reads for downstream analysis on brain samples.

Brain sample transcriptome merging

We first constructed the single cell RNA expression count matrix by calculating the number of RNA reads mapped to each human gene (GENCODE (Frankish et al., 2021), v36, chrM genes are excluded) for all single cells from all 14 brain samples. We then constructed one Seurat object

(Seurat 4.3.0 (Hao et al., 2021; Stuart et al., 2019)) for each brain sample with parameters min.cells = 2 and min.features = 200 (filtered out genes that express in no more than 2 cells, and filtered out cells that have no more than 200 expressed genes). The count matrix from all brains is then integrated together using *RunHarmony* from harmony R package (0.1.1 (Korsunsky et al., 2019)) based on STransform processed data and regressed out on factors including individual library and experimental batches.

Brain gene expression

To compare gene expression in brain cells, we use the LogNormalize method from the Seurat R package. Raw reads counts were first normalized by the library size and then log-transformed.

Single cell clustering and cell type identification

The integrated brain object is then subjected to dimensionality reduction using UMAP methods based on the first 20 principal components from PCA using the Seurat R package. All the cells are then clustered in an unsupervised method using a shared nearest neighbor graph based on k-nearest neighbors (k=20) calculated from the top two coordinates of UMAP. Then the clusters were derived by optimizing the modularity function using the function *FindClusters* with parameters resolution at 0.05. ExN and InN are clustered by first extracting the subset of cells and re-clustering at resolution of 1.

We assigned the cell types to each cluster by the known cell type specific marker gene expression level. Cell type specific marker genes are based on previous publications. For major brain cell types (Mathys et al., 2019), subclusters (Azimuth) (Bakken et al., 2021), and Vascular cell (Franzén et al., 2019). For each cluster, it is assigned to one cell type (A) if two times the average expression of all marker genes of cell type A plus the proportion of cells in the cluster expressing cell type A's marker genes are larger than any other cell types. We designed a score

that takes both the cell type marker gene expression level and proportion of cells expressing the genes into consideration but with a higher emphasis on expression level.

Single cell LCS-erosion score and transcriptomic age calculation

We calculated the Local Chromatin Structure (LCS) erosion score for each cell by determining the middle point of the genomic distances of the top 10 most frequently contacting genomic bins. To derive the contact frequency versus genomic distances heatmap for all frontal cortex cells, we first generated 149 genomic bins spanning from 5000 bp to 150M bp with the *n* th bin spanning the genomic region ($2^{log2(5000)+\frac{\{log_2(1.5*10^8\}-log_2(5000))}{150}\times(n-1)}$, $2^{log2(5000)+\frac{\{log_2(1.5*10^8\}-log_2(5000))}{150}\times n}$] where n=1,2,3,..,150. Subsequently, for DNA reads originating from DD or RD clusters, the genomic distances of all intrachromosomal pairwise interactions were determined, and the frequencies for each genomic distance bin were calculated. These frequencies were then normalized against the bin size. For each single cell the total contact frequency was normalized against the total frequency within each cell before plotting the heatmap. Cells exhibiting an LCS-erosion score exceeding 3e5 were classified as LCS-eroded cells, while others were considered LCS-preserved.

For assessing the transcriptomic-based biological age of cells, we implemented the SCALE methodology described by previous study (Mao et al., 2023). This approach involves using a list of human aging-associated genes obtained from https://sysomics.com/AgingMap/. Initially, we determined the direction (either +1 or -1) of each marker gene based on its correlation with the chronological age of the sample. This direction depended on whether the gene expression was positively (+1) or negatively (-1) correlated with age. Subsequently, for each marker gene, we assigned a weight computed as the proportion of cells expressing the gene, multiplied by its

directional value. Finally, we calculated the transcriptomic age for each cell using the dot product of the gene expression z-scores and the corresponding gene weights.

LCS-erosion score associated transcriptome functions

To identify genes whose expression is significantly associated with chromatin LCS-erosion score, we applied ANOVA test between each gene's expression and LCS-erosion score across all cells. Genes with p value < 0.01, F value > 1 and the absolute value of spearman correlation between LCS-erosion score and gene expression larger than 0.1 are considered significant. To identify cell-type specific LCS-erosion score associated genes, we applied ANOVA test within each cell type individually. We used R package *gprofiler2* (Kolberg et al., 2020) for pathway enrichment analysis. Enriched wikipathway, KEGG and REACtom pathways are shown.

eQTL analysis

All brain cell-type specific eQTL data were downloaded from (Bryois et al., 2022). We composed a combined dataset by selecting eQTLs with nominal p-value < 10^-4 and removing eQTLs in endothelial cells and pericytes. MUSIC DD contacts between eQTL and their target genes are read pairs where one end is mapped (1-base overlap) over the eQTL and the other end over the gene promoter (2.5 kb flanking region from the TSS). The global Chi-square test was performed on a 6 x 6 table to test the overall association of cell-type specific DD contacts with cell-type specific eQTL-gene pairs. The 6 x 6 table was then reduced to six 2 x 2 tables to test that association in each cell type (Chi-square test). The 95% confidence interval of the odds ratio (OR) was calculated as exp(log(OR) +/- 1.96*SELOR), where SELOR is the standard error of the log(OR).

XIST-chromatin interactions analysis

To derive the one dimensional XIST-genome attachment level plot (RAL), we extracted all clusters that contained at least one XIST RNA. For each cluster with M XIST RNA reads and

N DNA reads, we derived M*N paired XIST-DNA interactions, each of these interactions were then normalized according to their cluster size 1/(M+N). We then binned the whole genome with a 1Mb sized bin. Then the XIST RAL for each bin is the total weights of all XIST-DNA interactions with DNA ends overlapped with that bin.

To derive the 2D RNA-DNA contact map for chrX, we extracted all RNA and DNA reads that can map to chrX. Next, for any matched molecular complex barcode with M chrX RNA reads and N chrX DNA reads, we again first derive all combinations of RNA-DNA interactions with adjusted weight indicating the reverse of cluster size (1/M+N). We binned chrX at 1Mb resolution. The M[i, j] then represents the sum of weighted interactions whose RNA ends mapped to the i th bin and DNA ends mapped to the j th column.

XAL stratification and corresponding genomic distance versus contact frequency

XIST-chrX association levels (XAL) is numerically represented as the number of XIST attached chrX DNA bins, each individual cell will have one XAL value. To assess the differences of chromatin organization between XIST+ and XIST- clusters, we stratified all brain cells into four groups based on the XAL value. Group1 with XAL equals 0 will be all cells with no detectable XIST-chrX association. Group2 to Group4, which with increasing XAL value, are taking all cells that have XIST-chrX association and dividing them equally into 3 groups. To derive the genomic distances and contact frequency relationship we follow the same methods introduced in “Calculate genomic distance versus contact frequency curve from MUSIC data " section.

Acknowledgments

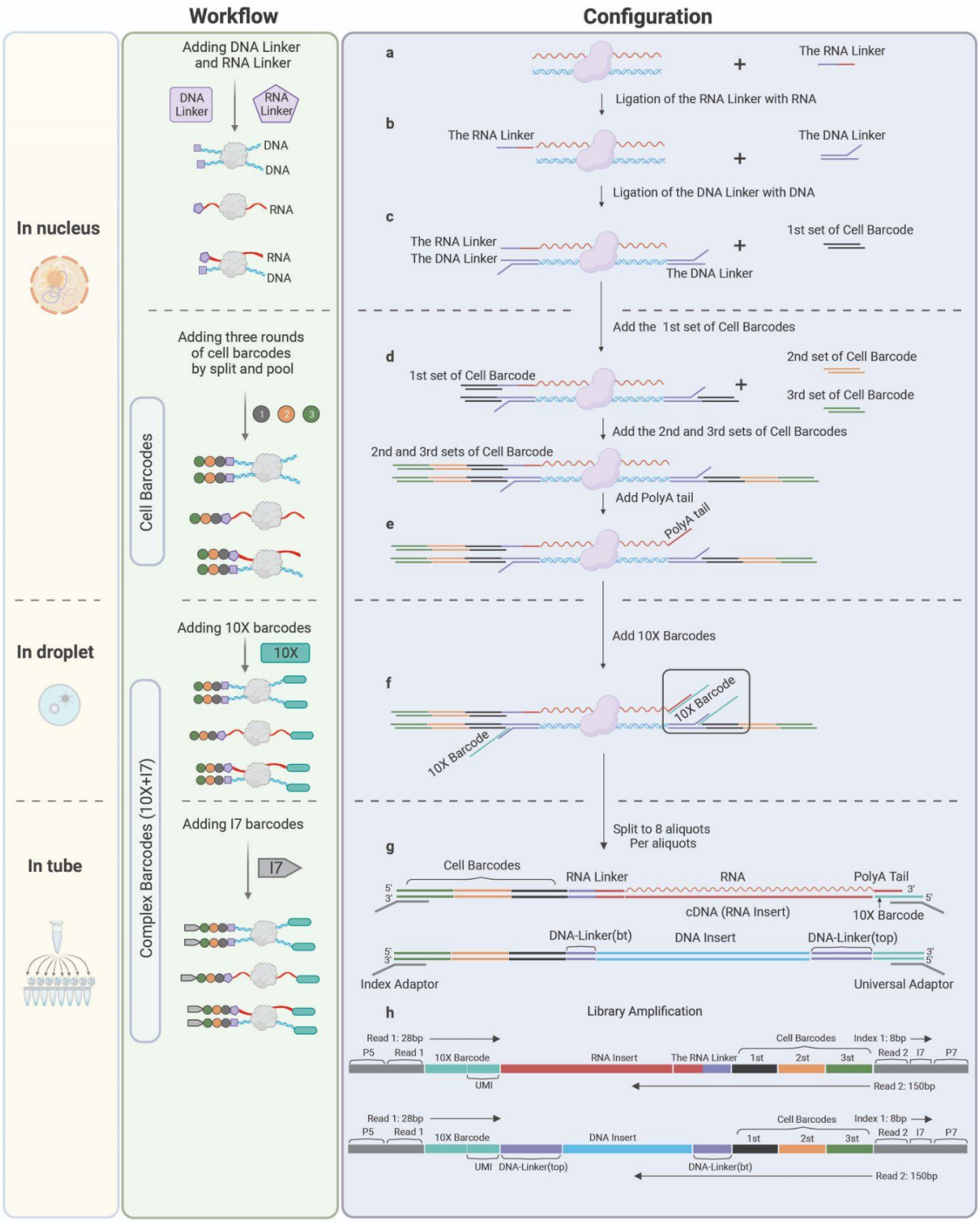
Chapter 3, in full, is a reprint of the material as it appears in Nature 2024. Wen, X., Luo, Z., Zhao, W., Calandrelli, R., Nguyen, T. C., Wan, X., Charles Richard, J. L., & Zhong, S. (2024).

Single-cell multiplex chromatin and RNA interactions in ageing human brain. The dissertation author was co-first author of this paper.

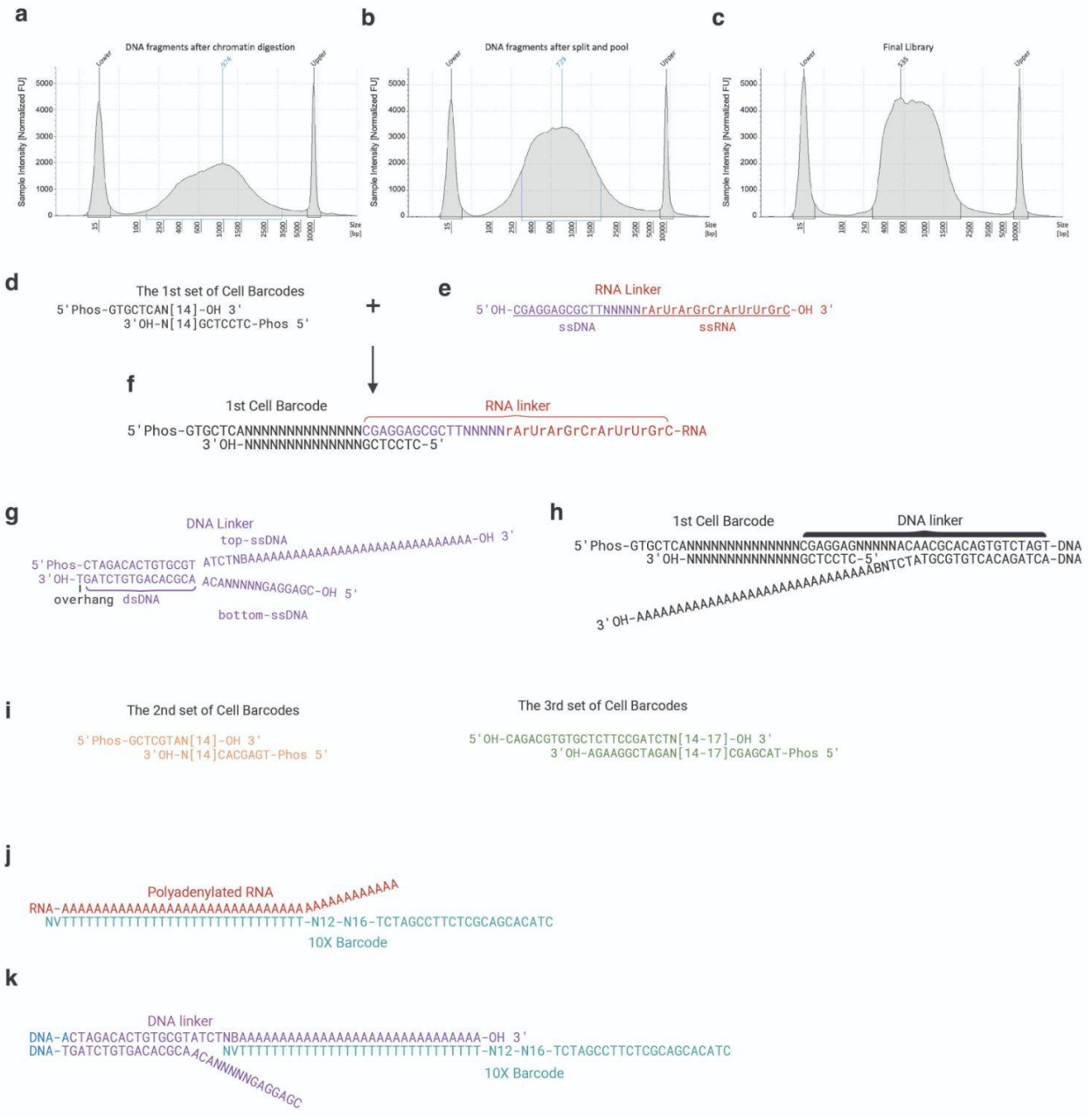
We are grateful to the Banner Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona, for the provision of human biological materials. The Brain and Body Donation Program has been supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinson's Disease and Related Disorders), the National Institute on Aging (P30 AG019610 and P30AG072980, Arizona Alzheimer's Disease Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimer's Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinson's Research. We thank UCSD IGM Genomics Center for support with sequencing. This work is funded by NIH grants DP1DK126138, R01GM138852, UH3CA256960, U01CA200147 and R01HD107206, and by a Kruger Research grant.

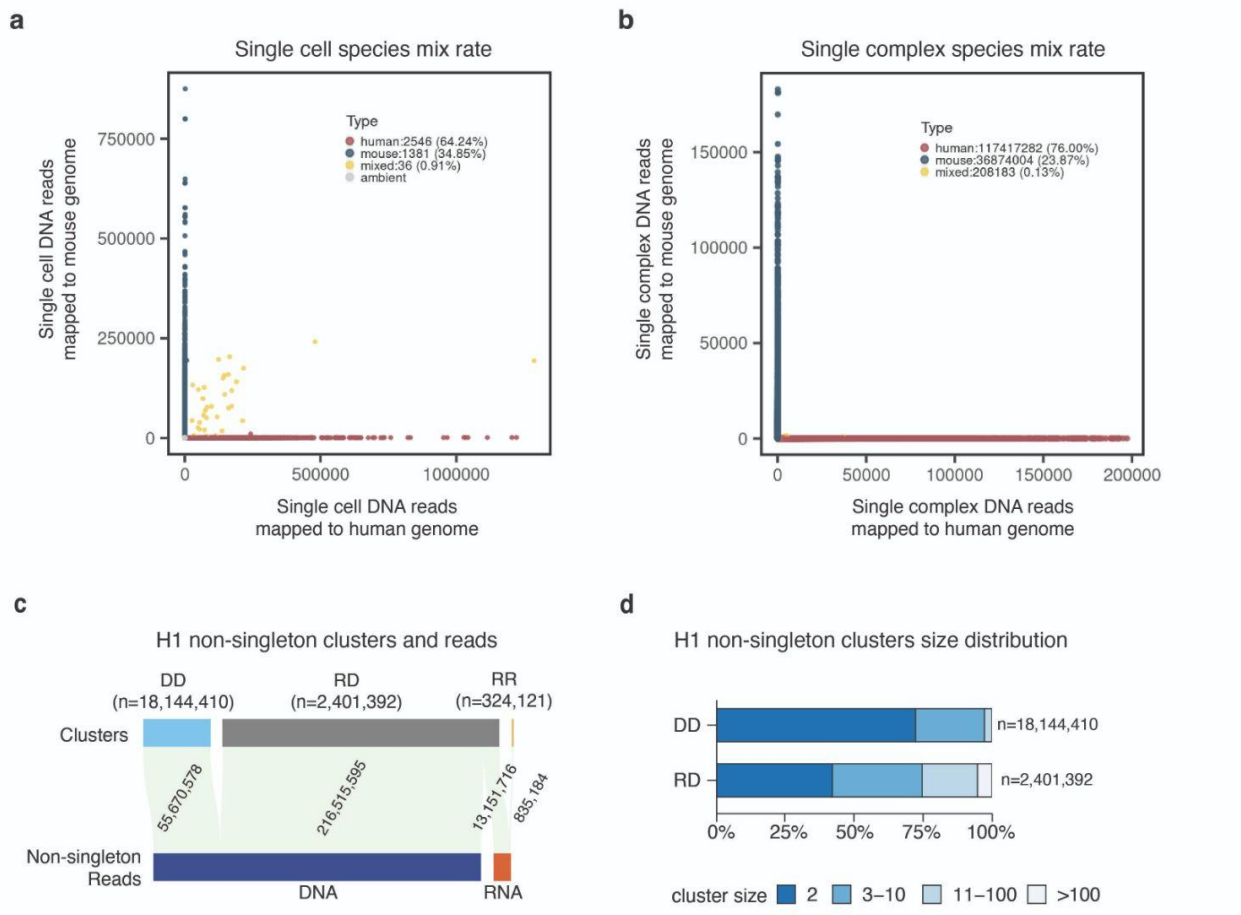
Supplementary information

Supplementary Figure 3-1: Overview of the MUSIC method.
(a) An example chromatin complex with one associated DNA fragment (blue) and one RNA molecule (red) is shown to illustrate the procedure. Other chromatin complexes with any number of associated DNA fragments or RNA molecules are expected to react accordingly. The RNA Linker with half single strand RNA (ssRNA, red) and half single strand DNA (ssDNA, purple). (b) The ligation product with a RNA linker conjugates with the Y-shaped DNA Linker that contains a double strand region (dsDNA) and single strand regions (top-ssDNA and bottom-ssDNA). (c) The ligation product with both DNA and RNA linkers. (d-e) Addition of the three sets of Cell Barcodes (color coded in black, yellow, green) and A-tail. (f) Reaction in a 10X GEM system. (g) Aliquoting the output of the 10X GEM system for library preparation. (h) The sequence configuration of the constructed sequencing library. For Illumina paired-end sequencing, Read1 covers the 28 bp 10X Barcode that consists of a 16 bp 10X GEM Barcode and 12 bp 10X UMI. Index 1 covers the 8 bp I7 Barcode, and Read2 covers the Cell Barcodes as well as the RNA Linker and the RNA insert or the DNA Linker and the DNA insert.

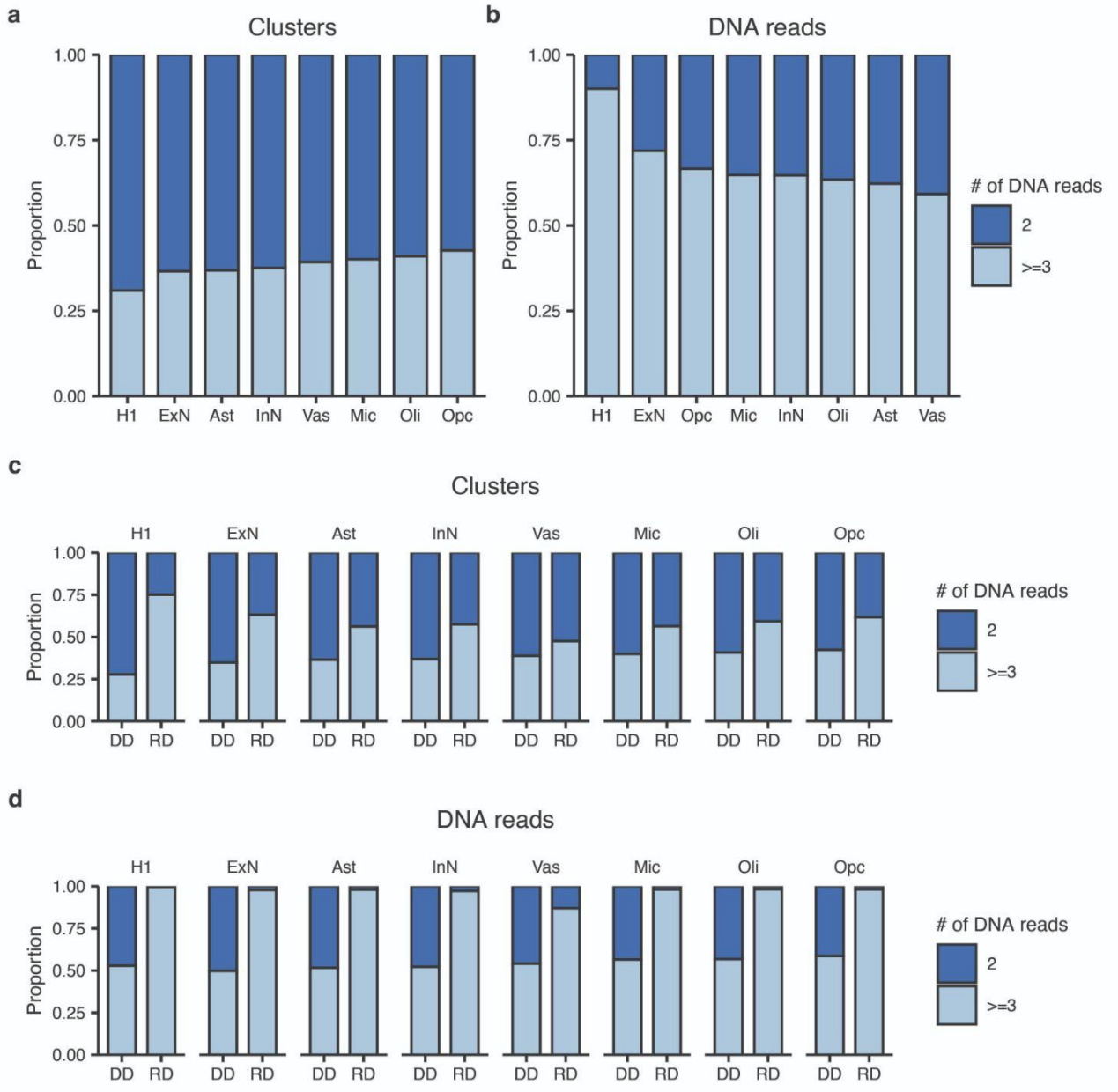

Workflow
Configuration
In nucleus
In droplet
In tube
Adding DNA Linker and RNA Linker
DNA Linker
RNA Linker
DNA
DNA
RNA
RNA
DNA
Adding three rounds of cell barcodes by split and pool
1
2
3
Cell Barcodes
Adding 10X barcodes
10X
Adding I7 barcodes
I7
Complex Barcodes (10X+I7)
a
The RNA Linker
Ligation of the RNA Linker with RNA
b
The RNA Linker
The DNA Linker
Ligation of the DNA Linker with DNA
c
1st set of Cell Barcode
The RNA Linker
The DNA Linker
The DNA Linker
Add the 1st set of Cell Barcodes
d
2nd set of Cell Barcode
1st set of Cell Barcode
3rd set of Cell Barcode
Add the 2nd and 3rd sets of Cell Barcodes
2nd and 3rd sets of Cell Barcode
Add PolyA tail
e
PolyA tail
Add 10X Barcodes
f
10X Barcode
10X Barcode
Split to 8 aliquots Per aliquots
g
Cell Barcodes
RNA Linker
RNA
PolyA Tail
cDNA (RNA Insert)
10X Barcode
DNA-Linker(bt)
DNA Insert
DNA-Linker(top)
Index Adaptor
Universal Adaptor
h
Library Amplification
Read 1: 28bp
P5
Read 1
10X Barcode
UMI
RNA Insert
The RNA Linker
Cell Barcodes
1st
2st
3st
Index 1: 8bp
Read 2
I7
P7
Read 2: 150bp
Read 1: 28bp
P5
Read 1
10X Barcode
UMI
DNA-Linker(top)
DNA Insert
Cell Barcodes
1st
2st
3st
Index 1: 8bp
Read 2
I7
P7
DNA-Linker(bt)
Read 2: 150bp

Supplementary Figure 3-2: Critical reagents and reactions.
(a-c) DNA size distribution of DNA fragments (a) after chromatin digestion, corresponding to Fig. 1a, (b) after three rounds of cell barcode ligation, corresponding to Supplementary Figure 1e, and (c) in the final library, corresponding to Supplementary Figure 1h. (d) the 1st set of Cell Barcodes. (e) The RNA Linker with half single strand RNA (ssRNA, red) and half single strand DNA (ssDNA, purple) (f) Base-pairing assisted ligation of the ssDNA (purple) in the RNA Linker with the 1st set of Cell Barcodes (black). (g) The Y-shaped DNA Linker with a double strand region (dsDNA) and single strand regions (top-ssDNA and bottom ssDNA) with a 3'dT overhang. (h) The DNA Linker (purple) is ligated with 3' dA-tailed double-stranded genomic DNA (blue) and with the 1st set of Cell Barcodes (black). (i) the 2nd and the 3rd sets of Cell Barcodes. The 3rd set of Cell Barcodes contains a fraction of Illumina Read2 adapter sequence. (j) Hybridization of the RNA's polyA tail (red) with the poly(dT) region of the 10X Barcodes (turquoise). (k) Hybridization of the top-ssDNA of the DNA linker (purple) with the poly(dT) region of the 10X Barcodes (turquoise).


a
DNA fragments after chromatin digestion
Sample Intensity [Normalized FU]
Size [bp]
b
DNA fragments after split and pool
c
Final Library
d
The 1st set of Cell Barcodes
5'Phos-GTGCTCAN[14]-OH 3'
3'OH-N[14]GCTCCTC-Phos 5'
+
e
RNA Linker
5'OH-CGAGGAGCGCTTNNNNNrArUrArGrCrArUrUrGrC-OH 3'
ssDNA
ssRNA
f
1st Cell Barcode
RNA linker
5'Phos-GTGCTCANNNNNNNNNNNNNNNCGAGGAGCGCTTNNNNNrArUrArGrCrArUrUrGrC-RNA
3'OH-NNNNNNNNNNNNNNNGCTCCTC-5'
g
DNA Linker
top-ssDNA
5'Phos-CTAGACACTGTGCGT ATCTNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA-OH 3'
3'OH-TGATCTGTGACACGCA ACANNNNNGAGGAGC-OH 5'
overhang dsDNA
bottom-ssDNA
h
1st Cell Barcode
DNA linker
5'Phos-GTGCTCANNNNNNNNNNNNNNNCGAGGAGNNNNNACAACGCACAGTGTCTAGT-DNA
3'OH-NNNNNNNNNNNNNNNGCTCCTC-5'
3'OH-AAAAAAAAAAAAAAAAAAAAAAAAAABNTCTATGCGTGTCACAGATCA-DNA
i
The 2nd set of Cell Barcodes
5'Phos-GCTCGTAN[14]-OH 3'
3'OH-N[14]CACGAGT-Phos 5'
The 3rd set of Cell Barcodes
5'OH-CAGACGTGTGCTCTTCCGATCTN[14-17]-OH 3'
3'OH-AGAAGGCTAGAN[14-17]CGAGCAT-Phos 5'
j
Polyadenylated RNA
RNA-AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
NVTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT-N12-N16-TCTAGCCTTCTCGCAGCACATC
10X Barcode
k
DNA linker
DNA-ACTAGACACTGTGCGTATCTNBAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA-OH 3'
DNA-TGATCTGTGACACGCA ACANNNNNGAGGAGC NVTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT-N12-N16-TCTAGCCTTCTCGCAGCACATC
10X Barcode

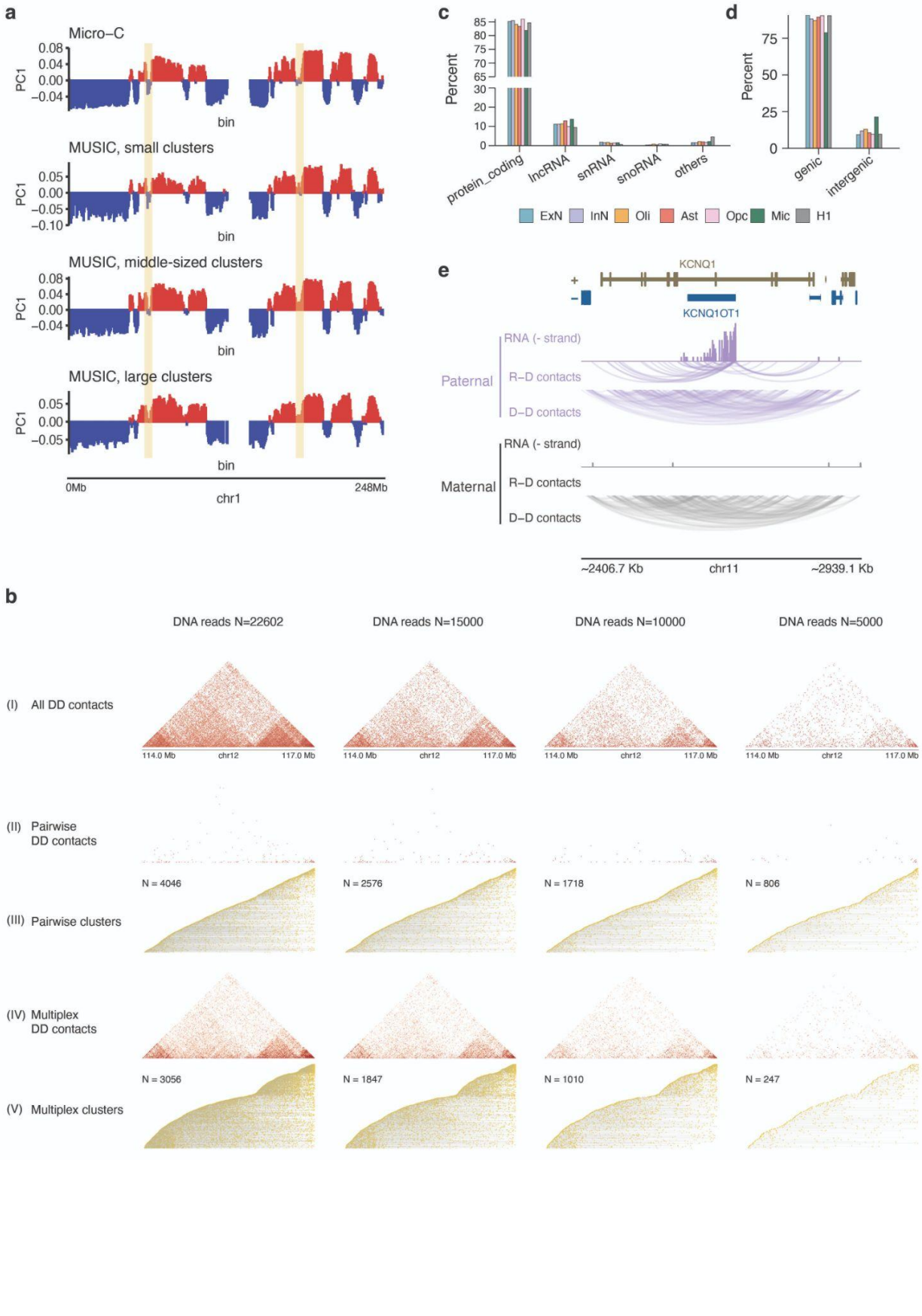


Supplementary Figure 3-3: Mixed-species analysis.
(a) The cells (dots) with more than 1000 reads (read count filter = 1000) are colored coded to red (human), blue (mouse) if 95% or more of the reads are mapped to a single species (purity filter = 95%), or yellow (mixed) if otherwise. The cells with less than 1000 reads (gray dots, ambient) are not used in the calculation of mixed-species rate. (b) The complexes (dots) are colored coded to red (human), blue (mouse) when 99% or more of the reads are mapped to a single species, or yellow (mixed) if otherwise. The complexes with 0 - 200000 DNA reads are plotted. (c) The distribution of DNA reads and RNA reads in DNA-DNA (DD), RNA-DNA (RD), and RNA-RNA (RR) clusters. (d) Cluster size distributions of DD and RD clusters.

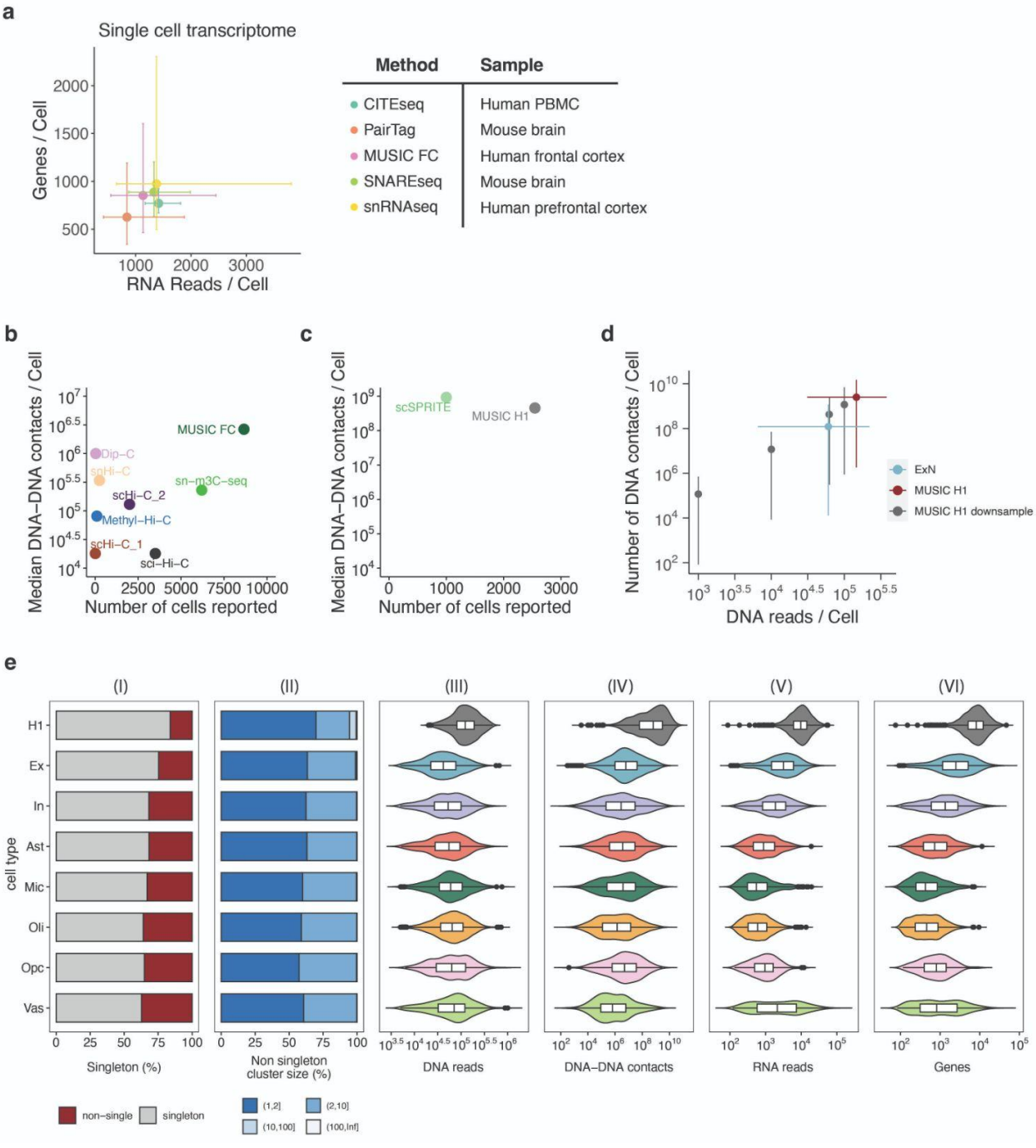


Supplementary Figure 3-4: Proportions of pairwise or multiplex interactions.
The proportions of clusters (a) and DNA read counts (b) that consist of multiplex chromatin interactions (DNA reads >=3, light blue) in each cell type. Both RD and DD clusters are included in this analysis. (c) Proportions of clusters with pairwise (2 DNA reads, dark blue) or multiplex chromatin interactions (3 or more DNA reads, light blue) in DNA-only (DD) and RNA-DNA (RD) clusters in each cell type. (d) Proportions of DNA read counts with clusters corresponding to pairwise (dark blue) or multiplex (light blue) chromatin interactions.

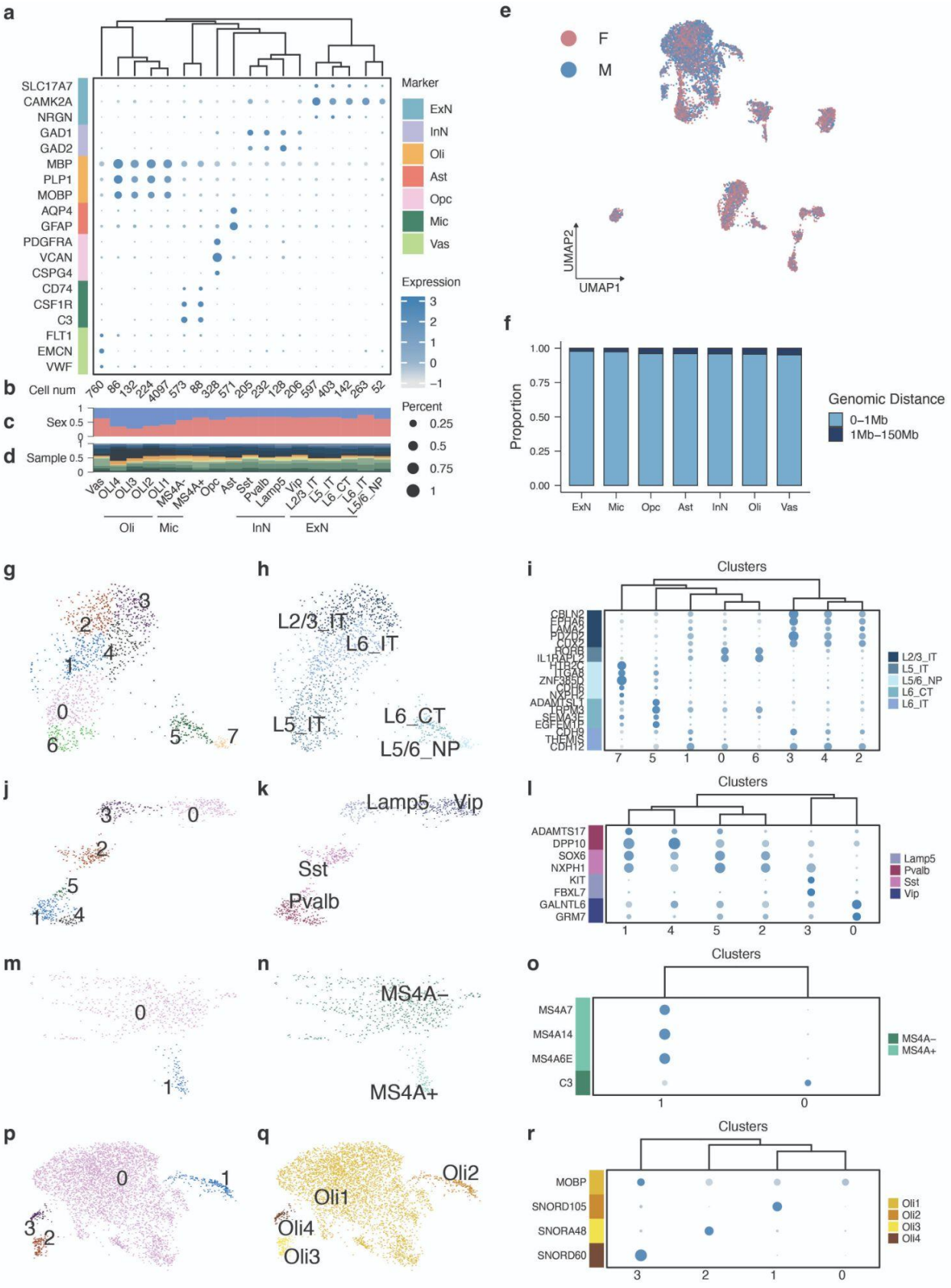
Supplementary Figure 3-5: Chromatin contacts and RNA in H1 cells.
(a) Compartment scores (PC1) across the entire Chromosome 1 at 500 Kb resolution. PC1s are calculated from Micro-C, MUSIC's small clusters (2–10 DNA reads), middle-sized clusters (10–50 DNA reads), and large clusters (50–100 DNA reads). Positive and negative PC1s correspond to A and B compartments. Two small B compartments in MUSICs' small clusters became indistinguishable in the larger clusters (vertical yellow stripes), corresponding to slightly fewer A/B compartment switches in the small clusters. (b) Downsampling the original 22,602 UMNDBC DNA reads mapped to this region (Chr12:114-117 Mb) to 15,000, 10,000, and 50,000 reads (columns). Rows from the top to the bottom are 2D contact map of all DD contacts that include multiplex and pairwise contacts (I), 2D contact map of pairwise DD contacts (II), a cumulative view of every DD cluster in pairwise complexes (III), 2D contact map of multiplex DD contacts (IV), and a cumulative view of every DD cluster in multiplex complexes (V). (c) The proportions of each type of RNA species (y axis) in each cell type (color coded). (d) The proportions of genic and intergenic RNA reads in each cell type. (e) Genome track view of the RNA reads (RNA), RNA-DNA (R-D) contacts, and DNA-DNA (D-D) contacts in the paternal (top) and maternal (bottom) allele in H1 cells. All the RNA reads shown are those mapped to the KCNQ1OT1 gene in the strand consistent with the direction of transcription of the KCNQ1OT1 gene (- strand).


a
Micro−C
MUSIC, small clusters
MUSIC, middle-sized clusters
MUSIC, large clusters
PC1
bin
0Mb
248Mb
chr1
c
Percent
protein_coding
lncRNA
snRNA
snoRNA
others
ExN
InN
Oli
Ast
Opc
Mic
H1
d
Percent
genic
intergenic
e
KCNQ1
KCNQ1OT1
RNA (- strand)
Paternal
R−D contacts
D−D contacts
Maternal
~2406.7 Kb
chr11
~2939.1 Kb
b
DNA reads N=22602
DNA reads N=15000
DNA reads N=10000
DNA reads N=5000
(I) All DD contacts
(II) Pairwise DD contacts
(III) Pairwise clusters
(IV) Multiplex DD contacts
(V) Multiplex clusters
N = 4046
N = 2576
N = 1718
N = 806
N = 3056
N = 1847
N = 1010
N = 247
114.0 Mb
chr12
117.0 Mb

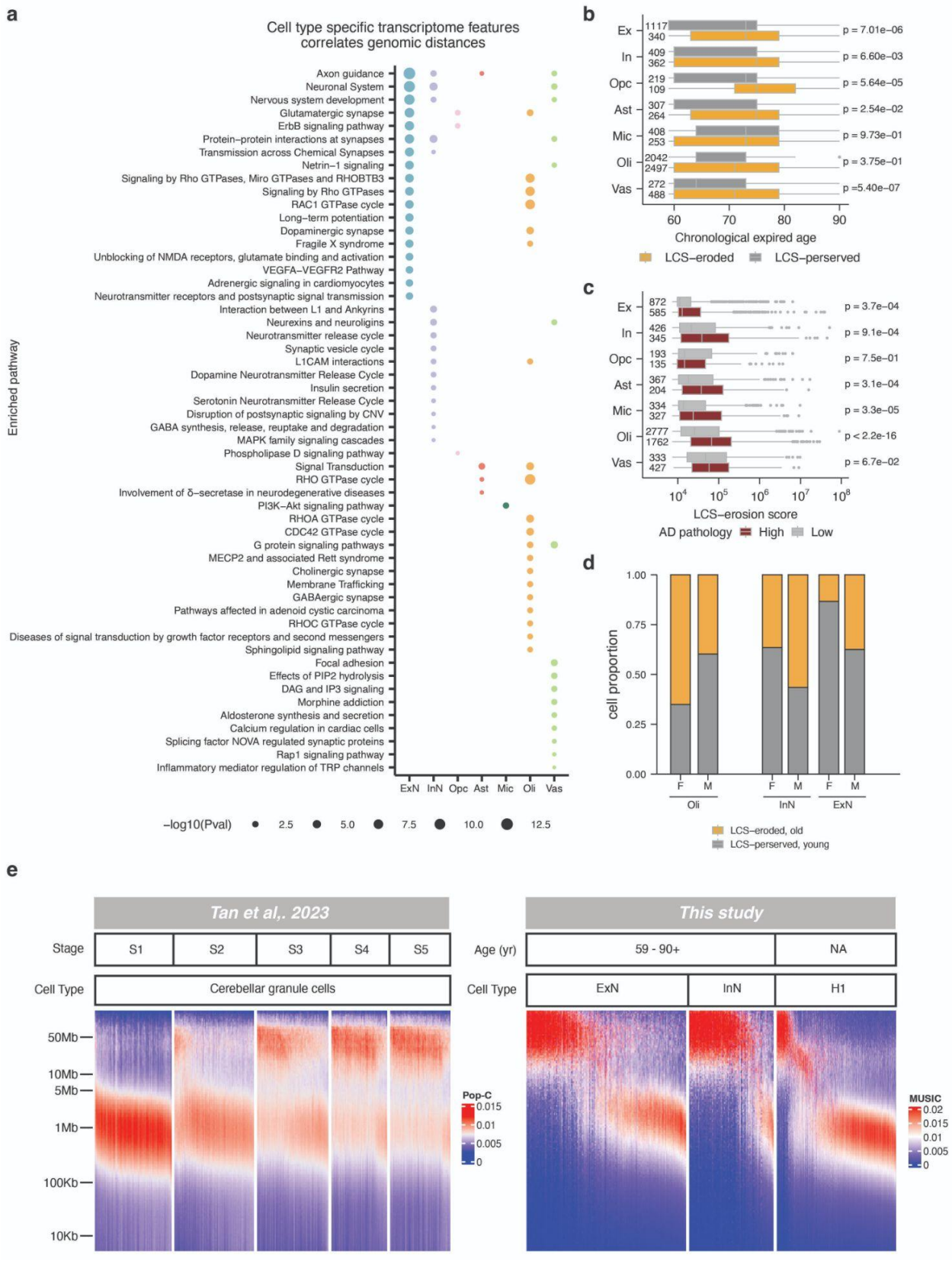
Supplementary Figure 3-6: Comparison of technologies.
(a) The number of RNA reads per cell (x axis) vs. the number of detected genes per cell (y axis) of CITE-seq, SNARE-seq, PairedTag, snRNA-seq, SNARE-seq and MUSIC frontal cortex data (MUSIC FC). The central dot represents the average number and the bars indicate the first and the third quartile of the distribution of all the cells (x axis) and per cell total detected genes (y axis). (b-c) The number of cells (x axis) vs. the median number of DNA-DNA contacts per cell (y axis) for each technique. For MUSIC and scSPRITE, the number of pairwise DNA-DNA contacts are decomposed from multiplex clusters. (d) Downsampling of the sequencing reads of the H1 library. Excitatory neuron (ExN) is used as an example cortical cell type to compare with the downsampled H1 data. The number of chromatin contacts per cell (y axis) is plotted against the number of DNA reads per cell (x axis). The mean (dot) and 95% range of the distribution (whiskers) of the original H1 (red), the downsampled H1 data (gray), and ExN (blue). (e) Summary of MUSIC data by cell type. (I) The proportions of singleton (gray) and non-singleton clusters (red). (II) Cluster sizes of the non-singleton clusters. (III-VI) The distributions of the numbers of DNA reads, DNA-DNA contacts, RNA reads, and detected genes in the single cells. Sample sizes for H1, Ex, In, Ast, Mic, Oli, Opc, Vas are 2267, 1457, 771, 571, 661, 4539, 328 and 760 respectively. For the middle boxplot, the left, center and right edge represent the 25th percentile, median and 75th percentile, respectively. Whiskers extend to 1.5 times the Interquartile Range from the box edges. Data points beyond the whiskers are outliers.


a
Single cell transcriptome
Genes / Cell
RNA Reads / Cell
Method
Sample
CITEseq
Human PBMC
PairTag
Mouse brain
MUSIC FC
Human frontal cortex
SNAREseq
Mouse brain
snRNAseq
Human prefrontal cortex
b
Median DNA–DNA contacts / Cell
Number of cells reported
MUSIC FC
Dip–C
snHi–C
sn–m3C–seq
scHi–C_2
Methyl–Hi–C
scHi–C_1
sci–Hi–C
c
Median DNA–DNA contacts / Cell
Number of cells reported
scSPRITE
MUSIC H1
d
Number of DNA contacts / Cell
DNA reads / Cell
ExN
MUSIC H1
MUSIC H1 downsample
e
(I)
(II)
(III)
(IV)
(V)
(VI)
cell type
H1
Ex
In
Ast
Mic
Oli
Opc
Vas
Singleton (%)
Non singleton cluster size (%)
DNA reads
DNA–DNA contacts
RNA reads
Genes
non–single
singleton
(1,2]
(2,10]
(10,100]
(100,Inf]

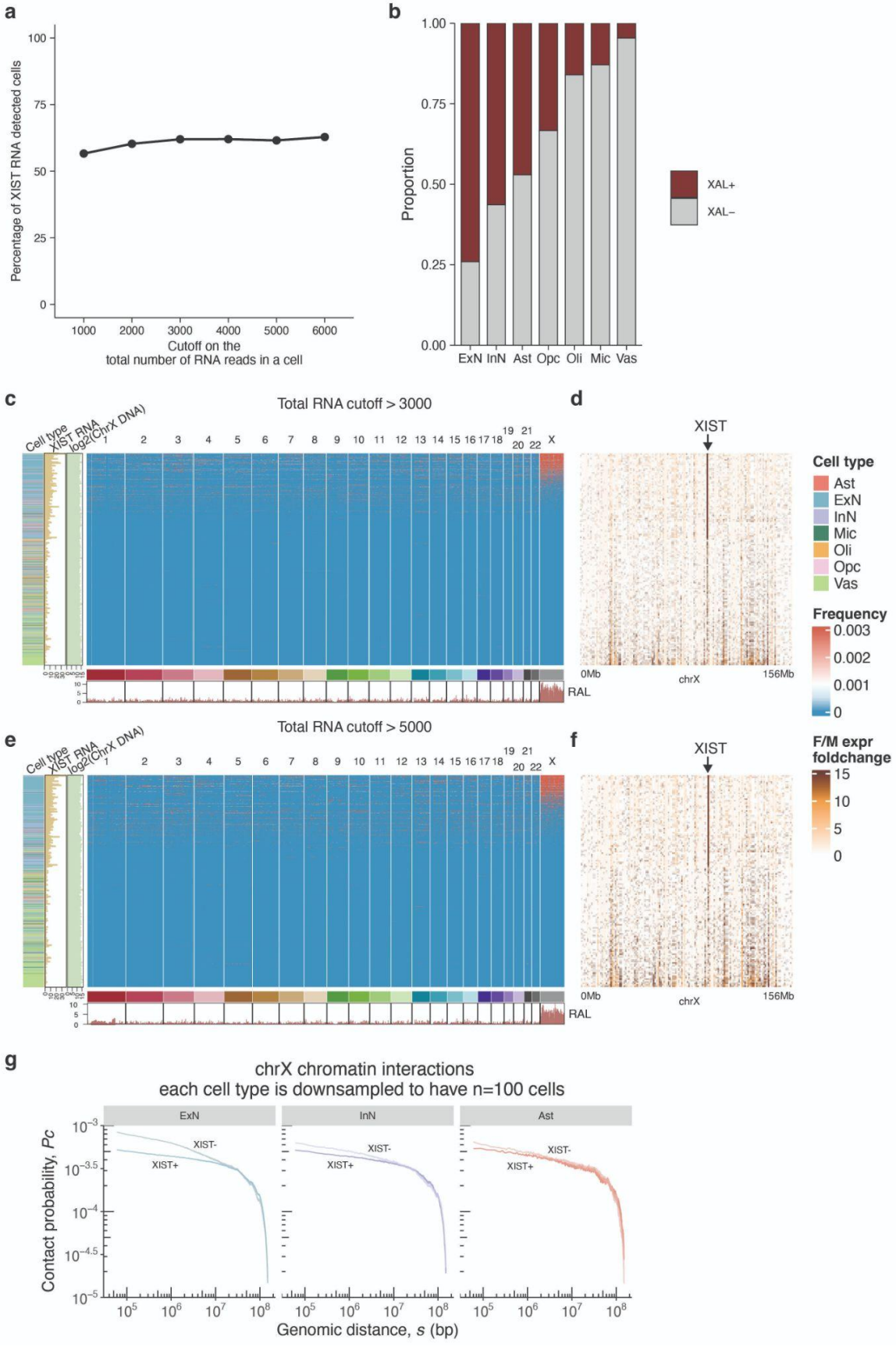
Supplementary Figure 3-7: MUSIC analysis of human frontal cortex.
Marker gene expression (a), number of cells (b), percent of female (pink) and male cells (blue) (c), and the relative proportion of each sample (d) in every single-cell cluster, and the assigned cell type and subtype for each cluster. Assigned subtypes and marker gene expression in excitatory neurons (g-i), inhibitory neurons (j-l), microglia (m-o), and oligodendrocytes (p-r). (e) Female (pink) and male (blue) cells in the UMAP embedding. (f) Proportions of single cells (y axis) with their most frequent chromatin interactions in each genomic bin (0-1 Mb, 1Mb-150Mb) (x axis) in each cell type.


a
SLC17A7
CAMK2A
NRGN
GAD1
GAD2
MBP
PLP1
MOBP
AQP4
GFAP
PDGFRA
VCAN
CSPG4
CD74
CSF1R
C3
FLT1
EMCN
VWF
Marker
ExN
InN
Oli
Ast
Opc
Mic
Vas
Expression
b
Cell num
760 86 132 224 4097 573 88 328 571 205 232 128 206 597 403 142 263 52
c
Sex
d
Sample
Vas OLI4 OLI3 OLI2 OLI1 MS4A- MS4A+ Opc Ast Sst Pvalb Lamp5 Vip L2/3_IT L5_IT L6_CT L6_IT L5/6_NP
Oli Mic InN ExN
Percent
0.25
0.5
0.75
1
e
F
M
UMAP2
UMAP1
f
Proportion
1.00
0.75
0.50
0.25
0.00
ExN Mic Opc Ast InN Oli Vas
Genomic Distance
0–1Mb
1Mb–150Mb
g
h
L2/3_IT
L6_IT
L5_IT
L6_CT
L5/6_NP
i
Clusters
L2/3_IT
L5_IT
L5/6_NP
L6_CT
L6_IT
7 5 1 0 6 3 4 2
j
k
Lamp5
Vip
Sst
Pvalb
l
Clusters
ADAMTS17
DPP10
SOX6
NXPH1
KIT
FBXL7
GALNTL6
GRM7
Lamp5
Pvalb
Sst
Vip
1 4 5 2 3 0
m
n
MS4A–
MS4A+
o
Clusters
MS4A7
MS4A14
MS4A6E
C3
MS4A–
MS4A+
1 0
p
q
Oli2
Oli1
Oli4
Oli3
r
Clusters
MOBP
SNORD105
SNORA48
SNORD60
Oli1
Oli2
Oli3
Oli4
3 2 1 0

Supplementary Figure 3-8: Correlation between Pc(s) and transcriptomic age among single cells. (a) The enriched pathways (row) in those genes that exhibit a correlated expression level with the LCS-erosion score across all the single cells in each cell type (column). The size of the dot represents the enrichment score (-log10(P value)) with larger dots indicating higher significance. (b) Distribution of the chronological ages (age at death, y axis) in the LCS-preserved (gray) and LCS-eroded cells (orange) in each cell type (column). Wilcoxon test, One-sided. Left numbers: sample size. Boxplot: the right edge, center line and left edge represent the 75th percentile, median and 25th percentile, respectively. Whiskers extend to 1.5 times the Interquartile Range from the box edges. Data points beyond the whiskers are outliers. (c) LCS-erosion scores (y axis) in the cells within the samples with high (Braak stage >=4) and low (Braak stage <3) AD pathology. Wilcoxon test, One-sided. Left numbers: sample size. Boxplot definition is the same as (e). (d) Proportion of LCS-eroded cells and LCS-preserved cells by gender (female (F), male (M)) within oligodendrocytes (Oli) and Neurons (InN, ExN). (e) Number of contacts (color) plotted against genomic distance (y-axis) in single nuclei (columns) using data from Tan et al. (left) and our study (right). Rows represent genomic bins with exponentially increasing sizes. The two datasets were plotted using the same computational program and genomic bin sizes. The 'Stage' labels correspond to cell groups identified by Tan et al., where Stages S1 to S5 are enriched with cells of chronological ages 0.2, 1, 10, 30, and 80, respectively. ExN: Frontal cortex excitatory neurons. InN: Frontal cortex inhibitory neurons.


a
Cell type specific transcriptome features correlates genomic distances
Enriched pathway
Axon guidance
Neuronal System
Nervous system development
Glutamatergic synapse
ErbB signaling pathway
Protein–protein interactions at synapses
Transmission across Chemical Synapses
Netrin–1 signaling
Signaling by Rho GTPases, Miro GTPases and RHOBTB3
Signaling by Rho GTPases
RAC1 GTPase cycle
Long–term potentiation
Dopaminergic synapse
Fragile X syndrome
Unblocking of NMDA receptors, glutamate binding and activation
VEGFA–VEGFR2 Pathway
Adrenergic signaling in cardiomyocytes
Neurotransmitter receptors and postsynaptic signal transmission
Interaction between L1 and Ankyrins
Neurexins and neuroligins
Neurotransmitter release cycle
Synaptic vesicle cycle
L1CAM interactions
Dopamine Neurotransmitter Release Cycle
Insulin secretion
Serotonin Neurotransmitter Release Cycle
Disruption of postsynaptic signaling by CNV
GABA synthesis, release, reuptake and degradation
MAPK family signaling cascades
Phospholipase D signaling pathway
Signal Transduction
RHO GTPase cycle
Involvement of δ–secretase in neurodegenerative diseases
PI3K–Akt signaling pathway
RHOA GTPase cycle
CDC42 GTPase cycle
G protein signaling pathways
MECP2 and associated Rett syndrome
Cholinergic synapse
Membrane Trafficking
GABAergic synapse
Pathways affected in adenoid cystic carcinoma
RHOC GTPase cycle
Diseases of signal transduction by growth factor receptors and second messengers
Sphingolipid signaling pathway
Focal adhesion
Effects of PIP2 hydrolysis
DAG and IP3 signaling
Morphine addiction
Aldosterone synthesis and secretion
Calcium regulation in cardiac cells
Splicing factor NOVA regulated synaptic proteins
Rap1 signaling pathway
Inflammatory mediator regulation of TRP channels
ExN InN Opc Ast Mic Oli Vas
−log10(Pval) 2.5 5.0 7.5 10.0 12.5
b
Ex 1117 340 p = 7.01e−06
In 409 362 p = 6.60e−03
Opc 219 109 p = 5.64e−05
Ast 307 264 p = 2.54e−02
Mic 408 253 p = 9.73e−01
Oli 2042 2497 p = 3.75e−01
Vas 272 488 p =5.40e−07
60 70 80 90
Chronological expired age
LCS−eroded LCS−perserved
c
Ex 872 585 p = 3.7e−04
In 426 345 p = 9.1e−04
Opc 193 135 p = 7.5e−01
Ast 367 204 p = 3.1e−04
Mic 334 327 p = 3.3e−05
Oli 2777 1762 p < 2.2e-16
Vas 333 427 p = 6.7e−02
10^4 10^5 10^6 10^7 10^8
LCS−erosion score
AD pathology High Low
d
cell proportion
1.00 0.75 0.50 0.25 0.00
F M F M F M
Oli InN ExN
LCS−eroded, old
LCS−perserved, young
e
Tan et al,. 2023
This study
Stage S1 S2 S3 S4 S5
Age (yr) 59 - 90+ NA
Cell Type Cerebellar granule cells
Cell Type ExN InN H1
50Mb 10Mb 5Mb 1Mb 100Kb 10Kb
Pop-C 0.015 0.01 0.005 0
MUSIC 0.02 0.015 0.01 0.005 0

Supplementary Figure 3-9: XIST-chromatin association in single female cells.
(a) The change of the percentage of XIST lncRNA detected cells (y axis) in the female cortical cells that satisfy the threshold on the total number of RNA reads in a cell (x axis). (b) The variation of the proportions of the observed XAL+ cells across cell types (columns) in the female cortex under total RNA larger than 5000 cutoff. (c-f) The female cells are filtered so that every cell has at least 3000 RNA reads (c-d) or at least 5000 RNA reads (e-f). (c, e) Genome-wide distribution of XIST-chromatin association in individual female cortical cells. Each row represents a single cell. The RNA attachment level (RAL) of the XIST lncRNA (intensity of red color) at any genomic region on any chromosome is plotted with the corresponding genomic coordinates (x axis). Resolution: 1 Mb. Track at the bottom: cumulative RAL of XIST. Tracks on the left indicate the cell type (color), XIST RNA read count (XIST RNA), X chromosomal DNA read counts (chr.X DNA) in log2 scale. (d, f) sex-fold-change of a previously identified gene with incomplete X chromosome inactivation (XCI) (column) between this female cell (row) and the average expression of the male cells of the matched cell type. Only the genes expressed in at least 20 cells are plotted. (g) Downsampling analysis of cell numbers. X chromosome Pc(s) curves in excitatory neurons (ExN), inhibitory neurons (InN), and astrocytes (Ast) are plotted from the same number of female cells (n = 100). The difference between XIST+ (darker color) and XIST- Pc(s) (lighter color) curves is most pronounced in ExN.


a
Percentage of XIST RNA detected cells
100
75
50
25
0
1000
2000
3000
4000
5000
6000
Cutoff on the
total number of RNA reads in a cell
b
Proportion
1.00
0.75
0.50
0.25
0.00
ExN InN Ast Opc Oli Mic Vas
XAL+
XAL−
c
Total RNA cutoff > 3000
Cell type
XIST RNA
log2(ChrX DNA)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
RAL
d
XIST
0Mb
chrX
156Mb
Cell type
Ast
ExN
InN
Mic
Oli
Opc
Vas
Frequency
0.003
0.002
0.001
0
e
Total RNA cutoff > 5000
Cell type
XIST RNA
log2(ChrX DNA)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
RAL
f
XIST
0Mb
chrX
156Mb
F/M expr
foldchange
15
10
5
0
g
chrX chromatin interactions
each cell type is downsampled to have n=100 cells
ExN
InN
Ast
XIST-
XIST+
Contact probability, Pc
10^{-3}
$10^{-3.5}$
10^{-4}
$10^{-4.5}$
10^{-5}
10^{5} 10^{6} 10^{7} 10^{8}
Genomic distance, s (bp)

Supplemental Note 1

Single-cell multi-omic technologies have revolutionized our ability to collectively analyze epigenomic or chromatin conformation features and gene expression (Chen et al., 2019; Z. Liu et al., 2023; Ma et al., 2020; Plongthongkum et al., 2021; Zhu et al., 2020). Various methodologies have been introduced to enable the joint profiling of specific aspects of the genome, transcriptome, or proteome. For instance, SNARE-seq (Chen et al., 2019) and SHARE-seq (Ma et al., 2020) allow for simultaneously profiling chromatin-accessible regions (CARs) and gene expression (Supplementary Table 1). Additionally, scNMT-seq (Clark et al., 2018) facilitates the combined analysis of CARs and DNA methylation, while scNOMeRe-seq (Yang Wang et al., 2021) enables the joint examination of CARs, DNA methylation, and gene expression. Techniques such as Pair-tag provide insights into histone modification profiles and transcription levels (Zhu et al., 2021), and sn-m3C-seq captures chromatin interactions and DNA methylation patterns (D.-S. Lee et al., 2019). CITE-seq, on the other hand, allows for the simultaneous measurement of surface protein and transcriptome (Stoeckius et al., 2017). Recently, HiRES technology has been introduced to the concurrent application of Hi-C and RNA-seq within individual cells (Z. Liu et al., 2023) (Supplementary Table 1). Furthermore, imaging tools like DNA seqFISH+ (Takei et al., 2021a), ORCA (Mateo et al., 2019), and MINA (M. Liu et al., 2020) have facilitated the joint analysis of chromatin traces and gene expression (Supplementary Table 1).

Supplemental Note 2

The MUSIC workflow contains two major steps.

Step 1: Adding Cell Barcodes

The first step ligates *the RNA Linker* to the RNA molecules and *the DNA Linker* to the fragmented DNA and adds the Cell Barcodes (Supplementary Figure 3-1a-e). This step accomplishes the first and the second design goals. The RNA Linker is a chimeric oligonucleotide

consisting of single-stranded DNA (ssDNA) at its 5' end and single-stranded RNA (ssRNA) at its 3' end (Supplementary Figure 3-2e). The ssRNA and ssDNA are designed to facilitate efficient ligations with RNA and the Cell Barcodes, respectively. The ligation of the RNA Linker with the RNA produces a chimeric product "RNA-Linker+RNA" with ssDNA on this 5' end (Supplementary Figure 3-1b).

The DNA Linker is partially double-stranded DNA (dsDNA) and partially single-stranded DNA (ssDNA). It is generated by hybridizing two DNA strands (Supplementary Figure 3-2g, Supplementary Table 2). The DNA Linker is designed to (1) efficiently ligate with fragmented chromosomal DNA through the dsDNA region via sticky-end ligation, (2) effectively ligate with the Cell Barcodes via the ssDNA region on the bottom strand, and (3) facilitate hybridization with the 10X Barcodes through the ssDNA region on the top strand. Ligation of the DNA Linker with fragmented DNA generates the DNA-Linker+DNA+DNA-Linker structure.

Subsequently, both RNA-Linker+RNA and DNA-Linker+DNA+DNA-Linker undergo ligation with the Cell Barcodes, which label the RNA and DNA originating from the same nuclei. An innovation in the MUSIC workflow is the ability to employ the same procedure for adding Cell Barcodes to both RNA and DNA. The ssDNA region on the 5' end of the RNA Linker (RNA-Linker_ssDNA) and the ssDNA region on the bottom strand of the DNA Linker (DNA-Linker_bt-ssDNA) share the same optimized DNA sequence, enabling efficient ligation with the Cell Barcodes.

The Cell Barcodes consist of three sets: the 1st, 2nd, and 3rd sets of Cell Barcodes (Supplementary Figure 3-2d, i). Each set comprises a top-strand overhang, a dsDNA region, and a bottom-strand overhang. The top-strand overhang and bottom-strand overhang are identical for all barcodes within a set, while the dsDNA region is unique for each barcode. The addition of the

three sets of barcodes is accomplished through three rounds of split-pooling of the nuclei (Arrastia et al., 2022) (Supplementary Figure 3-1d-e).

The three sets of Cell Barcodes are designed to maximize the ligation efficiency for sequentially ligating the 1st set of Cell Barcodes with the RNA Linker and the DNA Linker, the 2nd set with the 1st set, and the 3rd set with the 2nd set. The optimal ligation efficiency is achieved by the complementarity of the 1st set's overhang with RNA-Linker_ssDNA and DNA-Linker_bt-ssDNA and the complementarity of the overhang sequences of the sequentially ordered barcode sets (Supplementary Figure 3-1c, Supplementary Figure 3-2f, h). Out-of-order ligations, such as the ligation of the 2nd set of Cell Barcodes with the RNA Link are minimized due to lack of sequence complementarity in the overhangs by design (Bilotti et al., 2022; Potapov et al., 2018).

Step 2: Adding Complex Barcodes

The second step of the MUSIC workflow involves pooling the nuclei and releasing the molecular complexes from the nuclei. At this stage, the molecular complexes can be traced back to individual nuclei using their Cell Barcodes. The next objective is to assign the same Complex Barcode to any RNA or DNA within the same molecular complex. The combination of Cell Barcodes and Complex Barcodes specifies each molecular complex in a particular cell. It should be noted that the Complex Barcodes alone, without the Cell Barcodes, are not expected to be sufficient for identifying every molecular complex.

The Complex Barcode comprises two sets of barcodes: the 10X Barcodes and the I7 Barcodes that are embedded in the Index Adaptors, a.k.a. I7 Barcodes. We chose to use the combination of 10X Barcodes and I7 Barcodes rather than the 10X Barcodes alone as the Complex Barcodes because the combination offers an X-fold increase in the number of barcodes than the 10X Barcodes alone, where X is the number of distinct I7 Barcode sequences.

The 10X Genomics' Gel Bead-in-Emulsion (GEM) system is employed to add the 10X Barcodes. This system encapsulates molecular complexes into individual GEMs, also called droplets, where the complexes in each droplet are added with a unique 10X Barcode. Importantly, the MUSIC technology by design does not require most droplets to encapsulate no more than one molecular complex. As we will explain later, a single molecular complex is identified by the combination of Cell Barcodes and Complex Barcodes. This droplet-based addition of 10X Barcodes can be considered another round of split-pooling, where the units are molecular complexes, unlike the previous rounds where the units were nuclei.

The addition of 10X Barcodes is required for both RNA and DNA. The 10X system is optimized to add 10X Barcodes to RNA. The RNA, following polyadenylation, can be hybridized by the polyT sequence within the 10X Barcodes (Supplementary Figure 3-1f), initiating subsequent automated cDNA synthesis within each GEM. In the case of DNA, MUSIC leverages the design of the DNA-Linker to add 10X Barcodes. The DNA-Linker contains a 30 nt single-stranded polyA sequence (Supplementary Figure 3-2g), which can hybridize with the 30 nt PolyT sequence in the 10X Barcodes (Supplementary Figure 3-2k). This allows subsequent DNA synthesis using the fragmented DNA as a template.

The 10X GEM system comprises approximately 3.5 million unique 10X Barcodes. This number exceeds the expected number of molecules captured per cell by a single-cell assay. Consequently, there is no significant concern regarding the sufficiency of 10X Barcodes to differentiate molecular complexes that can be captured and assayed from a single cell. However, including I7 Barcodes in combination with the 10X Barcodes ensures the sufficient diversity of Complex Barcodes.

The I7 Barcode, a part of the Index Adaptor, is added during the PCR process that constructs the final sequencing library. The output of the 10X GEM system comprises two components: cDNA originating from RNA and double-stranded DNA (dsDNA) originating from chromosomal DNA. The cDNA takes the form of Cell Barcode+RNA-Linker+RNA Insert+10X Barcode, while the dsDNA takes the form of Cell Barcode+DNA-Linker(bt)+DNA Insert+DNA-Linker(top)+10X Barcode. DNA-Linker(top) and DNA-Linker(bt) denote the dsDNA originating from the top and the bottom strands of the DNA Linker. Both the cDNA and dsDNA contain the 3rd set of the Cell Barcode at one end, which includes a fraction of the Illumina Read2 sequence, and the 10X Barcode at the other end, including a fraction of the Illumina Read1 sequence. These components can be constructed into the Illumina sequencing library using PCR. The PCR primers, namely the Index Adaptor and the Universal Adaptor, common to both cDNA and dsDNA, are used to amplify them into the final sequencing library.

The addition of the different I7 Barcodes is achieved through another round of split-pooling. The output of the 10X GEM system is split into eight tubes. Each tube is added with an Index Adaptor that contains a specific I7 Barcode, for PCR-based library construction. The constructed sequencing libraries are combined to form the final sequencing library. In this round of split-pooling, the units are molecular complexes. To summarize, MUSIC employs two rounds of split-pooling to add the Complex Barcodes. The first round involves splitting by 10X GEMs, while the second round involves splitting by tubes. In the current implementation (MUSIC v1.0), the estimated number of possible barcode combinations is up to 25 trillion, approximated by multiplying the 884,736 distinct Cell Barcodes with 3.5 million available 10X Barcodes, and further multiplying this total by the 8 different I7 Barcodes. The 10X microfluidic system samples a random subset of the available 10X Barcodes.

Supplemental Note 3

We used the mix-species dataset to evaluate MUSIC's resolution at single-cell and single-complex levels. The mixed-species dataset resolved 372,878,969 DNA reads and 22,714,916 RNA reads for humans (hg38), 128,687,584 DNA reads and 8,951,899 RNA reads for mice (mm10). All the subsequent analyses are based on these UMNDBC reads.

First, MUSIC identifies the DNA reads and RNA reads that share the same Cell Barcodes (CB) as coming from the same cell. Based on an established approach and threshold (Arrastia et al., 2022), MUSIC resolved 2,546 human cells, 1,381 mouse cells, and 36 cells with mixed species content. This resulted in a species-mixing rate of 0.91% at the cell level (Supplementary Figure 3-3a). Comparatively, this species-mixing rate is within the same order of magnitude as the species-mixing rate of 3% reported for scSPRITE under the same thresholds (Arrastia et al., 2022). Next, MUSIC identifies the reads sharing the Cell Barcodes and the Complex Barcodes (CB+10X+I7) as a cluster. In the mixed-species dataset, 154,499,469 clusters were resolved, of which 154,291,286 clusters contained reads from a single species. This corresponds to a complex-level species-mixing rate of 0.13% (Supplementary Figure 3-3b). Taken together, the low species-mixing rates at the cell level and the complex level support the ability of MUSIC to generate data at single-cell and single-complex resolutions.

We analyzed the human H1 cells regarding the numbers of reads, clusters, and pairwise contacts. First, the data solved 2,546 human H1 cells. Each H1 cell contained an average of 144,049 UMNDBC DNA reads (blue dots) and 11,384 UMNDBC RNA reads (orange dots) (Figure 3.1b). Second, clusters were defined as collections of reads sharing the same Cell Barcode and Complex Barcode (CB+10X+I7). In the perfect scenario where cluster-level species-mixing rate = 0, a cluster reflects a molecular complex. A cluster is called a DNA-only, an RNA-only, or an RNA-DNA cluster if it contains only DNA reads, only RNA reads, or both DNA and RNA

reads. On average, each H1 cell had 7,036 DNA-only clusters (DD clusters), 232 RNA-only clusters (RR clusters), and 1,170 RNA-DNA clusters (RD clusters) (Figure 3.1c). Finally, co-complexed multi-way associations were projected to co-complexed read pairs based on a previously described procedure (Arrastia et al., 2022), resulting in an average of 2,639,302,084 co-complexed DNA-DNA pairs, 7,089,720 co-complexed RNA-RNA pairs, and 250,525,581 co-complexed RNA-DNA pairs per cell (Figure 3.1d).

Supplemental Note 4

We carried out a downsampling analysis to test whether the different read numbers in pairwise and multiplex interactions can explain the better representation of TAD structure in the multiplex interactions. We downsampled the total number of 22,602 UMNDBC DNA reads in this region (Chr12:114-117Mb) (complete data) to 15k, 10k, and 5k DNA reads (downsampled datasets). At the extreme end of the downsampled datasets (5k reads, pairwise and multiplex combined), the TAD structure becomes indiscernible, and accordingly there is no discernable TAD structure in either the pairwise or the multiplex clusters (Supplementary Figure 3-5b). In the other downsampled datasets (15k, 10k), the TAD structure remained discernable in the total DD contacts (pairwise and multiplex combined). Similar to what's observed in the complete data (without downsampling), pairwise clusters alone poorly reflected the TAD structure, while multiplex clusters recapitulated the TAD structure (Supplementary Figure 3-5b). Notably, the numbers of DNA reads in multiplex clusters in these downsampled datasets (7,717 multiplex DNA reads when downsampled to 15k; 3,782 multiplex DNA reads when downsampled to 10k) are smaller than the number of DNA reads in pairwise clusters in the complete data (8,092 pairwise DNA reads), and in contrast, the TADs structures are discernable the multiplex clusters in the downsampled datasets but not in the pairwise clusters in the complete data, suggesting that the stronger representation of TADs in the multiplex clusters is not due to a larger number of DNA reads in the multiplex clusters.

Supplemental Note 5

Data from bulk assays revealed that the probability of chromatin interactions (Pc) decreases as the genomic distance (s) between the two DNA fragments increases (Lieberman-Aiden et al., 2009; Mirny, 2011). Furthermore, Pc(s) almost linearly correlates with s at the log-log scale (Lieberman-Aiden et al., 2009; Mirny, 2011) (Figure 3.2e). In MUSIC data, the Pc(s) curve derived from size-2 DD clusters exhibited a linear relationship at the log-log scale (size=2 curve, Figure 3.2e), consistent with proximity-ligation-based assays. The DD clusters of size 3 or greater also exhibit decreasing Pc(s) curves, which are no longer linear at the log-log scale. The larger the cluster size, the greater Pc(s) deviates from a straight line (Figure 3.2e). The multiplex interactions displayed higher contact frequencies at submegabase to several megabase genomic distances as compared to the pairwise contacts, suggesting an enrichment of long-range chromatin interactions in the multiplex complexes.

Supplemental Note 6

Previous bulk assays studies have identified two key features of genome-wide RNA-chromatin associations. First, genomic regions with high levels of locally transcribed premature mRNA (pre-mRNA) (pre-mRNA-rich regions) tend to correspond to the A compartments defined by chromatin conformation assays such as Hi-C and Micro-C (Chen et al., 2018; Quinodoz et al., 2021). Second, these pre-mRNA-rich regions also serve as hotspots for long-range RNA-chromatin interactions involving non-coding RNAs MALAT1, 7SK, and snRNAs (collectively called nsaRNAs) (Chen et al., 2018). In the MUSIC-identified RNA-DNA contacts (ensemble RD), the pre-mRNA associated genomic regions correlated with Micro-C derived A compartment, and more specifically with the “Speckle” nuclear compartmentalization jointly inferred from TSA-seq, DamID, and Hi-C data (Yuchuan Wang et al., 2021) (Figure 3.2l). To categorize an RNA read as pre-mRNA, we required it to span an intron-exon junction and contain at least 15 intronic

nucleotides. Furthermore, the genomic regions targeted by long-range nsaRNA-chromatin associations (blue curve) overlapped with regions associated with high levels of locally transcribed pre-mRNA (p-value < 2.2e-16, red curve, Figure 3.2l, m). Thus, MUSIC recapitulated both prominent features of genome-wide RNA-chromatin associations. In addition, the MUSIC data reproduced the allele-specific association of the KCNQ1OT1 noncoding RNA with the genomic regions adjacent to the KCNQ1OT1 gene (Pandey et al., 2008) (Supplementary Figure 3-5e), affirming MUSIC's ability to recapitulate the known chromatin association pattern of a specific RNA.

Supplemental Note 7

We compared the number of RNA reads per cell between MUSIC and other technologies. We compared the RNA reads in MUSIC FC to single-nucleus RNA-seq data from human prefrontal cortex (snRNA-seq PFC) (Mathys et al., 2019) and multi-omic datasets including CITE-seq data generated from human PBMC (Stoeckius et al., 2017), SNARE-seq data generated from mouse brain (Chen et al., 2019), and PairTag data generated from mouse brain (Zhu et al., 2020) (Supplementary Table 5). snRNA-seq PFC resolved a median of 1,973 RNA reads per nucleus, corresponding to 1,348 genes. CITE-seq, SNARE-seq, and PairTag resolved a median of 1416, 1332 and 845 RNA reads per nucleus, corresponding to 770, 1203, and 626 genes (Supplementary Figure 3-6a). MUSIC FC exhibited a median of 1,136 RNA reads per nucleus, corresponding to 853 genes, which was within the range of the other methods (Supplementary Figure 3-6b). Thus, MUSIC demonstrated a comparable read depth to snRNA-seq and other multi-omic methods in mapping single-nucleus transcriptomes.

We also compared the numbers of resolved nuclei and DNA-DNA contacts. We compared MUSIC to published single-cell multi-omic datasets of scHiC_1(Nagano et al., 2013), sci-Hi-C(Ramani et al., 2017), scHi-C_2(Nagano et al., 2017), snHi-C(Flyamer et al., 2017), Dip-C(Tan

et al., 2018), scSPRITE(Arrastia et al., 2022) and multi-omics methods Methyl-Hi-C(G. Li et al., 2019), sn-m3C-seq(D.-S. Lee et al., 2019). Most of these datasets contained up to 2,000 nuclei, while the sn-m3C-seq dataset had 6,500 nuclei (Supplementary Figure 3-6b, c). We reused the numbers of median DNA-DNA contacts per cell of each technology from a recent review(Zhou et al., 2021). MUSIC FC had 9,087 nuclei and a median of 2,651,577 DNA-DNA contacts, surpassing most other datasets (Supplementary Figure 3-6b, c). Moreover, scSPRITE data generated from mouse embryonic stem cell (mESC) (Arrastia et al., 2022) resolved marginally fewer nuclei and more DNA-DNA contacts, both within 2-fold of difference as compared to MUSIC H1 data (Supplementary Figure 3-6c).

We compared the usable fractions of reads between MUSIC and scSPRITE. The scSRIPTE mESC dataset includes a total number of 1,269,693,929 raw sequencing reads, of which 632,242,064 (49.8%) contains all the six barcodes (Arrastia et al., 2022). scSPRITE has 50.2% loss due to incomplete cell barcodes, which is greater than MUSIC's total loss of Cell Barcodes and Complex Barcodes (35.94%). After read trimming, mapping, and removing PCR duplicates, the final yield of scSPRITE is 161,989,473 reads, which account for 12.75% of the raw sequencing reads. In comparison, MUSIC's final yield (UMNDBC reads) account for 17.39% of the raw sequencing reads. Thus, the usable fraction of the total sequencing reads in MUSIC is comparable, if not higher, than in scSPRITE. Taken together, these data suggest comparable throughputs between MUSIC and the other technologies. Of note, previous work projected scSRITE's multiplex interactions to pairwise contacts and reported the projected pairwise contacts as the number of DNA-DNA contacts of scSPRITE (Arrastia et al., 2022). We followed this approach to project MUSIC's multiplex data to pairwise contacts. The greater numbers reported here for scSPRITE and MUSIC are attributable to their ability to capture multi-way contacts, which does

not mean that they are more sensitive in revealing pairwise chromatin interactions than the other methods.

A downsampling analysis suggests that differences in sequencing depth partly account for the higher contact numbers observed in MUSIC H1 compared to MUSIC FC (Supplementary Figure 3-6d). Additionally, we acknowledge the potential influences of other biological or technical factors, such as the documented decrease in chromatin contacts during ES cell differentiation (Schlesinger & Meshorer, 2019)·(H. Liu et al., 2023), on these discrepancies.

Supplemental Note 8

We performed clustering analysis on the 9,087 cortical nuclei based on their RNA expression levels (Figure 3.3a). The clusters were then assigned to specific cell types using known marker genes from previous studies (Supplementary Table 6) (Franzén et al., 2019; Mathys et al., 2019). A cluster of 1,457 nuclei was categorized as excitatory neurons (ExN) based on the expression of marker genes SLC17A7, CAMK2A, and NRGN (Mathys et al., 2019) (Supplementary Figure 3-7a). Within this cluster, two subclusters were identified: intratelencephalic (IT) neurons (on the left) and corticothalamic (CT) and near-projecting (NP) neurons (on the right) (Supplementary Figure 3-7g-i). The IT subcluster further consisted of three subgroups: Layer 2/3 intratelencephalic (L2/3 IT) (597 nuclei), Layer 5 IT (403 nuclei), and Layer 6 IT (263 nuclei) neurons (Supplementary Figure 3-7g-i). The CT and NP subcluster contained Layer 6 corticothalamic (L6 CT) (142 nuclei) and Layer 5/6 near-projecting (L5/6 NP) (52 nuclei) neurons, respectively (Bakken et al., 2021) (Supplementary Figure 3-7g-i).

A cluster of 771 nuclei was categorized as inhibitory neurons (InN) based on the expression of marker genes GAD1 and GAD2 (Mathys et al., 2019) (Supplementary Figure 3-7a). This cluster formed four subclusters (Supplementary Figure 3-7j-l) corresponding to Lamp5, Vip, Pvalb, and Sst neurons. The two sub-clusters on the top corresponded to Lamp5 (128 nuclei) and Vip (206

nuclei) neurons that develop from caudal ganglionic eminence (Tasic et al., 2018). The two sub-clusters on the bottom corresponded to Pvalb (232 nuclei) and Sst (205 nuclei) neurons that develop from medial ganglionic eminence (Supplementary Figure 3-7j-l).

A cluster of 661 nuclei was categorized as microglia based on the expression of marker genes CD74, CSF1R, and C3 (Mathys et al., 2019) (Supplementary Figure 3-7a). This microglial cluster consisted of a major subcluster (573 nuclei) and a minor subcluster (88 nuclei) marked by different levels of Membrane-spanning 4A (MS4A) genes (Deming et al., 2019) (Supplementary Figure 3-7m-o). In mice, the activation of MS4A genes marks the transition of microglia from a chemokine state to an interferon state (Deming et al., 2019). Thus, the major and the minor sub-clusters may reflect the corresponding cellular states of microglia in humans.

A cluster of 4,539 nuclei was categorized as oligodendrocytes based on the expression of marker genes MBP, PLP1, and MOBP (Mathys et al., 2019), which included a major subcluster (4,097 nuclei) and three minor subclusters (224, 132, and 86 nuclei) (Supplementary Figure 3-7p-r). Another cluster of 760 nuclei was categorized as vascular cells based on the expression of marker genes EMCN, VWF, and FLT1 (Franzén et al., 2019) (Figure 3.3a). Lastly, two distinct clusters with 571 and 328 nuclei were categorized as astrocytes (Ast) based on marker genes AQP4 and GFAP, and oligodendrocyte precursors (Opc) based on marker genes VCAN, PDGFRA, and CSPG4 (Mathys et al., 2019). Additionally, a joint analysis of MUSIC FC data with a single-nucleus RNA-seq (snRNA-seq) dataset of human frontal cortex revealed highly consistent clustering structures and clustering-based cell type assignments between the two datasets (Supplemental Note 9).

Supplemental Note 9

We evaluated the consistency of single-cell clustering and cell type assignments between MUSIC FC and a snRNA-seq dataset of human frontal cortex (Mathys et al., 2019) (hereafter called Mathys' dataset). We integrated the two datasets using the Harmony software (Korsunsky et al., 2019) and carried out Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) analysis on the integrated dataset (Supplementary Fig. 2a-c). The single cells in the integrated dataset exhibited clear clustering structure, and the co-clustered single cells originated from MUSIC FC and Mathys' dataset tended to share the same marker gene expression profiles (Supplementary Fig. 2d-g), suggesting a high degree of consistency between MUSIC FC and Mathys' dataset.

MUSIC FC's oligodendrocyte cluster included three previously uncharacterized sub-clusters. These sub-clusters (Oli2, Oli, Oli4) are separated from the main oligodendrocyte cluster (Oli1) due to their cluster-specific expression of snoRNAs (Supplementary Figure 3-7r). snoRNAs are non-polyA RNAs that cannot be captured by the the snRNA-seq technique used by Mathys et al., whereas they are included in MUSIC FC because MUSIC captures both polyA and non-polyA RNA types (Supplementary Figure 3-5c). Accordingly, we expect these oligodendrocyte sub-clusters to be grouped in the main oligodendrocyte cluster in the integrated dataset, which indeed was the case (Supplementary Fig. 3d-g). Furthermore, 76.62%, 73.21%, 68.18% and 81.39% cells from MUSIC FC's Oli1 (the main cluster), Oli2, Oli3, and Oli4 were included in the oligodendrocyte cluster of the integrated dataset, revealing a high degree of consistency between Oli2, Oli3, Oli4 and the oligodendrocyte cluster of the integrated dataset.

Supplemental Note 10

We analyzed the contributions of transcriptomic age, chronological age, sex, AD pathology (AD or non-AD), and cell type to explain the variation of LCS erosion score. To ensure the

robustness of the results, we ran two linear models, namely a full model (Full Model) and a model selection identified model (Selected Model). The two models led to essentially the same results.

First, we specified the Full Model as:

$log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_{6-31}(X_1 * X_2 * X_3 * X_4 * X_5) + \varepsilon$, where Y denote the LCS erosion score, X_1 denotes Transcriptomic age, X_2 denotes Chronological Age, X_3 denotes AD pathology, X_4 denotes Sex, X_5 denotes Cell type, and $X_1 * X_2 * X_3 * X_4 * X_5$ denotes the collection of all the pairwise and higher order interaction terms. The covariates collectively explain 27.25% of the variance in $log(Y)$. Six terms exhibit the ability to explain 1.0% or more variance, which are:

Transcriptomic age, explaining 14.38% of the variance (p value < 2.2e-16),

Sex * Cell type, explaining 2.09% of the variance (p value < 2.2e-16),

Cell type, explaining 1.83% of the variance (p value < 2.2e-16),

Sex, explaining 1.67% of the variance (p value < 2.2e-16),

AD pathology, explaining 1.57% of the variance (p value < 2.2e-16),

Transcriptomic age * Cell type, explaining 1.10% of variance (p value < 2.2e-16),

None of the other covariates explains 1% of the total variance.

Next, we identified the Selected Model using stepwise model selection with the optimal Akaike information criterion (AIC), as:

$log(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_1 X_3 + \beta_9 X_1 X_5 + \beta_{15} X_4 X_5 + \varepsilon$,

i.e.

$$log(LCS\ erosion\ score) \sim \beta_1 Transcriptomic\ age + \beta_3 Pathology + \beta_4 Sex + \beta_5 Cell\ type + \beta_7\ Transcriptomic\ age * Pathology + \beta_9\ Transcriptomic\ age * Cell\ type + \beta_{15}\ Sex * Cell\ type$$

The Selected Model explains 23.34% of the variance in $log(Y)$, where Transcriptomic age, Sex*Cell Type, Cell type, AD pathology, Sex, Transcripomic age*Cell type, and Transcriptomic age*Pathology explains 14.38%, 2.65%, 2.23%, 1.70%, 1.15%, 0.72%, 0.52%, respectively. This Selected Model includes all the top six terms with the largest explanatory abilities in the Full Model, suggesting the two models prioritized the same explanatory variables. In summary, LCS erosion score correlates with transcriptomic age, cell type, AD pathology, sex and the interactions of sex and cell type, transcriptomic age and cell type, and transcriptomic age and pathology.

Supplemental Note 11

MUSIC did not detect XIST lncRNA in every female cell. There are two probable causes for this variation. Technically, MUSIC may not be sensitive enough to detect XIST lncRNA in every nucleus. Biologically, the expression level of XIST may be heterogeneous among the female cells. Supporting the coexistence of both causes, XIST lncRNA is exclusively detected in female cells, and as the threshold on the number of total RNA reads in a cell is increased, the proportion of XIST-detected cells among the remaining female cells remains relatively stable with a modestly increasing trend (Supplementary Figure 3-9a). This analysis suggests that the RNA read count in a cell is influenced by both the expression level of the molecule and the sensitivity of the technique itself. To mitigate the impact of false negatives in our analyses, we refined the dataset by filtering based on RNA read count, and furthermore, we refrained from drawing conclusions about individual cells.

Supplemental Note 12

We compared XIST-chrX association with chrX gene expression. Without a phased genome, MUSIC data cannot infer allelic expression and thus cannot directly pinpoint any gene expressed from the Xi chromosome (incomplete XCI gene). However, a survey conducted by the Genotype-Tissue Expression (GTEx) consortium, involving 5,500 transcriptomes from 449

individuals across 29 tissues, reported that incomplete XCI genes generally exhibit higher expression in females compared to males in the corresponding tissue, suggesting sex bias can be used as “a proxy for XCI status" (Tukiainen et al., 2017). The MUSIC data can infer sex-difference in gene expression at the single-cell level. To illustrate this point, for every female cell and every previously annotated incomplete XCI gene (Tukiainen et al., 2017), we calculated the fold change between the RNA read count of this gene in this female cell and the average RNA read counts of this gene in all the male cells of the matching cell type (sex-fold-change) (Supplementary Figure 3-9d, f). XAL- female cells exhibited greater sex-fold-changes in most of the incomplete XCI genes compared to XAL+ female cells, suggesting the loss of XIST-chrX association in single female cells correlates with larger sex-difference in the gene expression in the human cortex (Supplementary Figure 3-9d, f).

Supplemental Note 13

Aligned with our findings, seqFISH+ analysis in the female mouse cortex revealed that Xa has smaller 3D spatial distances than Xi when the genomic sequence is below ~10Mb, and vice versa when the genomic distance is above ~10Mb (Fig. S24g in (Takei et al., 2021b)). Moreover, the female mouse cortex demonstrated a clearer difference between Xa and Xi in ExN compared to InN and Ast (Fig. S24 in (Takei et al., 2021b)). Specifically, most single ExN cells exhibited a smaller intra-chromosomal spatial distance in Xi than Xa at genomic distances greater than ~10Mb. In contrast, a notable fraction of InN and Ast cells showed non-distinguishable differences in intra-chromosomal spatial distance between Xi and Xa at any genomic distance (Fig. S24g in (Takei et al., 2021b)).

Supplemental Note 14

We developed a data processing pipeline, called MUSIC-docker for passing the MUSIC sequencing data. We packaged the data processing snakemake pipeline (Köster and Rahmann,

2012) with the Docker platform to enable execution of this data processing pipeline in any operating system without the need for re-compilation of the source codes. MUSIC-docker takes in i7 index splitted pair-end fastq files and process them into two sequence files (.fastq), corresponding to the RNA inserts and the DNA inserts, based on the RNA Linker or the DNA-Linker(bt) sequence identified in each read pair. MUSIC-docker adds the Cell Barcodes, the 10X Barcode, and the I7 Barcode into the sequence name of each RNA or DNA sequence. MUSIC maps the RNA sequences and the DNA sequences to the genome to obtain two mapped files (.BAM), where every sequence is associated with its genomic coordinate and with its Cell Barcodes and Complex Barcodes (see Methods). A full documentation of the MUSIC-docker pipeline is available at: http://sysbiocomp.ucsd.edu/public/wenxingzhao/MUSIC_docker/intro.html.

Convert the raw sequencing output to FASTQ files

The sequencing output file (.bcl) is provided together with the eight I7 Barcode sequences to the *bcl2fastq* program to yield eight FASTQ files. Each FASTQ file contains read pairs with a 28bp Read1 and a 150bp Read2. The 28bp Read1 sequence contains the 10X Barcode, which is composed of a 16bp 10X GEM Barcode and a 12bp 10X UMI sequence. The 150bp Read2 sequence contains the 3rd, 2nd, 1st Cell Barcodes, the RNA Linker sequence or the DNA-Linker(bt) sequence, the RNA insert or the DNA insert.

Extracting Cell Barcodes

Demultiplexing is the step to extract the fragment identity information (FMI) and the fragment sequence information (insert) from paired-end FASTQ files and reformat them into two new FASTQ files for RNA and DNA reads individually where the read sequence will be the insert,

and the read name will be the fragment identity information. This step is applied individually to each I7 index separated library.

FMI includes three cell barcodes, 10X GEM barcodes and 10X UMI. The 3rd Cell Barcode (CB3) contains a 14-17bp variable-length unique sequence region, followed by a 7bp overhang sequence, while the 2nd (CB2) and the 1st (CB1) Cell Barcodes contain a 14bp unique sequence region, followed by 7bp overhangs. CB3 has varied length to introduce complexity of the base composition at the same position which will help improve the sequencer's base calling quality. The linker sequences followed by the Cell Barcodes in Read2 are either the RNA Linker or the DNA-Linker(bt). The RNA Linker contains a 7bp overhang-complementary sequence to CB1 and a 20bp sequence that is unique to the RNA Linker. We note that RNA Linker's 7bp overhang has been counted in CB1's overhang sequence. The DNA-Linker(bt) is the double-stranded DNA produced from the bottom strand of the DNA Linker. The DNA-Linker(bt) is 31bp in length, including a 7bp overhang sequence and a 24bp sequence that is unique to the DNA Linker. We note that the 7bp overhang sequence has been counted in CB1's overhang. Based on whether it is the RNA Linker or the DNA-Linker(bt) sequence, we separately output to two .fastq files.

Following the RNA Linker sequence is the RNA insert, which is output as the sequence of the .fastq file for the RNA. The output read name for this RNA sequence is in the format of [previous read name]|[CB3]-[CB2]-[CB1]-[10X GEM Barcode]-[I7 Barcode]#[10X UMI], where [I7 Barcode] is a choice of "lib1", …, "lib8". The following is an example output sequence.

@**MN00185**:343:000H532W5:1:11102:14470:1060|BC3_68-BC2_89-BC1_46-TGACGCGGTACACGTT-lib1#*GTAAG GGTGAAG*

AATGGCAGAATGTTTTGGTCAAAATTTTTTAGCCAGCACCCCAATTAAAAAAAAAAAAAAAAAA +

FAFFFF==AFFFFFAFAFFAAF6=AFFFFFFA=F=/FAF=FFF=A=FFFF=FFFFF=/FF=//6

Following the DNA-Linker(bt) sequence is the DNA insert, which is output as the sequence of the .fastq file for the DNA. The output read name for this DNA sequence is in the format of [previous read name]|[CB3]-[CB2]-[CB1]-[10X GEM Barcode]-[I7 Barcode]#[10X UMI], where "I7 Barcode" is a choice of "lib1", …, "lib8". The following is an example output sequence.

@**MN00185**:343:000H532W5:1:11102:15055:1047|BC3_75-BC2_51-BC1_11-CGGAACCAGGCATCGA-lib1#*GGTCT GTACCTG*

TCATAAAGAACTCTACCAACTTAAGAAGAAGAAAGCAAGCAACCCCATACAAAAATGGCGA +

=F/FA=/F=/F=FF/FF///6/A=AAFFF/A/=//F///FF/FFFFF/FF///FFF/F///

The Cell Barcodes are designed such that the Levenshtein distance between any two of the Cell Barcodes is greater than 2. For Cell Barcode or Linker sequences, we allow at most 2 mismatches (Hamming distance) to the reference sequence. The whitelist of the "10X Single Cell 3' v3.1 Chemistry" - is provided for matching the 10X Barcodes, where a match is defined as a perfect match. The demultiplexing step is implemented in python with customized code.

Reads cleaning

Based on our experimental design and sequencing errors from the Illumina machine, it is possible that we introduced some artificial sequences into the library even if the reads could have perfectly matched the three cell barcodes and 10X barcodes. For DNA reads, we will remove the consecutive As or consecutive Gs from the 3' of the DNA insert if there are more than 20 consecutive As or consecutive Gs. For RNA reads, we will detect the ssDNA region of the RNA Linker sequence "CGAGGAGCGCTT" and remove any sequence following it. We use *cutadaptor* (v2.8) with parameter `-q 15 -m 20` to ensure the remaining sequences have a base quality score larger than 15 and with length at least 20bp.

Mapping DNA and RNA ends

After raw reads demultiplexing and reads cleaning, we will map DNA and RNA inserts to the reference genome. We used *bowtie2* (v5.4.0) for DNA inserts genome mapping with the command "bowtie2 -p 10 -t --phred33 -x ". We used *bwa mem* to map RNA inserts to the reference genome in a splice aware manner with parameters "-SP5M". Only uniquely mapped reads were selected and will be used for downstream analysis.

Reads deduplication

Following read mapping, the subsequent crucial task involves the removal of PCR duplicates, which are identified by sharing the same Unique Molecular Identifier (UMI). In our study, we utilized the 10X UMI. Deduplication based on mapped coordinates is a widely employed approach, as alignment software typically accounts for sequencing errors present in the reads. To adapt to our specific requirements, we have developed a customized deduplication tool capable of accommodating the customized read name format. The deduplication process involves initially sorting the BAM file according to the mapping coordinates. Subsequently, we perform a single pass scan of the BAM file. Any read that satisfies all the following criteria is flagged as a duplicate: 1) it maps to a location within 8 base pairs of the previous read, 2) it shares the same cell barcode and molecular barcode with the previous read, and 3) its UMI exhibits a Levenshtein distance of less than 2 base pairs from the UMI of the previous read. All identified PCR duplicates are subsequently removed to ensure the integrity of downstream analyses.

Merging I7 index separated libraries

In the final step, we consolidate all the deduplicated BAM files obtained from each I7 index-marked library into a unified, comprehensive BAM file. This final output is generated in a sorted BAM format, which captures essential information including the cell barcode, molecular

barcode (comprising the 10X and I7 index), and the mapping location of the insert for both DNA and RNA (distinctly stored in separate files).

Supplementary Table 3-1: A partial list of single-cell multi-modal technologies.
"CARs": chromatin accessible regions. "Co-complex pairs": variations of Hi-C technologies that sample a pair of DNA fragments from a chromatin complex involving either pairwise or multiplex interactions. "Multiplex interactions": technologies that can measure multiplex interactions and thus differentiate pairwise and multiplex interactions.

Method	Imaging/Omics	Surface protein	CARs	Histone modification	DNA methylation	3D genome organization			Gene expression	RNA-chromatin association
						Chromatin traces	Co-complex pairs	Multiplex interactions		
DNA seqFISH+	Imaging			✓		✓			✓	
ORCA	Imaging					✓			✓	
MINA	Imaging					✓			✓	
Snare-seq	Omics		✓						✓	
SHARE-seq	Omics		✓						✓	
scNMT-seq	Omics		✓		✓					
scNOMeRe-seq	Omics		✓		✓				✓	
Pair-tag	Omics			✓					✓	
sn-m3C-seq	Omics				✓		✓			
CITE-seq	Omics	✓							✓	
HiRES	Omics						✓		✓	
MUSIC	Omics							✓	✓	✓

Supplementary Table 3-2: Sequences of the RNA Linker, the DNA Linker, Universal adaptor, Index adaptors, and Cell Barcodes.
https://docs.google.com/spreadsheets/d/1SiIrvWFqIWpHL9TDgs2Rd1RxiE08vFvq9kHXLxWHEOc/edit?usp=sharing

Supplementary Table 3-3: Number of reads retained in every data processing step in the mixed-species library.
(b) Imperfect_10x indicates the number of MUSIC sequencing reads that do not have the 16bp 10x barcode sequence that can perfectly align to the whitelist provided by 10X genomics. (c) Incomplete_CB: number of reads that don't have three cell barcode sequences that can map to the cell barcodes whitelist each within 2 edit distances. (h, i) DNA or RNA reads that are uniquely mapped to the human genome. (j, k) PCR duplicates removed uniquely mapped barcode-complete reads. Details of sequencing reads processing steps can be found in the Methods section.

Steps	Reads Number	Ratio
(a) Raw_reads	3,067,956,666	
(b) Imperfect_10x	101,139,708	(b)/(a) 3.30%
(c) Incomplete_CB	1,001,280,905	(c)/(a) 32.64%
(d) Adaptor_removed_DNA	9,591,538	(d)/(a) 0.31%
(e) Adaptor_removed_RNA	29,416,328	(e)/(a) 0.96%
(f) Barcode complete, adaptor removed DNA reads	1,291,977,026	(f)/(a) 42.11%
(g) Barcode complete, adaptor removed RNA reads	195,044,878	(g)/(a) 6.36%
(h) DNA uniquely mapped reads	732,019,497	(h)/(a) 23.86%
(i) RNA uniquely mapped reads	45,719,121	(i)/(a) 1.50%
(j) DNA uniquely mapped, non-duplicate, and barcode-complete reads (UMNDBC)	501,566,553	(j)/(a) 16.35%
(k) RNA uniquely mapped, non-duplicate, and barcode-complete reads (UMNDBC)	31,666,815	(k)/(a) 1.03%

Supplementary Table 3-4: Brain sample metadata.

Sample ID	Case ID	Sex	Age at death	ApoE	PlaqueTotal	TangleTotal	Braak score
B1	12-49	M	63	3/4	14.5	15	VI
B2	11-38	M	64	3/3	14.5	15	VI
B3	18-78	M	71	3/3	0	4	III
B4	13-40	M	73	3/3	0	2.25	II
B5	19-40	F	75	3/4	15	15	VI
B6	13-49	F	75	3/3	0	2.5	II
B7	17-60	F	82	3/3	15	14.5	VI
B8	15-46	F	82	2/3	1	1.5	I
B9	11-107	M	71	4/4	14	15	VI
B10	16-32	M	79	2/3	0	1	II
B11	10-22	F	59	3/3	1	0.5	I
B12	09-46	F	60	3/3	15	15	VI
B13	16-18	M	≥90	2/4	14.5	14	VI
B14	10-26	F	≥90	3/3	0	6.5	III

Supplementary Table 3-5: Transcriptome profiling statistics from multi-modal single cell sequencing technologies.
All statistics for each technique are computed based on the raw count matrix downloaded from each published dataset (CITE-seq: GSE100866_PBMC, SNARE-seq: GSE126074_AdBrainCortex, PairTag: GSE152020, snRNA-seq: Synapse: syn18485175).

Method	Median RNA reads / cell	Medium genes / cell	Number of cells in Raw count table	Number of genes in Raw count table	Sample
snRNA-seq	1,973	1,348	80,660	18,192	Human prefrontal cortex
CITE-seq	1,416	770	7,985	29,929	Human PBMC cells
SNARE-seq	1,332	887	10,309	33,160	Mouse adult brain cerebral cortex
Pair-Tag	845	626	64,849	52,781	Mouse frontal cortex and hippocampus
MUSIC H1	6,559	3,684	2,546	39,007	Human embryonic H1 cell line
MUSIC FC	1,136	853	9,067	22,373	Frozen human frontal cortex

Supplementary Table 3-6: Marker genes for each cell type and subtype.

Type	Marker genes	Subtype	Marker genes
ExN	SLC17A7, CAMK2A, NRGN	L2/3_IT	CBLN2, EPHA6, LAMA2, PDZD2, CUX2
		L5_IT	RORB, IL1RAPL2
		L6_IT	CDH9, THEMIS, CDH12
		L5/6_NP	NPSR1.AS1, HTR2C, ITGA8, ZNF385D, CDH6, NXPH2
		L6_CT	ADAMTSL1, TRPM3, SEMA3E, EGFEM1P
InN	GAD1, GAD2	Vip	GALNTL6, GRM7
		Lamp5	KIT, FBXL7
		Sst	SOX6, NXPH1
		Pvalb	ADAMTS17, DPP10
Mic	CD74, CSF1R, C3	MS4A-	N/A
		MS4A+	MS4A7, MS4A14, MS4A6E
Oli	MBP, PLP1, MOBP		
Opc	PDGFRA, VCAN, CSPG4		
Ast	AQP4, GFAP		
Vas	FLT1, EMCN, VWF		

REFERENCES

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

Akgol Oksuz, B., Yang, L., Abraham, S., Venev, S.V., Krietenstein, N., Parsi, K.M., Ozadam, H., Oomen, M.E., Nand, A., Mao, H., Genga, R.M.J., Maehr, R., Rando, O.J., Mirny, L.A., Gibcus, J.H., Dekker, J., 2021. Systematic evaluation of chromosome conformation capture assays. Nat. Methods 18, 1046–1055.

Allou, L., Balzano, S., Magg, A., Quinodoz, M., Royer-Bertrand, B., Schöpflin, R., Chan, W.-L., Speck-Martins, C.E., Carvalho, D.R., Farage, L., Lourenço, C.M., Albuquerque, R., Rajagopal, S., Nampoothiri, S., Campos-Xavier, B., Chiesa, C., Niel-Bütschi, F., Wittler, L., Timmermann, B., Spielmann, M., Robson, M.I., Ringel, A., Heinrich, V., Cova, G., Andrey, G., Prada-Medina, C.A., Pescini-Gobert, R., Unger, S., Bonafé, L., Grote, P., Rivolta, C., Mundlos, S., Superti-Furga, A., 2021. Non-coding deletions identify Maenli lncRNA as a limb-specific En1 regulator. Nature 592, 93–98.

Arrastia, M.V., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A., Ismagilov, R.F., Guttman, M., 2022. Single-cell measurement of higher-order 3D genome organization with scSPRITE. Nat. Biotechnol. 40, 64–73.

Aumiller, W.M., Jr, Keating, C.D., 2016. Phosphorylation-mediated RNA/peptide complex coacervation as a model for intracellular liquid organelles. Nat. Chem. 8, 129–137.

Bakken, T.E., Jorstad, N.L., Hu, Q., Lake, B.B., Tian, W., Kalmbach, B.E., Crow, M., Hodge, R.D., Krienen, F.M., Sorensen, S.A., Eggermont, J., Yao, Z., Aevermann, B.D., Aldridge, A.I., Bartlett, A., Bertagnolli, D., Casper, T., Castanon, R.G., Crichton, K., Daigle, T.L., Dalley, R., Dee, N., Dembrow, N., Diep, D., Ding, S.-L., Dong, W., Fang, R., Fischer, S., Goldman, M., Goldy, J., Graybuck, L.T., Herb, B.R., Hou, X., Kancherla, J., Kroll, M., Lathia, K., van Lew, B., Li, Y.E., Liu, C.S., Liu, H., Lucero, J.D., Mahurkar, A., McMillen, D., Miller, J.A., Moussa, M., Nery, J.R., Nicovich, P.R., Niu, S.-Y., Orvis, J., Osteen, J.K., Owen, S., Palmer, C.R., Pham, T., Plongthongkum, N., Poirion, O., Reed, N.M., Rimorin, C., Rivkin, A., Romanow, W.J., Sedeño-Cortés, A.E., Siletti, K., Somasundaram, S., Sulc, J., Tieu, M., Torkelson, A., Tung, H., Wang, X., Xie, F., Yanny, A.M., Zhang, R., Ament, S.A., Behrens, M.M., Bravo, H.C., Chun, J., Dobin, A., Gillis, J., Hertzano, R., Hof, P.R., Höllt, T., Horwitz, G.D., Keene, C.D., Kharchenko, P.V., Ko, A.L., Lelieveldt, B.P., Luo, C., Mukamel, E.A., Pinto-Duarte, A., Preissl, S., Regev, A., Ren, B., Scheuermann, R.H., Smith, K., Spain, W.J., White, O.R., Koch, C., Hawrylycz, M., Tasic, B., Macosko, E.Z., McCarroll, S.A., Ting, J.T., Zeng, H., Zhang, K., Feng, G., Ecker, J.R., Linnarsson, S., Lein, E.S., 2021. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature 598, 111–119.

Barutcu, A.R., Blencowe, B.J., Rinn, J.L., 2019. Differential contribution of steady-state RNA and active transcription in chromatin organization. EMBO Rep. 20, e48068.

Beach, T.G., Adler, C.H., Sue, L.I., Serrano, G., Shill, H.A., Walker, D.G., Lue, L., Roher, A.E., Dugger, B.N., Maarouf, C., Birdsill, A.C., Intorcia, A., Saxon-Labelle, M., Pullen, J., Scroggins, A., Filon, J., Scott, S., Hoffman, B., Garcia, A., Caviness, J.N., Hentz, J.G., Driver-Dunckley, E., Jacobson, S.A., Davis, K.J., Belden, C.M., Long, K.E., Malek-Ahmadi, M., Powell, J.J., Gale, L.D., Nicholson, L.R., Caselli, R.J., Woodruff, B.K., Rapscak, S.Z., Ahern, G.L., Shi, J., Burke, A.D., Reiman, E.M., Sabbagh, M.N., 2015.

Arizona Study of Aging and Neurodegenerative Disorders and Brain and Body Donation Program. Neuropathology 35, 354–389.
Bell, J.C., Jukam, D., Teran, N.A., Risca, V.I., Smith, O.K., Johnson, W.L., Skotheim, J.M., Greenleaf, W.J., Straight, A.F., 2018. Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. Elife 7, e27024.
Bhadra, M., Howell, P., Dutta, S., Heintz, C., Mair, W.B., 2020. Alternative splicing in aging and longevity. Hum. Genet. 139, 357–369.
Bilotti, K., Potapov, V., Pryor, J.M., Duckworth, A.T., Keck, J.L., Lohman, G.J.S., 2022. Mismatch discrimination and sequence bias during end-joining by DNA ligases. Nucleic Acids Res. 50, 4647–4658.
Bonetti, A., Agostini, F., Suzuki, A.M., Hashimoto, K., Pascarella, G., Gimenez, J., Roos, L., Nash, A.J., Ghilotti, M., Cameron, C.J.F., Valentine, M., Medvedeva, Y.A., Noguchi, S., Agirre, E., Kashi, K., Samudyata, Luginbühl, J., Cazzoli, R., Agrawal, S., Luscombe, N.M., Blanchette, M., Kasukawa, T., Hoon, M. de, Arner, E., Lenhard, B., Plessy, C., Castelo-Branco, G., Orlando, V., Carninci, P., 2020. RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. Nat. Commun. 11, 1018.
Bose, D.A., Donahue, G., Reinberg, D., Shiekhattar, R., Bonasio, R., Berger, S.L., 2017. RNA Binding to CBP Stimulates Histone Acetylation and Transcription. Cell 168, 135-149.e22.
Brahma, S., Henikoff, S., 2023. RNA Polymerase II, the BAF remodeler and transcription factors synergize to evict nucleosomes. bioRxiv. https://doi.org/10.1101/2023.01.22.525083
Bryois, J., Calini, D., Macnair, W., Foo, L., Urich, E., Ortmann, W., Iglesias, V.A., Selvaraj, S., Nutma, E., Marzin, M., Amor, S., Williams, A., Castelo-Branco, G., Menon, V., De Jager, P., Malhotra, D., 2022. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. Nat. Neurosci. 25, 1104–1112.
Calandrelli, R., Wen, X., Charles Richard, J.L., Luo, Z., Nguyen, T.C., Chen, C.-J., Qi, Z., Xue, S., Chen, W., Yan, Z., Wu, W., Zaleta-Rivera, K., Hu, R., Yu, M., Wang, Y., Li, W., Ma, J., Ren, B., Zhong, S., 2023. Genome-wide analysis of the interplay between chromatin-associated RNA and 3D genome organization in human cells. Nat. Commun. 14, 6519.
Calandrelli, R., Xu, L., Luo, Y., Wu, W., Fan, X., Nguyen, T., Chen, C.-J., Sriram, K., Tang, X., Burns, A.B., Natarajan, R., Chen, Z.B., Zhong, S., 2020. Stress-induced RNA-chromatin interactions promote endothelial dysfunction. Nat. Commun. 11, 5211.
Cerghet, M., Skoff, R.P., Bessert, D., Zhang, Z., Mullins, C., Ghandour, M.S., 2006. Proliferation and death of oligodendrocytes and myelin proteins are differentially regulated in male and female rodents. J. Neurosci. 26, 1439–1447.
Chao, S.H., Price, D.H., 2001. Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. J. Biol. Chem. 276, 31793–31799.
Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., Cui, Q., 2013. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res. 41, D983-6.
Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., Zhuang, X., 2015. Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090.

Chen, S., Lake, B.B., Zhang, K., 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat. Biotechnol. 37, 1452–1457.
Chen, W., Yan, Z., Li, S., Huang, N., Huang, X., Zhang, J., Zhong, S., 2018. RNAs as Proximity-Labeling Media for Identifying Nuclear Speckle Positions Relative to the Genome. iScience 4, 204–215.
Chu, H.-P., Cifuentes-Rojas, C., Kesner, B., Aeby, E., Lee, H.-G., Wei, C., Oh, H.J., Boukhali, M., Haas, W., Lee, J.T., 2017. TERRA RNA Antagonizes ATRX and Protects Telomeres. Cell 170, 86-101.e16.
Clark, S.J., Argelaguet, R., Kapourani, C.-A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C., Stegle, O., Reik, W., 2018. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. Nat. Commun. 9, 781.
Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., Lawrence, J.B., 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell 33, 717–726.
Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J., Meyer, B.J., 2015. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 523, 240–244.
Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'Shea, C.C., Park, P.J., Ren, B., Politz, J.C.R., Shendure, J., Zhong, S., 4D Nucleome Network, 2017. The 4D nucleome project. Nature 549, 219–226.
Dekker, J., Rippe, K., Dekker, M., Kleckner, N., 2002. Capturing chromosome conformation. Science 295, 1306–1311.
Deming, Y., Filipello, F., Cignarella, F., Cantoni, C., Hsu, S., Mikesell, R., Li, Z., Del-Aguila, J.L., Dube, U., Farias, F.G., Bradley, J., Budde, J., Ibanez, L., Fernandez, M.V., Blennow, K., Zetterberg, H., Heslegrave, A., Johansson, P.M., Svensson, J., Nellgård, B., Lleo, A., Alcolea, D., Clarimon, J., Rami, L., Molinuevo, J.L., Suárez-Calvet, M., Morenas-Rodríguez, E., Kleinberger, G., Ewers, M., Harari, O., Haass, C., Brett, T.J., Benitez, B.A., Karch, C.M., Piccio, L., Cruchaga, C., 2019. The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. Sci. Transl. Med. 11. https://doi.org/10.1126/scitranslmed.aau2291
Depp, C., Sun, T., Sasmita, A.O., Spieth, L., Berghoff, S.A., Nazarenko, T., Overhoff, K., Steixner-Kumar, A.A., Subramanian, S., Arinrad, S., Ruhwedel, T., Möbius, W., Göbbels, S., Saher, G., Werner, H.B., Damkou, A., Zampar, S., Wirths, O., Thalmann, M., Simons, M., Saito, T., Saido, T., Krueger-Burg, D., Kawaguchi, R., Willem, M., Haass, C., Geschwind, D., Ehrenreich, H., Stassart, R., Nave, K.-A., 2023. Myelin dysfunction drives amyloid-β deposition in models of Alzheimer's disease. Nature 618, 349–357.
Deshpande, A.S., Ulahannan, N., Pendleton, M., Dai, X., Ly, L., Behr, J.M., Schwenk, S., Liao, W., Augello, M.A., Tyer, C., Rughani, P., Kudman, S., Tian, H., Otis, H.G., Adney, E., Wilkes, D., Mosquera, J.M., Barbieri, C.E., Melnick, A., Stoddart, D., Turner, D.J., Juul, S., Harrington, E., Imieliński, M., 2022. Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nat. Biotechnol. 40, 1488–1499.

Dhanoa, J.K., Sethi, R.S., Verma, R., Arora, J.S., Mukhopadhyay, C.S., 2018. Long non-coding RNA: its evolutionary relics and biological implications in mammals: a review. Hanguk Tongmul Chawon Kwahakhoe Chi 60, 25.

Disteche, C.M., 2012. Dosage compensation of the sex chromosomes. Annu. Rev. Genet. 46, 537–560.

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

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R., 2012. Landscape of transcription in human cells. Nature 489, 101–108.

Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., Aiden, E.L., 2016a. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst 3, 99–101.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., Aiden, E.L., 2016b. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst 3, 95–98.

ENCODE Project, 2012. The ENCODE Project Encyclopedia of DNA Elements.

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

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., Plath, K., Guttman, M., 2013. The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science 341, 1237973.

Engreitz, J.M., Sirokman, K., McDonel, P., Shishkin, A.A., Surka, C., Russell, P., Grossman, S.R., Chow, A.Y., Guttman, M., Lander, E.S., 2014. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. Cell 159, 188–199.

Flyamer, I.M., Gassler, J., Imakaev, M., Brandão, H.B., Ulianov, S.V., Abdennur, N., Razin, S.V., Mirny, L.A., Tachibana-Konwalski, K., 2017. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature 544, 110–114.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I.T., García Girón, C., Gonzalez, J.M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O.G., Lagarde,

J., Martin, F.J., Martínez, L., Mohanan, S., Muir, P., Navarro, F.C.P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B.M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J.S., Gerstein, M., Guigó, R., Hubbard, T.J.P., Kellis, M., Paten, B., Reymond, A., Tress, M.L., Flicek, P., 2019. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773.

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I.T., García Girón, C., Gonzalez, J.M., Grego, T., Hardy, M., Hourlier, T., Howe, K.L., Hunt, T., Izuogu, O.G., Johnson, R., Martin, F.J., Martínez, L., Mohanan, S., Muir, P., Navarro, F.C.P., Parker, A., Pei, B., Pozo, F., Riera, F.C., Ruffier, M., Schmitt, B.M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczynska-Ratajczak, B., Wolf, M.Y., Xu, J., Yang, Y.T., Yates, A., Zerbino, D., Zhang, Y., Choudhary, J.S., Gerstein, M., Guigó, R., Hubbard, T.J.P., Kellis, M., Paten, B., Tress, M.L., Flicek, P., 2021. GENCODE 2021. Nucleic Acids Res. 49, D916–D923.

Franzén, O., Gan, L.-M., Björkegren, J.L.M., 2019. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019. https://doi.org/10.1093/database/baz046

Fu, X.D., Maniatis, T., 1990. Factor required for mammalian spliceosome assembly is localized to discrete regions in the nucleus. Nature 343, 437–441.

G Hendrickson, D., Kelley, D.R., Tenen, D., Bernstein, B., Rinn, J.L., 2016. Widespread RNA binding by chromatin-associated proteins. Genome Biol. 17, 28.

Gavrilov, A.A., Sultanov, R.I., Magnitov, M.D., Galitsyna, A.A., Dashinimaev, E.B., Lieberman Aiden, E., Razin, S.V., 2022. RedChIP identifies noncoding RNAs associated with genomic sites occupied by Polycomb and CTCF proteins. Proc. Natl. Acad. Sci. U. S. A. 119. https://doi.org/10.1073/pnas.2116222119

Genga, R.M.J., Kernfeld, E.M., Parsi, K.M., Parsons, T.J., Ziller, M.J., Maehr, R., 2019. Single-Cell RNA-Sequencing-Based CRISPRi Screening Resolves Molecular Drivers of Early Human Endoderm Development. Cell Rep. 27, 708-718.e10.

González, F., Zhu, Z., Shi, Z.-D., Lelli, K., Verma, N., Li, Q.V., Huangfu, D., 2014. An iCRISPR platform for rapid, multiplexable, and inducible genome editing in human pluripotent stem cells. Cell Stem Cell 15, 215–226.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., van Steensel, B., 2008. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature 453, 948–951.

Guo, C.-J., Xu, G., Chen, L.-L., 2020. Mechanisms of Long Noncoding RNA Nuclear Retention. Trends Biochem. Sci. 45, 947–960.

Hall, L.L., Lawrence, J.B., 2016. RNA as a fundamental component of interphase chromosomes: could repeats prove key? Curr. Opin. Genet. Dev. 37, 137–147.

Hansen, A.S., Hsieh, T.-H.S., Cattoglio, C., Pustova, I., Saldaña-Meyer, R., Reinberg, D., Darzacq, X., Tjian, R., 2019. Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF. Mol. Cell 76, 395-411.e13.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou,

E.P., Jain, J., Srivastava, A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R., Smibert, P., Satija, R., 2021. Integrated analysis of multimodal single-cell data. Cell 184, 3573-3587.e29.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589.

Henninger, J.E., Oksuz, O., Shrinivas, K., Sagi, I., LeRoy, G., Zheng, M.M., Andrews, J.O., Zamudio, A.V., Lazaris, C., Hannett, N.M., Lee, T.I., Sharp, P.A., Cissé, I.I., Chakraborty, A.K., Young, R.A., 2021. RNA-Mediated Feedback Control of Transcriptional Condensates. Cell 184, 207-225.e24.

Holmes, D.S., Mayfield, J.E., Sander, G., Bonner, J., 1972. Chromosomal RNA: its properties. Science 177, 72–74.

Holmes, Z.E., Hamilton, D.J., Hwang, T., Parsonnet, N.V., Rinn, J.L., Wuttke, D.S., Batey, R.T., 2020. The Sox2 transcription factor binds RNA. Nat. Commun. 11, 1805.

Huang, W., Kew, C., Fernandes, S. de A., Löhrke, A., Han, L., Demetriades, C., Antebi, A., 2022. Decreased spliceosome fidelity and egl-8 intron retention inhibit mTORC1 signaling to promote longevity. Nat Aging 2, 796–808.

Ilik, İ.A., Malszycki, M., Lübke, A.K., Schade, C., Meierhofer, D., Aktaş, T., 2020. SON and SRRM2 are essential for nuclear speckle formation. Elife 9. https://doi.org/10.7554/eLife.60579

Jain, A., Vale, R.D., 2017. RNA phase transitions in repeat expansion disorders. Nature 546, 243–247.

Ji, Z., Song, R., Huang, H., Regev, A., Struhl, K., 2016. Transcriptome-scale RNase-footprinting of RNA-protein complexes. Nat. Biotechnol. 34, 410–413.

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

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J.M., Ouellette, S.B., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B.H., Pfister, H., Mirny, L.A., Park, P.J., Gehlenborg, N., 2018. HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biol. 19, 125.

Kim, B.R., Coyaud, E., Laurent, E.M.N., St-Germain, J., Van de Laar, E., Tsao, M.-S., Raught, B., Moghal, N., 2017. Identification of the SOX2 Interactome by BioID Reveals EP300 as a Mediator of SOX2-dependent Squamous Differentiation and Lung Squamous Cell Carcinoma Growth. Mol. Cell. Proteomics 16, 1864–1888.

Kim, J., Han, K.Y., Khanna, N., Ha, T., Belmont, A.S., 2019. Nuclear speckle fusion via long-range directional motion regulates speckle morphology after transcriptional inhibition. J. Cell Sci. 132. https://doi.org/10.1242/jcs.226563

Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J., Peterson, H., 2020. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. F1000Res. 9. https://doi.org/10.12688/f1000research.24956.2

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., Raychaudhuri, S., 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods 16, 1289–1296.

Köster, J., Rahmann, S., 2012. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522.

Krietenstein, N., Abraham, S., Venev, S.V., Abdennur, N., Gibcus, J., Hsieh, T.-H.S., Parsi, K.M., Yang, L., Maehr, R., Mirny, L.A., Dekker, J., Rando, O.J., 2020. Ultrastructural Details of Mammalian Chromosome Architecture. Mol. Cell 78, 554-565.e7.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P.V., Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22, 1813–1831.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

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., 2019. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat. Methods 16, 999–1006.

Lee, J.T., 2012. Epigenetic regulation by long noncoding RNAs. Science 338, 1435–1439.

Lee, J.T., 2011. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. Nat. Rev. Mol. Cell Biol. 12, 815–826.

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

Li, L., Williams, P., Ren, W., Wang, M.Y., Gao, Z., Miao, W., Huang, M., Song, J., Wang, Y., 2021. YY1 interacts with guanine quadruplexes to regulate DNA looping and gene expression. Nat. Chem. Biol. 17, 161–168.

Li, X., Fu, X.-D., 2019. Chromatin-associated RNAs as facilitators of functional genomic interactions. Nat. Rev. Genet. 20, 503–519.

Li, X., Zhou, B., Chen, L., Gou, L.-T., Li, H., Fu, X.-D., 2017a. GRID-seq reveals the global RNA-chromatin interactome. Nat. Biotechnol. 35, 940–950.

Li, X., Zhou, B., Chen, L., Gou, L.-T., Li, H., Fu, X.-D., 2017b. GRID-seq reveals the global RNA–chromatin interactome. Nat. Biotechnol. 35, 940–950.

Liang, L., Cao, C., Ji, L., Cai, Z., Wang, D., Ye, R., Chen, J., Yu, X., Zhou, J., Bai, Z., Wang, R., Yang, X., Zhu, P., Xue, Y., 2023. Complementary Alu sequences mediate enhancer-promoter selectivity. Nature 619, 868–875.

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, 289–293.

Limouse, C., Smith, O.K., Jukam, D., Fryer, K.A., Greenleaf, W.J., Straight, A.F., 2023. Global mapping of RNA-chromatin contacts reveals a proximity-dominated connectivity model for ncRNA-gene interactions. Nat. Commun. 14, 6073.

Liu, H., Ma, Y., Yu, J., Chen, X., Wang, S., Jia, Y., Ding, N., Jin, X., Zhang, Y., Xu, J., Li, X., 2023. Insight into the regulatory mechanism of dynamic chromatin 3D interactions during cardiomyocyte differentiation in human. Mol. Ther. Nucleic Acids 33, 629–641.

Liu, M., Lu, Y., Yang, B., Chen, Y., Radda, J.S.D., Hu, M., Katz, S.G., Wang, S., 2020. Multiplexed imaging of nucleome architectures in single cells of mammalian tissue. Nat. Commun. 11, 2907.

Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y., Pape, U.J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., Liu, X.S., 2011. Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol. 12, R83.

Liu, Z., Chen, Y., Xia, Q., Liu, M., Xu, H., Chi, Y., Deng, Y., Xing, D., 2023. Linking genome structures to functions by simultaneous single-cell Hi-C and RNA-seq. Science 380, 1070–1076.

Long, Y., Hwang, T., Gooding, A.R., Goodrich, K.J., Rinn, J.L., Cech, T.R., 2020. RNA is essential for PRC2 chromatin occupancy and function in human pluripotent stem cells. Nat. Genet. 52, 931–938.

Lu, L., Liu, X., Huang, W.-K., Giusti-Rodríguez, P., Cui, J., Zhang, S., Xu, W., Wen, Z., Ma, S., Rosen, J.D., Xu, Z., Bartels, C.F., Kawaguchi, R., Hu, M., Scacheri, P.C., Rong, Z., Li, Yun, Sullivan, P.F., Song, H., Ming, G.-L., Li, Yan, Jin, F., 2020. Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. Mol. Cell 79, 521-534.e15.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., Mundlos, S., 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161, 1012–1025.

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, 1103-1116.e20.

Mao, S., Su, J., Wang, L., Bo, X., Li, C., Chen, H., 2023. A transcriptome-based single-cell biological age model and resource for tissue-specific aging measures. Genome Res. 33, 1381–1394.

Margueron, R., Reinberg, D., 2011. The Polycomb complex PRC2 and its mark in life. Nature 469, 343–349.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10–12.

Mateo, L.J., Murphy, S.E., Hafner, A., Cinquini, I.S., Walker, C.A., Boettiger, A.N., 2019. Visualizing DNA folding and RNA in embryos at single-cell resolution. Nature 568, 49–54.

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A.J., Ransohoff, R.M., Hafler, B.P., Bennett, D.A., Kellis, M., Tsai, L.-H., 2019. Single-cell transcriptomic analysis of Alzheimer's disease. Nature 570, 332–337.

Mazuir, E., Fricker, D., Sol-Foulon, N., 2021. Neuron-Oligodendrocyte Communication in Myelination of Cortical GABAergic Cells. Life 11. https://doi.org/10.3390/life11030216

Miao, Y., Ajami, N.E., Huang, T.-S., Lin, F.-M., Lou, C.-H., Wang, Y.-T., Li, S., Kang, J., Munkacsi, H., Maurya, M.R., Gupta, S., Chien, S., Subramaniam, S., Chen, Z., 2018.

Enhancer-associated long non-coding RNA LEENE regulates endothelial nitric oxide synthase and endothelial function. Nat. Commun. 9, 292.
Miele, A., Dekker, J., 2009. Mapping cis- and trans- chromatin interaction networks using chromosome conformation capture (3C). Methods Mol. Biol. 464, 105–121.
Miele, A., Gheldof, N., Tabuchi, T.M., Dostie, J., Dekker, J., 2006. Mapping chromatin interactions by chromosome conformation capture. Curr. Protoc. Mol. Biol. Chapter 21, Unit 21.11.
Mirny, L.A., 2011. The fractal globule as a model of chromatin architecture in the cell. Chromosome Res. 19, 37–51.
Morris, K.V., Chan, S.W.-L., Jacobsen, S.E., Looney, D.J., 2004. Small interfering RNA-induced transcriptional gene silencing in human cells. Science 305, 1289–1292.
Mourad, R., Cuvier, O., 2015. Predicting the spatial organization of chromosomes using epigenetic data. Genome Biol.
Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., Fraser, P., 2013. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502, 59–64.
Nagano, T., Lubling, Y., Várnai, C., Dudley, C., Leung, W., Baran, Y., Mendelson Cohen, N., Wingett, S., Fraser, P., Tanay, A., 2017. Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature 547, 61–67.
Nalavade, R., Griesche, N., Ryan, D.P., Hildebrand, S., Krauss, S., 2013. Mechanisms of RNA-induced toxicity in CAG repeat disorders. Cell Death Dis. 4, e752.
Nguyen, D.K., Disteche, C.M., 2006. Dosage compensation of the active X chromosome in mammals. Nat. Genet. 38, 47–53.
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., 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385.
Nozawa, R.-S., Boteva, L., Soares, D.C., Naughton, C., Dun, A.R., Buckle, A., Ramsahoye, B., Bruton, P.C., Saleeb, R.S., Arnedo, M., Hill, B., Duncan, R.R., Maciver, S.K., Gilbert, N., 2017. SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin-Associated RNAs. Cell 169, 1214-1227.e18.
Nozawa, R.-S., Gilbert, N., 2019. RNA: Nuclear Glue for Folding the Genome. Trends Cell Biol. 29, 201–211.
Oksuz, O., Henninger, J.E., Warneford-Thomson, R., Zheng, M.M., Erb, H., Vancura, A., Overholt, K.J., Hawken, S.W., Banani, S.F., Lauman, R., Reich, L.N., Robertson, A.L., Hannett, N.M., Lee, T.I., Zon, L.I., Bonasio, R., Young, R.A., 2023. Transcription factors interact with RNA to regulate genes. Mol. Cell 83, 2449-2463.e13.
Open2C, Abdennur, N., Abraham, S., Fudenberg, G., Flyamer, I.M., Galitsyna, A.A., Goloborodko, A., Imakaev, M., Oksuz, B.A., Venev, S.V., 2022. Cooltools: enabling high-resolution Hi-C analysis in Python. bioRxiv. https://doi.org/10.1101/2022.10.31.514564
Palazzo, A.F., Lee, E.S., 2018. Sequence Determinants for Nuclear Retention and Cytoplasmic Export of mRNAs and lncRNAs. Front. Genet. 9, 440.
Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., Kanduri, C., 2008. Kcnq1ot1 antisense noncoding RNA mediates

lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell 32, 232–246.
Pederson, T., 2019. Half a century of locating DNA and RNA in cells. FASEB J. 33, 8693–8694.
Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., Brockdorff, N., 1996. Requirement for Xist in X chromosome inactivation. Nature 379, 131–137.
Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., Furlong, L.I., 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. 48, D845–D855.
Place, R.F., Li, L.-C., Pookot, D., Noonan, E.J., Dahiya, R., 2008. MicroRNA-373 induces expression of genes with complementary promoter sequences. Proc. Natl. Acad. Sci. U. S. A. 105, 1608–1613.
Plongthongkum, N., Diep, D., Chen, S., Lake, B.B., Zhang, K., 2021. Scalable dual-omics profiling with single-nucleus chromatin accessibility and mRNA expression sequencing 2 (SNARE-seq2). Nat. Protoc. 16, 4992–5029.
Potapov, V., Ong, J.L., Kucera, R.B., Langhorst, B.W., Bilotti, K., Pryor, J.M., Cantor, E.J., Canton, B., Knight, T.F., Evans, T.C., Jr, Lohman, G.J.S., 2018. Comprehensive Profiling of Four Base Overhang Ligation Fidelity by T4 DNA Ligase and Application to DNA Assembly. ACS Synth. Biol. 7, 2665–2674.
Quinodoz, S.A., Jachowicz, J.W., Bhat, P., Ollikainen, N., Banerjee, A.K., Goronzy, I.N., Blanco, M.R., Chovanec, P., Chow, A., Markaki, Y., Thai, J., Plath, K., Guttman, M., 2021. RNA promotes the formation of spatial compartments in the nucleus. Cell 184, 5775-5790.e30.
Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., Trinh, V., Aznauryan, E., Russell, P., Cheng, C., Jovanovic, M., Chow, A., Cai, L., McDonel, P., Garber, M., Guttman, M., 2018. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell 174, 744-757.e24.
Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., Young, R.A., 2010. c-Myc regulates transcriptional pause release. Cell 141, 432–445.
Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z., Shendure, J., 2017. Massively multiplex single-cell Hi-C. Nat. Methods 14, 263–266.
Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., Huang, X., Shamim, M.S., Shin, J., Turner, D., Ye, Z., Omer, A.D., Robinson, J.T., Schlick, T., Bernstein, B.E., Casellas, R., Lander, E.S., Aiden, E.L., 2017. Cohesin Loss Eliminates All Loop Domains. Cell 171, 305-320.e24.
Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., Aiden, E.L., 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159, 1665–1680.
Reiff, S.B., Schroeder, A.J., Kırlı, K., Cosolo, A., Bakker, C., Mercado, L., Lee, S., Veit, A.D., Balashov, A.K., Vitzthum, C., Ronchetti, W., Pitman, K.M., Johnson, J., Ehmsen, S.R., Kerpedjiev, P., Abdennur, N., Imakaev, M., Öztürk, S.U., Çamoğlu, U., Mirny, L.A., Gehlenborg, N., Alver, B.H., Park, P.J., 2022. Author Correction: The 4D Nucleome

Data Portal as a resource for searching and visualizing curated nucleomics data. Nat. Commun. 13, 6561.
Rinn, J.L., Chang, H.Y., 2012. Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. 81, 145–166.
Rodríguez-Campos, A., Azorín, F., 2007. RNA is an integral component of chromatin that contributes to its structural organization. PLoS One 2, e1182.
Rosenberg, M., Blum, R., Kesner, B., Aeby, E., Garant, J.-M., Szanto, A., Lee, J.T., 2021. Motif-driven interactions between RNA and PRC2 are rheostats that regulate transcription elongation. Nat. Struct. Mol. Biol. 28, 103–117.
Saha, S., Hyman, A.A., 2017. RNA gets in phase. J. Cell Biol.
Saldaña-Meyer, R., Rodriguez-Hernaez, J., Escobar, T., Nishana, M., Jácome-López, K., Nora, E.P., Bruneau, B.G., Tsirigos, A., Furlan-Magaril, M., Skok, J., Reinberg, D., 2019. RNA Interactions Are Essential for CTCF-Mediated Genome Organization. Mol. Cell 76, 412-422.e5.
Schlesinger, S., Meshorer, E., 2019. Open Chromatin, Epigenetic Plasticity, and Nuclear Organization in Pluripotency. Dev. Cell 48, 135–150.
Sigova, A.A., Abraham, B.J., Ji, X., Molinie, B., Hannett, N.M., Guo, Y.E., Jangi, M., Giallourakis, C.C., Sharp, P.A., Young, R.A., 2015. Transcription factor trapping by RNA in gene regulatory elements. Science 350, 978–981.
Smith, K.P., Hall, L.L., Lawrence, J.B., 2020. Nuclear hubs built on RNAs and clustered organization of the genome. Curr. Opin. Cell Biol. 64, 67–76.
Sridhar, B., Rivas-Astroza, M., Nguyen, T.C., Chen, W., Yan, Z., Cao, X., Hebert, L., Zhong, S., 2017. Systematic Mapping of RNA-Chromatin Interactions In Vivo. Curr. Biol. 27, 602–609.
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.S., Lehner, B., Di Croce, L., Wutz, A., Hendrich, B., Klenerman, D., Laue, E.D., 2017. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature 544, 59–64.
Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P.K., Swerdlow, H., Satija, R., Smibert, P., 2017. Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods 14, 865–868.
Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., Satija, R., 2019. Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902.e21.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 102, 15545–15550.
Sun, J.H., Zhou, L., Emerson, D.J., Phyo, S.A., Titus, K.R., Gong, W., Gilgenast, T.G., Beagan, J.A., Davidson, B.L., Tassone, F., Phillips-Cremins, J.E., 2018. Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. Cell 175, 224-238.e15.
Takei, Y., Yun, J., Zheng, S., Ollikainen, N., Pierson, N., White, J., Shah, S., Thomassie, J., Suo, S., Eng, C.-H.L., Guttman, M., Yuan, G.-C., Cai, L., 2021a. Integrated spatial genomics reveals global architecture of single nuclei. Nature 590, 344–350.

Takei, Y., Zheng, S., Yun, J., Shah, S., Pierson, N., White, J., Schindler, S., Tischbirek, C.H., Yuan, G.-C., Cai, L., 2021b. Single-cell nuclear architecture across cell types in the mouse brain. Science 374, 586–594.
Tan, L., Shi, J., Moghadami, S., Parasar, B., Wright, C.P., Seo, Y., Vallejo, K., Cobos, I., Duncan, L., Chen, R., Deisseroth, K., 2023. Lifelong restructuring of 3D genome architecture in cerebellar granule cells. Science 381, 1112–1119.
Tan, L., Xing, D., Chang, C.-H., Li, H., Xie, X.S., 2018. Three-dimensional genome structures of single diploid human cells. Science 361, 924–928.
Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., Penn, O., Bakken, T., Menon, V., Miller, J., Fong, O., Hirokawa, K.E., Lathia, K., Rimorin, C., Tieu, M., Larsen, R., Casper, T., Barkan, E., Kroll, M., Parry, S., Shapovalova, N.V., Hirschstein, D., Pendergraft, J., Sullivan, H.A., Kim, T.K., Szafer, A., Dee, N., Groblewski, P., Wickersham, I., Cetin, A., Harris, J.A., Levi, B.P., Sunkin, S.M., Madisen, L., Daigle, T.L., Looger, L., Bernard, A., Phillips, J., Lein, E., Hawrylycz, M., Svoboda, K., Jones, A.R., Koch, C., Zeng, H., 2018. Shared and distinct transcriptomic cell types across neocortical areas. Nature 563, 72–78.
Thakur, J., Henikoff, S., 2020. Architectural RNA in chromatin organization. Biochem. Soc. Trans. 48, 1967–1978.
Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., Blencowe, B.J., Prasanth, S.G., Prasanth, K.V., 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell 39, 925–938.
Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M.A., Marshall, J.L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B.B., Castel, S.E., Karczewski, K.J., Aguet, F., Byrnes, A., GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, NIH/NCI, NIH/NHGRI, NIH/NIMH, NIH/NIDA, Biospecimen Collection Source Site—NDRI, Biospecimen Collection Source Site—RPCI, Biospecimen Core Resource—VARI, Brain Bank Repository—University of Miami Brain Endowment Bank, Leidos Biomedical—Project Management, ELSI Study, Genome Browser Data Integration &Visualization—EBI, Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz, Lappalainen, T., Regev, A., Ardlie, K.G., Hacohen, N., MacArthur, D.G., 2017. Landscape of X chromosome inactivation across human tissues. Nature 550, 244–248.
Urlaub, H., Raker, V.A., Kostka, S., Lührmann, R., 2001. Sm protein-Sm site RNA interactions within the inner ring of the spliceosomal snRNP core structure. EMBO J. 20, 187–196.
Vohhodina, J., Goehring, L.J., Liu, B., Kong, Q., Botchkarev, V.V., Jr, Huynh, M., Liu, Z., Abderazzaq, F.O., Clark, A.P., Ficarro, S.B., Marto, J.A., Hatchi, E., Livingston, D.M., 2021. BRCA1 binds TERRA RNA and suppresses R-Loop-based telomeric DNA damage. Nat. Commun. 12, 3542.
Wang, J., Lunyak, V.V., Jordan, I.K., 2013. BroadPeak: a novel algorithm for identifying broad peaks in diffuse ChIP-seq datasets. Bioinformatics 29, 492–493.
Wang, J., Vicente-García, C., Seruggia, D., Moltó, E., Fernandez-Miñán, A., Neto, A., Lee, E., Gómez-Skarmeta, J.L., Montoliu, L., Lunyak, V.V., Jordan, I.K., 2015. MIR

retrotransposon sequences provide insulators to the human genome. Proc. Natl. Acad. Sci. U. S. A. 112, E4428-37.
Wang, Yang, Yuan, P., Yan, Z., Yang, M., Huo, Y., Nie, Y., Zhu, X., Qiao, J., Yan, L., 2021. Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos. Nat. Commun. 12, 1247.
Wang, Yuchuan, Zhang, Y., Zhang, R., van Schaik, T., Zhang, L., Sasaki, T., Peric-Hupkes, D., Chen, Y., Gilbert, D.M., van Steensel, B., Belmont, A.S., Ma, J., 2021. SPIN reveals genome-wide landscape of nuclear compartmentalization. Genome Biol. 22, 36.
Weakley, S.M., Wang, H., Yao, Q., Chen, C., 2011. Expression and function of a large non-coding RNA gene XIST in human cancer. World J. Surg. 35, 1751–1756.
Wen, X., Zhong, S., 2023. Alu transposable elements rewire enhancer-promoter network through RNA pairing. Mol. Cell 83, 3234–3235.
Wienert, B., Shin, J., Zelin, E., Pestal, K., Corn, J.E., 2018. In vitro-transcribed guide RNAs trigger an innate immune response via the RIG-I pathway. PLoS Biol. 16, e2005840.
Wu, W., Yan, Z., Nguyen, T.C., Bouman Chen, Z., Chien, S., Zhong, S., 2019. Mapping RNA–chromatin interactions by sequencing with iMARGI. Nat. Protoc. 14, 3243–3272.
Xiao, R., Chen, J.-Y., Liang, Z., Luo, D., Chen, G., Lu, Z.J., Chen, Y., Zhou, B., Li, H., Du, X., Yang, Y., San, M., Wei, X., Liu, W., Lécuyer, E., Graveley, B.R., Yeo, G.W., Burge, C.B., Zhang, M.Q., Zhou, Y., Fu, X.-D., 2019. Pervasive Chromatin-RNA Binding Protein Interactions Enable RNA-Based Regulation of Transcription. Cell 178, 107-121.e18.
Xiong, X., James, B.T., Boix, C.A., Park, Y.P., Galani, K., Victor, M.B., Sun, N., Hou, L., Ho, L.-L., Mantero, J., Scannail, A.N., Dileep, V., Dong, W., Mathys, H., Bennett, D.A., Tsai, L.-H., Kellis, M., 2023. Epigenomic dissection of Alzheimer's disease pinpoints causal variants and reveals epigenome erosion. Cell 186, 4422-4437.e21.
Yan, Y., Wang, X., Chaput, D., Shin, M.-K., Koh, Y., Gan, L., Pieper, A.A., Woo, J.-A.A., Kang, D.E., 2022. X-linked ubiquitin-specific peptidase 11 increases tauopathy vulnerability in women. Cell 185, 3913-3930.e19.
Yan, Z., Huang, N., Wu, W., Chen, W., Jiang, Y., Chen, J., Huang, X., Wen, X., Xu, J., Jin, Q., Zhang, K., Chen, Z., Chien, S., Zhong, S., 2019. Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs. Proceedings of the National Academy of Sciences 116, 3328–3337.
Yang, F., Deng, X., Ma, W., Berletch, J.B., Rabaia, N., Wei, G., Moore, J.M., Filippova, G.N., Xu, J., Liu, Y., Noble, W.S., Shendure, J., Disteche, C.M., 2015. The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. Genome Biol. 16, 52.
Yesbolatova, A., Saito, Y., Kitamoto, N., Makino-Itou, H., Ajima, R., Nakano, R., Nakaoka, H., Fukui, K., Gamo, K., Tominari, Y., Takeuchi, H., Saga, Y., Hayashi, K.-I., Kanemaki, M.T., 2020. The auxin-inducible degron 2 technology provides sharp degradation control in yeast, mammalian cells, and mice. Nat. Commun. 11, 5701.
Yin, Y., Lu, J.Y., Zhang, X., Shao, W., Xu, Y., Li, P., Hong, Y., Cui, L., Shan, G., Tian, B., Zhang, Q.C., Shen, X., 2020. U1 snRNP regulates chromatin retention of noncoding RNAs. Nature 580, 147–150.
Yoh, S.M., Lucas, J.S., Jones, K.A., 2008. The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. Genes Dev. 22, 3422–3434.

Yu, R., McCauley, B., Dang, W., 2020. Loss of chromatin structural integrity is a source of stress during aging. Hum. Genet. 139, 371–380.
Zhang, S., Plummer, D., Lu, L., Cui, J., Xu, W., Wang, M., Liu, X., Prabhakar, N., Shrinet, J., Srinivasan, D., Fraser, P., Li, Y., Li, J., Jin, F., 2022. DeepLoop robustly maps chromatin interactions from sparse allele-resolved or single-cell Hi-C data at kilobase resolution. Nat. Genet. 54, 1013–1025.
Zhang, S., Übelmesser, N., Josipovic, N., Forte, G., Slotman, J.A., Chiang, M., Gothe, H.J., Gusmao, E.G., Becker, C., Altmüller, J., Houtsmuller, A.B., Roukos, V., Wendt, K.S., Marenduzzo, D., Papantonis, A., 2021. RNA polymerase II is required for spatial chromatin reorganization following exit from mitosis. Sci Adv 7, eabg8205.
Zhang, Y., Li, T., Preissl, S., Amaral, M.L., Grinstein, J.D., Farah, E.N., Destici, E., Qiu, Y., Hu, R., Lee, A.Y., Chee, S., Ma, K., Ye, Z., Zhu, Q., Huang, H., Fang, R., Yu, L., Izpisua Belmonte, J.C., Wu, J., Evans, S.M., Chi, N.C., Ren, B., 2019. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat. Genet. 51, 1380–1388.
Zheng, M., Tian, S.Z., Capurso, D., Kim, M., Maurya, R., Lee, B., Piecuch, E., Gong, L., Zhu, J.J., Li, Z., Wong, C.H., Ngan, C.Y., Wang, P., Ruan, X., Wei, C.-L., Ruan, Y., 2019. Multiplex chromatin interactions with single-molecule precision. Nature 566, 558–562.
Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.-H., Brown, M., Zhang, X., Meyer, C.A., Liu, X.S., 2019. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. 47, D729–D735.
Zhou, T., Zhang, R., Ma, J., 2021. The 3D Genome Structure of Single Cells. Annu Rev Biomed Data Sci 4, 21–41.
Zhu, C., Preissl, S., Ren, B., 2020. Single-cell multimodal omics: the power of many. Nat. Methods 17, 11–14.
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. Nat. Methods 18, 283–292.
Zufferey, M., Tavernari, D., Oricchio, E., Ciriello, G., 2018. Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 19, 217.