Higher-Order Chromatin States and Nuclear Structures Regulating Gene Expression

Thesis by

Isabel Nadine Goronzy

In Partial Fulfillment of the Requirements for the

Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2024

(Defended July 12, 2023)

Isabel Goronzy
ORCID: 0000-0002-6713-9192

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Professor Mitchell Guttman for nurturing my love for science, for fostering my curiosity and for challenging me to become a better scientist. Many thanks to the members of my thesis committee, Professors Elowitz, Chong, and Patcher, for their continuous support. I would also like to thank the members of the Guttman lab, who gave me guidance at every step and were my companions along this journey. I am incredibly proud to be a member of the academic community at CalTech, a very special environment for the pursuit of basic research. Finally, my deepest gratitude goes to my parents, who have always encouraged me to follow my dreams.

ABSTRACT

Although the same genome is present in every cell, each cell type orchestrates a distinct gene expression program, which can be rapidly adapted in response to stimuli. Accordingly, gene regulation is a highly complex, context-specific process that involves the dynamic interplay between numerous regulatory factors. Most methods to study these regulatory factors only measure pairwise interactions between molecules and are limited to mapping one regulatory protein at a time. Consequently, the combinatorial complexity of gene regulation at individual genomic loci and the functional consequence of many regulatory factors remain underexplored. To address this, we have developed new sequencing-based approaches and computational analyses to comprehensively profile, at unprecedented scale, the diverse gene regulatory landscape and directly establish the link between regulatory factors and transcriptional outcomes. In Chapter 2, we present Chromatin Immunoprecipitation Done-In-Parallel (ChIP-DIP), a highly multiplexed method for mapping hundreds of proteins to DNA within a single sample. ChIP-DIP increases the throughput of existing methods by > 100-fold and enables the production of consortium-scale, cell type-specific data within a single lab. Capitalizing on the scale and diversity provided by ChIP-DIP, we uncover unique quantitative combinations of histone modifications that define distinctive classes of regulatory elements. Specifically, we find features distinguishing classes of promoters that correspond to different polymerase activity, transcriptional levels, and gene types and find acetylation patterns distinguishing classes of enhancers that exhibit distinct activity states, induction potential, and regulatory potential. Next, in Chapter 3, we apply RNA-DNA SPRITE (RD-SPRITE), a method for simultaneous measurement of RNA and DNA organization, to investigate the functional relationship between genome structure and transcription. We demonstrate that RD-SPRITE precisely detects individual, nascent pre-mRNAs at their transcriptional locus and, as a result, can be used to assess the 3D genome structure present during active transcription. We find that RNA polymerase II transcription occurs within genomic structures previously thought to be inactive, such as the B compartment and DNA regions near the nucleolus. This suggests that active transcription can occur throughout the nucleus and argues against structural domains that preclude transcription. Overall, our findings highlight the ability of

RD-SPRITE to establish a structure-function link. Finally, in Chapter 4, we apply RD-SPRITE to study the transcriptional dependence of nuclear organization. We demonstrate that transcriptional inhibition leads to the loss of high-order structure around multiple RNA-processing bodies — the nucleolus, the scaRNA hub and the histone locus body — that are responsible for essential nuclear functions such as RNA processing and gene regulation. These findings suggest a role for RNA and nascent transcription in the formation and maintenance of long-range 3D contacts and critical nuclear compartments. In summary, we have developed new approaches to explore epigenomic and organizational complexity within the mammalian nucleus and have uncovered genome-wide principles of gene regulation.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. Goronzy IN, Quinodoz SA, Jachowicz JW, Ollikainen N, Bhat P, Guttman M. Simultaneous mapping of 3D structure and nascent RNAs argues against nuclear compartments that preclude transcription. Cell Rep. 2022 Nov 29;41(9):111730. doi: 10.1016/j.celrep.2022.111730. PMID: 36450242; PMCID: PMC9793828.

I.N.G. conceived of this project with S.A.Q. and M.G., led the effort to analyze and interpret data, performed all final computation analysis of RD-SPRITE data, generated figures, and wrote the paper.

2. Quinodoz SA, Jachowicz JW, Bhat P, Ollikainen N, Banerjee AK, Goronzy IN, Blanco MR, Chovanec P, Chow A, Markaki Y, Thai J, Plath K, Guttman M. RNA promotes the formation of spatial compartments in the nucleus. Cell. 2021 Nov 11;184(23):5775-5790.e30. doi: 10.1016/j.cell.2021.10.014. Epub 2021 Nov 4. PMID: 34739832; PMCID: PMC9115877.

I.N.G. performed the data processing and analysis of the ActD-treated SPRITE datasets; developed statistical methods for multiway RNA analyses; created new methods for data analysis and visualization; wrote scripts and constructed pipelines that enabled data curation; and wrote and provided comments and edits for the manuscript, figures, supplemental tables, and methods.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Number *Page*

Chapter 2

Number *Page*

Chapter 3

Number *Page*

Chapter 4

Number *Page*

Chapter 1

INTRODUCTION

1.1. THE MOTIVATION: WHY STUDY GENE REGULATION?

Although all cells in the human body contain an identical copy of the genome, every cell type orchestrates a distinctive gene expression program to take on a characteristic phenotype. Moreover, these cell-type-specific expression programs can be rapidly adapted to respond to stimuli; this modularity underpins both physiological (e.g. development) and pathological (e.g. malignant) transformation in cellular state. Accordingly, understanding how a single genomic sequence encodes complex and diverse functional outcomes and how variants within the genome (e.g., polymorphisms, mutations) generate phenotypic heterogeneity or pathology is the fundamental goal for the study of gene regulation. Simply put, how does the genome guide the spatial- and temporal-specific gene expression programs behind human health and disease?

National and international consortia have been and continue to be launched in an attempt to answer this question (**Figure 1**). In 2003, the Human Genome Project completed the first sequence of the human genome, providing a fundamental blueprint and exposing the "alphabet" of gene regulation[1–3]. Surprisingly, 95% of the genome was found to be non-coding, opening questions about the number of regulatory elements encoded in the genome and their roles in controlling gene expression. Simultaneously, growing interest was directed to epigenomics – the study of reversible modifications on DNA or histone proteins that affect gene transcription. Capitalizing on the advancements of sequencing technology, projects such as The Encyclopedia of DNA Elements (ENCODE) were designed to catalogue the regulome – a variety of DNA elements, cis-regulatory sequences and regions of chromatin structure that modulate gene expression[4,5]. Their goal was to "build a comprehensive parts list of functional elements in the human genome", exposing the "words and phrases" of gene regulation to reveal the links between DNA sequence, variable gene expression patterns, and the development of disease. Shortly following, advancements in RNA-sequencing technology revealed the vast non-coding elements of the transcriptome, and molecular studies of these RNAs demonstrated their functional relevance as regulators of gene transcription, both in healthy and disease states[6–8]. Finally,

multiple studies highlighted the role of nuclear organization (how the genome is folded, stored, and unpacked in the nucleus) in controlling genome function. The 4D Nucleome Program was established to map three-dimensional genome organization, its dynamics across time, and its relation to disease[9].

However, despite these tremendous efforts to decode the genome, the ability to predict gene expression from genomic and epigenomic measurements has remained frustratingly elusive[2]. In this perspective, we discuss current technical limitations and propose a new experimental and conceptual framework for achieving this goal.

1.2. THE APPROACH: STRATEGIES TO STUDY GENE REGULATION

To date, most approaches to characterize the epigenome and other regulatory features have relied on measuring pairwise interactions between molecules (e.g., Protein-DNA, Protein-RNA, DNA-DNA; **Figure 2**)[10–17]. Methods that map protein-nucleic acid interactions (e.g., ChIP-Seq, CLIP-Seq) are largely limited to studying the interactions of one regulatory protein at a time. In contrast, methods that map only nucleic acid interactions (e.g., HiC, Ric-seq) can provide a comprehensive readout of genomic and/or transcriptomic pairwise interactions. Finally, imaging modalities (e.g., immunofluorescence, RNA/DNA FISH) have complemented sequencing-based strategies by visualizing spatial localization of regulatory factors within single cells. Such methods cannot directly measure interactions but can provide relative distances between molecules and/or genomic regions of interest.

1.3. THE PROBLEM: ESTABLISHING THE LINK

While individual, pairwise methods have been incredibly useful for cataloguing regulatory elements and characterizing the behaviors of individual regulatory factors, the links between genome, epigenome, and molecular phenotype have remained nebulous. Sequencing the genome provided an “alphabet” and cataloguing regulatory elements and

factors provided “words and phrases” but the “grammar” of gene regulation is missing. That is, how do genomic elements and regulatory factors act in processes? How do they integrate diverse signals across time and space? How do they coordinate and adapt, rapidly and specifically, in response to stimuli? With the generation of larger mapping datasets, there has been growing appreciation that gene regulation involves the coordinated interplay between numerous regulatory factors, dynamically localizing at specific genomic regions and at specific times. These regulatory factors are thought to work in concert within high-order, multimodal assemblies and engage in a complex logic process to control gene expression. What these logic circuits and higher-order regulatory structures underlying gene regulation are has been underexplored.

Fundamentally, measurements of pairwise interactions cannot teach us the grammar of gene regulation. This is because pairwise measurements cannot provide a comprehensive picture of the context in which the regulatory factor acts. For instance, independently generated measurements of different regulatory factors cannot inform upon the link between these factors. Although various factors may have similar pairwise interaction profiles (e.g., two DNA-binding proteins localizing at the same region of DNA), it remains unclear whether these factors frequently co-occur, co-occur only in specific circumstances, or are mutually exclusive. Similarly, independently generated datasets cannot inform upon the functional link between a single factor and a phenotypic outcome. Correlations between a regulatory factor and independently generated RNA-seq datasets cannot establish the context-specific regulatory function of that factor.

To illustrate these limitations with pairwise measurements, consider the following simple example: ensemble measurements demonstrate that Protein A, Protein B, and Protein C bind to the promoter region of an actively transcribed gene. One possible ensemble model is that all three proteins co-bind to cause transcription (A+B+C → transcription). An alternative ensemble model would be that proteins A and B co-bind to cause high levels of transcription, protein C binds alone to cause low levels of transcription, but all three proteins bind together to inhibit transcription. (A+B→high transcription, C→low transcription, A+B+C→no transcription). With current pairwise methods, it is impossible

to determine which model is correct. Generalizing this example, even the simplest of logic gates (AND|OR|NOT) cannot be resolved using measurements of the components one-by-one.

In part, the inability to compare across independent measurements stems from the cell-to-cell heterogeneity underlying ensemble measurements. Growing appreciation for the degree of cell-to-cell variation has emphasized the need for single cell resolution and catapulted the development of single-cell sequencing-based approaches for measuring pairwise interactions (**Figure 2**). While these method variants have been incredibly useful for revealing the heterogeneous behavior of single regulatory factors, single cell-based approaches do not solve the underlying challenge of linking regulatory factors to each other in a context-specific manner. Single cell technologies still do not allow assessment of multiple factors on the same cell, and different regulatory factors are still measured using different populations of cells. At any point in time, regulatory factor A may be present in a subset of cell population A while regulatory factor B may be present in a subset of cell population B. Comparing regulatory factors A and B is done using the ensemble model of regulatory factor A in cell population A versus the ensemble model of regulatory factor B in cell population B. The direct relationship (e.g. co-occurrence frequency) between A and B remains unclear.

In summary, because of measurement limitations and a reliance on pairwise measurements, regulatory grammar remains challenging to study. The pairwise interactome catalogue for regulatory factors, even at the single-cell level, has been insufficient to understand the regulatory complexity underlying human health and disease.

Here, we highlight several well-known regulatory themes that involve spatial and/or temporal dynamics to demonstrate the difficulties presented by existing approaches.

1. **Multi-step Sequence (A→B→C)**

One relatively common circuit in gene regulation involves a sequential progression of events. For example, in transcribing a gene, RNA Polymerase II undergoes initiation, pausing, elongation, and finally termination.[18] When measured by ensemble ChIP-Seq, negative elongation factors responsible for pausing RNAP II such as DRB sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) co-localized with the polymerase peak genome wide. Separate in-vitro transcription experiments using crude nuclear extracts found that DSIF and NELF reduced the elongation rate of Pol II, functionally assigning them as negative factors.[19] Together, these independent observations were used to build an ensemble model for transcriptional pausing. The positional information from ChIP-seq was, on its own, insufficient to describe the link between these factors and their functional outcomes (e.g. transcriptional pausing) and construct the logic circuit of transcriptional steps.

2. **Feedback, Loops, and Oscillations (A↔B)**

Another simple circuit in gene regulation is feedback regulation or loops responsible for maintenance of steady state equilibria. For example, histone acetylation is modulated by two antagonizing classes of enzymes: histone acetyltransferases (HATs) which add acetyl groups, and histone deacetylases (HDACs) which remove them. It is well established that histone acetylation is marker of active transcription; HATs and HDACs have been found to be transcriptional co-activators and co-repressors, respectively. Perplexingly, upon genome-wide mapping, HATs and HDACs are both found at active genes marked with acetylated histones and positively correlated with transcription.[20] To explain this finding, the authors proposed a model in which, at active genes, the main function of HDACs is to ‘reset’ chromatin and, at inactive but poised genes, a dynamic cycle of transient HAT/HDAC binding maintains the poised state. Other studies have proposed a role for HDACs in regulating pause release based on responses to HDAC inhibitors.[21] However, without concurrent, real-time measurement of HDAC/HAT occupancy, histone acetylation, and transcriptional output, the molecular model controlling expression of these genes remains speculative.

3. Cis-regulatory Networks (B←A→C)

Long-range contacts between cis-regulatory elements and promoters are thought to regulate gene expression and facilitate co-regulation of multiple genes.[22] For example, promoter-capture Hi-C analysis has shown that certain enhancers contact multiple promoters, leading to the proposal that these enhancers form complex spatial networks connecting the regulation of these genes.[23] Complementarily, individual promoters have also been shown to contact multiple enhancers, leading to the proposal that multiple regulators co-regulate a single gene. Importantly, it remains unclear whether spatial proximity between enhancer and promoter is sufficient for functional activity.[24] Assignment of enhancers to promoters has also been attempted using computational methods instead of proximity methods, searching for target genes near enhancers that share chromatin state or accessibility.[22,25,26] However, this alternative strategy has struggled when genes are regulated by different tissue-specific enhancers.[24] In summary, because a direct link between a promoter-enhancer interaction and transcriptional output has been challenging to measure, much remains unclear. Other related questions such as how many of promoter-enhancer interactions occur simultaneously, what regulatory proteins are present during these interactions, and what is the contribution of each enhancer-promoter interaction on transcriptional output also remain unanswered.

4. High Affinity Macromolecular Complexes (A+B+C)

Regulatory proteins (e.g., chromatin regulators, RNA polymerase) commonly assemble into high affinity, macromolecular complexes. These complexes are frequently modular and combinatorial, with the same component associated with various forms. These forms may be cell-type specific but there are also known instances were heterogeneity of a complex leads to distinct localization and functionality within the same cell. For example, the BAF regulator complex can include up to 15 subunits, many of which are encoded by gene families, leading to hundreds of possible predicted assemblies.[27] Three of these (esBAF, npBAF, nBAF) have been explored for their tissue-specific functions, but simultaneously occurring non-canonical assemblies have also been observed. For instance,

in mouse embryonic stem cells, Glioma tumor suppressor candidate region gene 1 (GLTSCR1) or its paralog GLTSCR1-like (GLTSCR1L) and Bromodomain-containing protein 9 (BRD9) define a non-canonical BAF complex, which localizes to genomic features distinct from those targeted by canonical ESC BAF.[28] Another example, the nucleosome remodeling and deacetylase activities (NuRD) complex, contains the CHD proteins, CHD3 and CHD4, which are co-expressed in many cell lines and localize to distinct regions of the genome. In-vitro assays of CHD3-NuRD and CHD4-NuRD demonstrate differences in remodeling behavior.[29] However, the in-situ actions of heterogenous, combinatorial complexes cannot be studied easily and as a result much about their in-context biological functions remains unclear.

5. **Biomolecular Condensates (A+A+A)**

In contrast to high-affinity macromolecular complexes, other key molecular assemblies in gene regulation form through concentration-dependent, multivalent, cooperative associations.[30,31] These biomolecular condensates are characterized by spatial enrichment relative to the cellular surroundings and can have variable stoichiometries. For example, imaging studies have found that RNA Polymerase II and Mediator form clusters within the nucleus (referred to as transcriptional condensates), and the size of these clusters correlates with the level of nascent transcription.[32] Many individual nuclear proteins have been shown to undergo condensate formation through liquid-liquid phase separation; these include chromatin regulators (e.g. HP1)[33], transcription factors (e.g. OCT4)[34] and other RNA processing factors (e.g. SRSF1)[35]. While imaging-based methodologies have potentiated the study of condensates, they are undetectable by pairwise-interaction sequencing techniques.

6. **Nuclear Compartments**

Recently, there has been growing appreciation for the role of nuclear compartmentalization in multiple nuclear processes, including transcription regulation, co-transcriptional and post-transcriptional RNA processing, and higher-order chromatin regulation.[30] For

example, compartmentalization is thought to be important for spatially organizing enhancers, promoters, and transcription factors to drive transcription initiation. In addition, several nuclear structures have been shown to form membrane-less compartments (e.g., nucleolus). Such higher order spatial organizations are impossible to study using traditional pairwise measurements.

The examples above highlight the challenges associated with studying gene regulation principles using traditional genomics/epigenomics methods. Additionally, countless observations have been made using these approaches whose functional significance and mechanism remain unclear. Without a direct link between the observed phenomena, regulatory context, and transcriptional output, the biological relevance is impossible to decipher. Finally, there likely exist many novel mechanisms of gene regulation, which have yet not been uncovered and may be impossible to uncover using the current strategies.

1.4. THE SOLUTION: EVERYTHING, EVERYWHERE, ALL AT ONCE

While it is theoretically possible to study multiple factors simultaneously by enumerating their combinations, this strategy is limited by the poor scalability of combinatorics (**Figure 3**). The number of possible combinations quickly becomes unmanageable, and more importantly, only a small subset of this combinatorial space is biologically relevant. For example, as described for the BAF chromatin regulator complex, while there are hundreds of theoretical combinations, only a handful of cell-type specific modules (e.g., esBAF, npBAF, nBAP) have been observed to exist.

To discover the rules of gene regulation, we require experimental strategies that can directly observe the in-situ complexity of gene regulation. We need a comprehensive "snap-shot", including the set of regulatory factors (e.g., transcription factors, chromatin regulators), the genomic context (e.g., 3D genomic structure, chromatin state) and the transcriptional output, of an individual locus at a specific time (**Figure 4**). Such "snapshots" are multi-component, multi-modal, spatial measurements that capture the

direct associations between regulatory factors and transcriptional output. By generating many of such "snapshots", we may be able to appreciate the heterogeneity and context-specificity of regulatory events, discover the common rule sets and build more comprehensive models.

Here, we present split pool recognition of interactions by tag extension (SPRITE)[36] as a potential enabling technology for learning the grammar of gene regulation. Specific methods where SPRITE has been used include DNA SPRITE (a method for genome-wide mapping of higher-order DNA interactions)[37], RNA-DNA SPRITE (a method for simultaneous measurement of multiway RNA-RNA, RNA-DNA, and DNA-DNA contacts)[38], and SPRITE-IP (a method for mapping the DNA contacts surrounding a protein of interest)[39]. SPRITE can resolve the in-situ, multi-way interactions (including both long and short distance) between molecules within the nucleus. The output of SPRITE is a collection of measurements that identify 'clusters', each of which represents a multi-component, single time point, single cell event (**Figure 4**). While current applications of SPRITE have focused on simultaneous measurement of nucleic acids (e.g., RNA and/or DNA) and cannot simultaneously capture all critical element of gene regulation (e.g., multiple RNAs, DNAs and proteins), future adaptations could be developed to include the missing elements.

To highlight the potential of SPRITE-based measurements, here we describe examples where D-SPRITE, RD-SPRITE, or SPRITE-IP have been used to study the regulatory structures that are undetectable using other sequencing-based approaches:

1. **Higher Order Spatial Compartments: Interchromosomal DNA, RNA and Enhancer-Promoter Hubs**

Utilizing D-SPRITE, Quinodoz et al. (2018) uncovered that higher-order interchromosomal hubs shape 3D genome organization within the nucleus.[37] Specifically,

the authors define two hubs of interchromosomal DNA-DNA interactions that are arranged around the nucleolus and the nuclear speckle.

Utilizing RD-SPRITE, we described multiple higher-order RNA-chromatin structures that are involved in diverse classes of nuclear functions, including RNA processing, heterochromatin assembly and gene transcription.[38] Specifically, we defined hubs (e.g., nucleolar hub, centromeric hub, spliceosomal hub, scaRNA hub, histone locus body hub) using RNA-RNA interactions. Coupling RD-SPRITE with polymerase inhibition treatment, we then investigated the mechanisms behind formation of these RNA-mediated nuclear compartments which are involved in essential nuclear functions and gene regulation.

Utilizing SPRITE-IP, Vangala et al. (2020) characterized multi-way enhancer-promoter (E-P) interactions.[39] The authors found that E-P interactions can form transcriptional hubs involving multiple genes and that the stability of E-P hub predicts the stability of gene expression across a cell population.

2. 3D Genome Structure in Context

Capitalizing on the ability of RD-SPRITE to simultaneously measure DNA-DNA interactions and non-coding RNAs, we explored the 3D genome structures surrounding RNA-mediated nuclear bodies.[38] Specifically, we showed the genomic structures associated with active rRNA transcription and processing (nucleolar hub), active snRNA transcription and processing (scaRNA hub), or active histone pre-mRNA transcription and processing (histone locus hub). In each of these cases, we focused our analyses on clusters containing the relevant set of RNAs (pre-rRNA/snoRNAs, snRNA/scaRNA, and U7 snRNA, respectively) and mapped the concomitant DNA structure. We visualized complex genomic structures, such as long-distance DNA-DNA loops and interchromosomal associations, present at these RNA-mediated nuclear bodies.

Capitalizing on the ability of RD-SPRITE to simultaneously measure DNA-DNA interactions and nascent RNAs, we explored the functional link between 3D genome

structure and transcription.[40] We demonstrated that active transcription occurs within genomic structures that have previously thought to be inactive, including B compartments and DNA regions near the nucleolus, and argue against structural domains that preclude transcription.

In addition to these published examples, existing SPRITE-based data could be used to explore other aspects of regulatory grammar. For example, because each SPRITE cluster represents a distinct in-situ observation, examining SPRITE-based data could profile the heterogeneity of multi-way interactions. Alternatively, because SPRITE clusters can contain multiple copies of a particular molecule and cluster size (e.g., the number of molecules within a cluster) corresponds to volume, SPRITE-based data may be used to explore local concentrations within the nucleus.

1.5. THESIS CONTENTS

The central goal of my graduate work has been the development, application, and analysis of new genomics methods to study the fundamental principles of gene regulation genome-wide and to explore the interplay of regulatory proteins, chromatin state, genome structure, and transcription within the mammalian nucleus.

In Chapter 2, I describe a newly developed protocol, ChIP-DIP, for highly multiplexed mapping of hundreds of regulatory proteins to genomic DNA in a single experiment. ChIP-DIP increases the throughput of existing methods by > 100-fold and enables the production of consortium-scale data by a single lab. By dramatically increasing scale, ChIP-DIP facilitates the rapid characterization of hundreds of individual regulatory proteins within any experimental system of interest and enables a fundamental shift from consortium-generated cell-line reference maps to cell-type and cell-state specific maps. In addition, ChIP-DIP enables the rapid screening of protein affinity reagents, which are essential for studying regulatory proteins. Given the context-dependent nature of gene expression, comprehensive understanding of regulatory proteins within specific contexts is necessary

for learning the rules of gene regulation. Methodologically, ChIP-DIP employs SPRITE for multiplexing; instead of split-and-pool barcoding to resolve the in-situ interactions between molecules within the nucleus, ChIP-DIP uses split-and-pool barcoding to link together immunoprecipitated chromatin and a synthetic identifier sequence specific to the antibody used for immunoprecipitation. Capitalizing on the scale and diversity provided by ChIP-DIP, we uncover unique quantitative combinations of histone modifications that define distinctive classes of regulatory elements. Specifically, we find features distinguishing classes of promoters that correspond to different polymerase activity, transcriptional levels, and gene types and find acetylation patterns distinguishing classes of enhancers that exhibit distinct activity states, induction potential, and regulatory potential.

In Chapter 3, I describe the application of RNA-DNA SPRITE to investigate the functional relationship between genome structural organization and transcription and, specifically, argue against nuclear compartments that preclude transcription. Certain nuclear structures, such as B compartments or the nucleolus, have been associated with transcriptional inactivity; whether these structures are truly impermissive to transcription or simply correlated with inactivity has remained unknown due to a lack of methods that simultaneously measure genome structure and transcription. We demonstrate that RNA-DNA-SPRITE, a method for simultaneous measurement of RNA and DNA organization, precisely detects individual, nascent pre-mRNAs at their transcriptional locus and, as a result, can be used to assess the 3D genomic structure present during active transcription. We find that genes located in B compartments as well as genes located in proximity to the nucleolus can be actively transcribed, and we argue against a mechanistic model requiring drastic changes in genome organization (e.g., “looping out”) for active transcription of these genes. In addition, we measure the genome-wide organization of nascent pre-mRNAs for the first time and uncover structures reminiscent of DNA organization, such as chromosomal territories and A/B compartments. Our results highlight the power of RD-SPRITE to answer outstanding questions in gene regulation because it detects higher order

(e.g., long-distance and multiway) interactions involving nascent RNAs and concurrently measures transcriptional output and 3D genome organization.

Finally, in Chapter 4, I describe the application of RNA-DNA SPRITE to investigate the mechanistic dependence of genome structural organization on transcription and, specifically, demonstrate that transcriptional inhibition results in the loss of high-order genomic structure at RNA-mediated nuclear bodies.

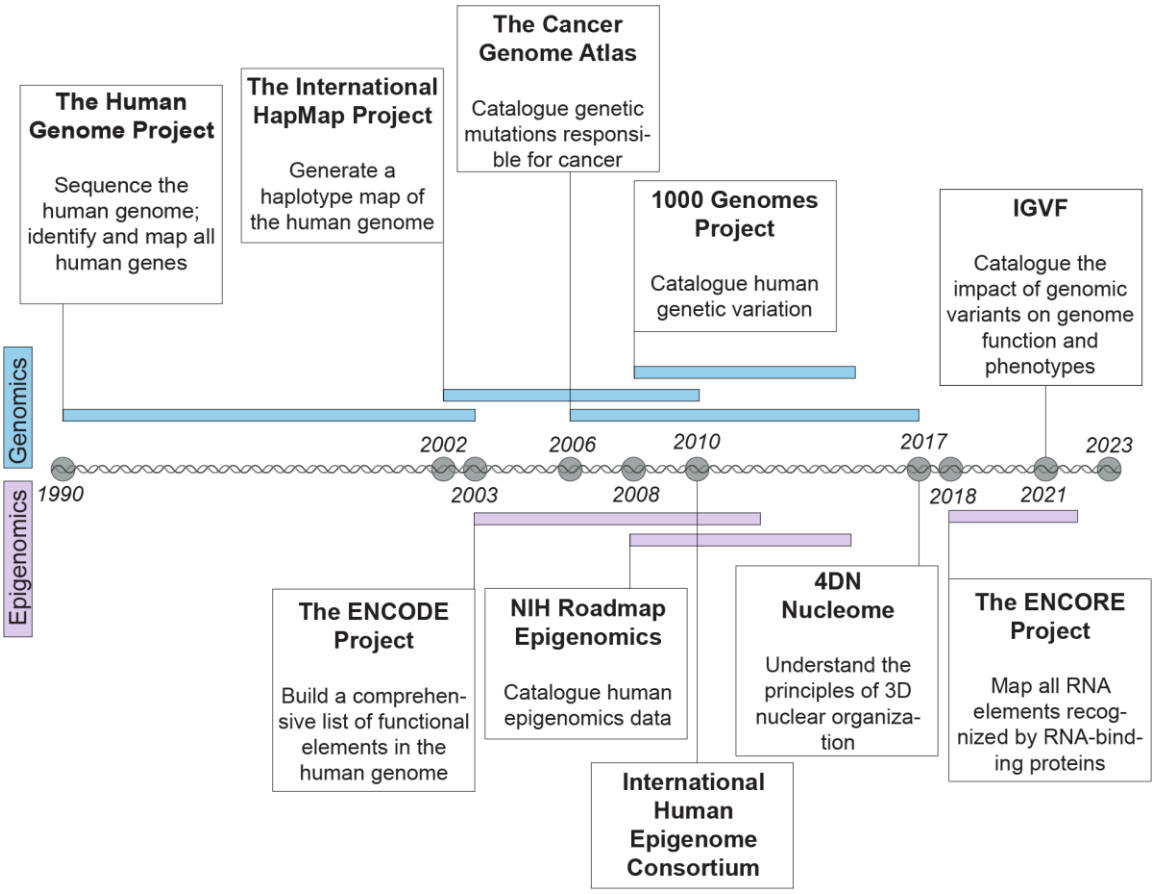


Figure 1: Timeline of large-scale national and international genomics and epigenomics research initiatives.

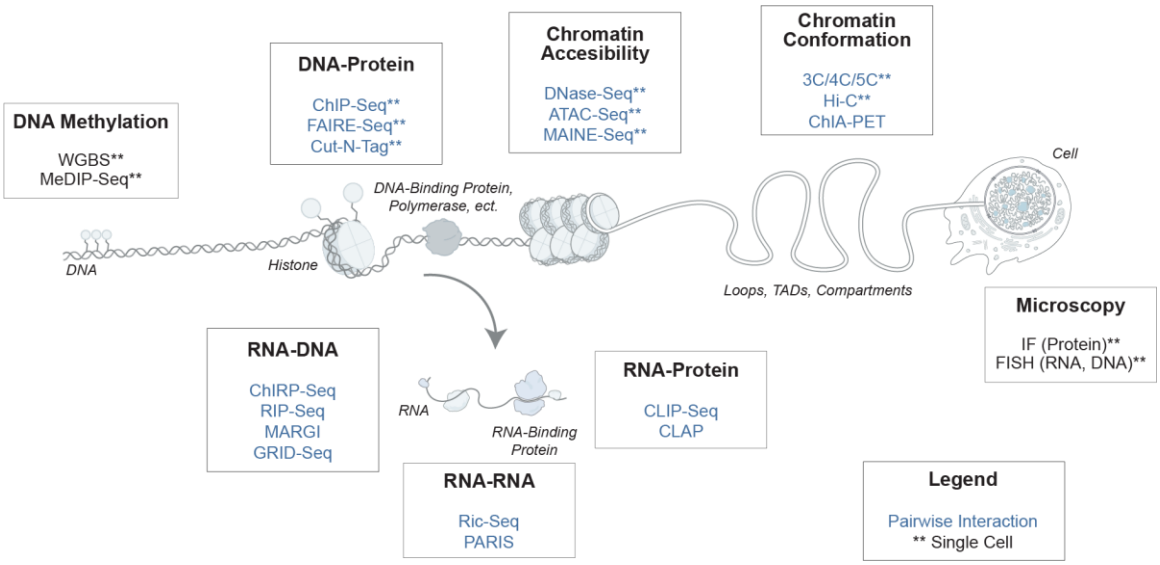


Figure 2: Methods to study gene regulatory factors.

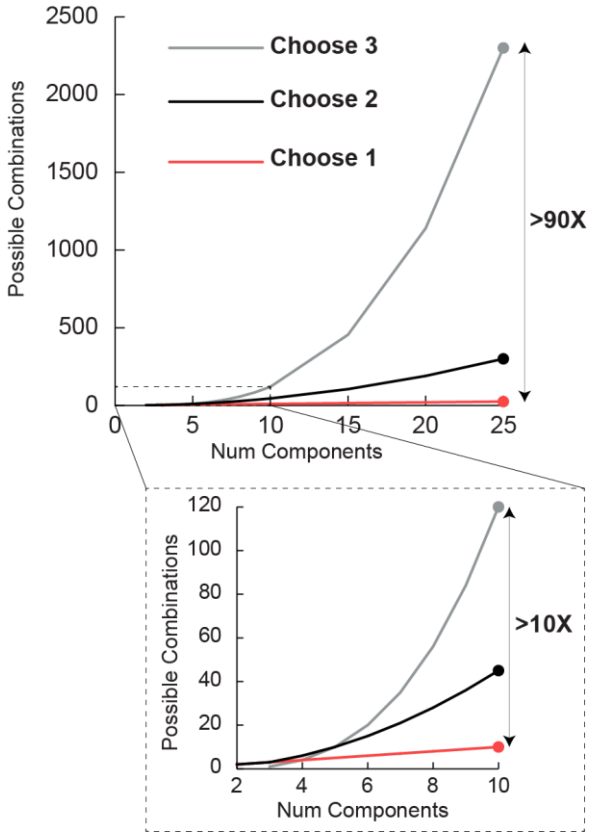


Figure 3: Scaling comparison of combinatoric sampling.

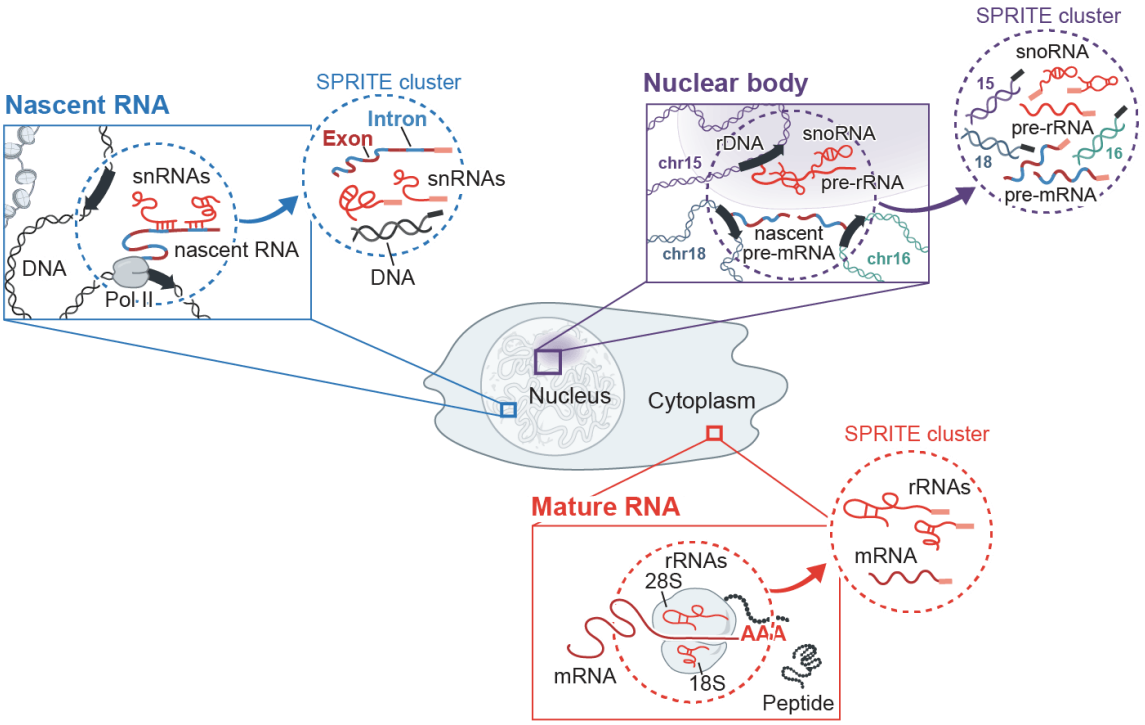


Figure 4: Schematic of in-situ molecular interactions and corresponding SPRITE clusters.

1.6. REFERENCES

1. Consortium, I.H.G.S. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945. 10.1038/nature03001.

2. Lander, E.S. (2011). Initial impact of the sequencing of the human genome. Nature *470*, 187–197. 10.1038/nature09792.

3. Consortium, I.H.G.S., Research:, W.I. for B.R., Center for Genome, Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. 10.1038/35057062.

4. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

5. Abascal, F., Acosta, R., Addleman, N.J., Adrian, J., Afzal, V., Aken, B., Ai, R., Akiyama, J.A., Jammal, O.A., Amrhein, H., et al. (2020). Perspectives on ENCODE. Nature *583*, 693–698. 10.1038/s41586-020-2449-8.

6. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227. 10.1038/nature07672.

7. Guo, J.K., and Guttman, M. (2022). Regulatory non-coding RNAs: everything is possible, but what is important? Nat. Methods *19*, 1156–1159. 10.1038/s41592-022-01629-6.

8. Lorenzi, L., Chiu, H.-S., Cobos, F.A., Gross, S., Volders, P.-J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., Lefever, S., et al. (2021). The RNA Atlas expands the catalog of human non-coding RNAs. Nat. Biotechnol. *39*, 1453–1465. 10.1038/s41587-021-00936-1.

9. 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., et al. (2017). The 4D nucleome project. Nature *549*, 219–226. 10.1038/nature23884.

10. Mehrmohamadi, M., Sepehri, M.H., Nazer, N., and Norouzi, M.R. (2021). A Comparative Overview of Epigenomic Profiling Methods. Front. Cell Dev. Biol. *9*, 714687. 10.3389/fcell.2021.714687.

11. Minnoye, L., Marinov, G.K., Krausgruber, T., Pan, L., Marand, A.P., Secchia, S., Greenleaf, W.J., Furlong, E.E.M., Zhao, K., Schmitz, R.J., et al. (2021). Chromatin accessibility profiling methods. Nat. Rev. Methods Prim. *1*, 10. 10.1038/s43586-020-00008-9.

12. Ramanathan, M., Porter, D.F., and Khavari, P.A. (2019). Methods to study RNA–protein interactions. Nat. Methods *16*, 225–234. 10.1038/s41592-019-0330-1.

13. Guh, C.-Y., Hsieh, Y.-H., and Chu, H.-P. (2020). Functions and properties of nuclear lncRNAs—from systematically mapping the interactomes of lncRNAs. J. Biomed. Sci. *27*, 44. 10.1186/s12929-020-00640-3.

14. Cao, C., Cai, Z., Ye, R., Su, R., Hu, N., Zhao, H., and Xue, Y. (2021). Global in situ profiling of RNA-RNA spatial interactions with RIC-seq. Nat. Protoc. *16*, 2916–2946. 10.1038/s41596-021-00524-2.

15. Nguyen, T.C., Zaleta-Rivera, K., Huang, X., Dai, X., and Zhong, S. (2018). RNA, Action through Interactions. Trends Genet. *34*, 867–882. 10.1016/j.tig.2018.08.001.

16. Khelifi, G., and Hussein, S.M.I. (2020). A New View of Genome Organization Through RNA Directed Interactions. Front. Cell Dev. Biol. *8*, 517. 10.3389/fcell.2020.00517.

17. Jerkovic´, I., and Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. Nat. Rev. Mol. Cell Biol. *22*, 511–528. 10.1038/s41580-021-00362-w.

18. Cramer, P. (2019). Organization and regulation of gene transcription. Nature *573*, 45–54. 10.1038/s41586-019-1517-4.

19. Zhou, Q., Li, T., and Price, D.H. (2012). RNA Polymerase II Elongation Control. Biochemistry *81*, 119–143. 10.1146/annurev-biochem-052610-095910.

20. Wang, Z., Zang, C., Cui, K., Schones, D.E., Barski, A., Peng, W., and Zhao, K. (2009). Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes. Cell *138*, 1019–1031. 10.1016/j.cell.2009.06.049.

21. Vaid, R., Wen, J., and Mannervik, M. (2020). Release of promoter–proximal paused Pol II in response to histone deacetylase inhibition. Nucleic Acids Res. *48*, 4877–4890. 10.1093/nar/gkaa234.

22. Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120. 10.1038/nature11243.

23. Novo, C.L., Javierre, B.-M., Cairns, J., Segonds-Pichon, A., Wingett, S.W., Freire-Pritchett, P., Furlan-Magaril, M., Schoenfelder, S., Fraser, P., and Rugg-Gunn, P.J. (2018). Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition. Cell Rep. *22*, 2615–2627. 10.1016/j.celrep.2018.02.040.

24. Rivera, C., and Ren, B. (2013). Mapping Human Epigenomes. Cell.

25. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82. 10.1038/nature11232.

26. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49. 10.1038/nature09906.

27. Alfert, A., Moreno, N., and Kerl, K. (2019). The BAF complex in development and disease. Epigenetics Chromatin *12*, 19. 10.1186/s13072-019-0264-y.

28. Gatchalian, J., Malik, S., Ho, J., Lee, D.-S., Kelso, T.W.R., Shokhirev, M.N., Dixon, J.R., and Hargreaves, D.C. (2018). A non-canonical BRD9-containing BAF chromatin remodeling complex regulates naive pluripotency in mouse embryonic stem cells. Nat. Commun. *9*, 5139. 10.1038/s41467-018-07528-9.

29. Hoffmeister, H., Fuchs, A., Erdel, F., Pinz, S., Gröbner-Ferreira, R., Bruckmann, A., Deutzmann, R., Schwartz, U., Maldonado, R., Huber, C., et al. (2017). CHD3 and CHD4 form distinct NuRD complexes with different yet overlapping functionality. Nucleic Acids Res. *45*, gkx711-. 10.1093/nar/gkx711.

30. Bhat, P., Honson, D., and Guttman, M. (2021). Nuclear compartmentalization as a mechanism of quantitative control of gene expression. Nat. Rev. Mol. Cell Biol. *22*, 653–670. 10.1038/s41580-021-00387-1.

31. Sabari, B.R., Dall'Agnese, A., and Young, R.A. (2020). Biomolecular Condensates in the Nucleus. Trends Biochem. Sci. *45*, 961–977. 10.1016/j.tibs.2020.06.007.

32. Cho, W.-K., Jayanth, N., English, B.P., Inoue, T., Andrews, J.O., Conway, W., Grimm, J.B., Spille, J.-H., Lavis, L.D., Lionnet, T., et al. (2016). RNA Polymerase II cluster dynamics predict mRNA output in living cells. eLife *5*, e13617. 10.7554/elife.13617.

33. Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241–245. 10.1038/nature22989.

34. Boija, A., Klein, I.A., Sabari, B.R., Dall'Agnese, A., Coffey, E.L., Zamudio, A.V., Li, C.H., Shrinivas, K., Manteiga, J.C., Hannett, N.M., et al. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. Cell *175*, 1842-1855.e16. 10.1016/j.cell.2018.10.042.

35. Guo, Y.E., Manteiga, J.C., Henninger, J.E., Sabari, B.R., Dall'Agnese, A., Hannett, N.M., Spille, J.-H., Afeyan, L.K., Zamudio, A.V., Shrinivas, K., et al. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature *572*, 543–548. 10.1038/s41586-019-1464-0.

36. Quinodoz, S.A., Bhat, P., Chovanec, P., Jachowicz, J.W., Ollikainen, N., Detmar, E., Soehalim, E., and Guttman, M. (2022). SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. Nat. Protoc. *17*, 36–75. 10.1038/s41596-021-00633-y.

37. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744-757.e24. 10.1016/j.cell.2018.05.024.

38. 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., et al. (2021). RNA promotes the formation of spatial compartments in the nucleus. Cell *184*, 5775-5790.e30. 10.1016/j.cell.2021.10.014.

39. Vangala, P., Murphy, R., Quinodoz, S.A., Gellatly, K., McDonel, P., Guttman, M., and Garber, M. (2020). High-Resolution Mapping of Multiway Enhancer-Promoter Interactions Regulating Pathogen Detection. Mol. Cell *80*, 359-373.e8. 10.1016/j.molcel.2020.09.005.

40. Goronzy, I.N., Quinodoz, S.A., Jachowicz, J.W., Ollikainen, N., Bhat, P., and Guttman, M. (2022). Simultaneous mapping of 3D structure and nascent RNAs argues against nuclear compartments that preclude transcription. Cell Rep. *41*, 111730. 10.1016/j.celrep.2022.111730.

Chapter 2

CHIP-DIP: A MULTIPLEXED METHOD FOR MAPPING HUNDREDS OF PROTEINS TO DNA UNCOVERS DIVERSE REGULATORY ELEMENTS CONTROLLING GENE EXPRESSION

Andrew A. Perez*, Isabel N. Goronzy*, Mario R. Blanco, Jimmy K. Guo, and Mitchell Guttman

2.1. SUMMARY

Gene regulation is governed by the complex interplay between thousands of regulatory proteins and chromatin states; understanding how these dynamics give rise to precisely controlled, cell type-specific gene expression has been a central goal of molecular biology. Yet, addressing this goal remains challenging because current methods for mapping proteins to DNA are labor-intensive, resource-demanding, and limited to studying a single or a small number of proteins at a time. To overcome this, we developed ChIP-DIP (ChIP Done In Parallel), a novel split-pool based method that enables simultaneous, genome-wide mapping of hundreds of diverse regulatory proteins in a single experiment. We demonstrate that ChIP-DIP generates highly accurate maps equivalent to traditional approaches, with data quality unaffected by the number of distinct proteins or the composition of proteins measured within a single experiment. We show that, because of this multiplexed capability, ChIP-DIP enables generation of highly accurate maps using several orders of magnitude fewer cells per protein compared to traditional approaches (~30,000 fold), making it ideal for studying primary and rare cell populations. In addition, we show that ChIP-DIP can generate high-quality maps for all classes of DNA-associated proteins, including histone modifications, chromatin regulators, transcription factors, and RNA Polymerases. Using these data, we explore quantitative combinations of histone modifications and integrate these signatures with RNA Polymerase activity, chromatin regulatory protein binding, and transcription factor binding to define distinct classes of regulatory elements (e.g. distinct types of enhancer elements), their functional activity (e.g. transcriptional activity), and their regulatory potential (e.g. poised for activation upon stimulation or differentiation). Together, our results demonstrate that ChIP-DIP enables generation of consortium level data within a single lab and highlight the importance of this approach for studying mechanisms of gene regulation in a context and cell type-specific manner.

2.2. INTRODUCTION

Cell type-specific gene regulation is controlled by thousands of regulatory proteins which dynamically localize at precise DNA regions within distinct chromatin states[1]. Chromatin states are comprised of distinct post-translational modifications on histone proteins[2] and critical for controlling which genomic regions can be bound by transcription factors[3] and the localization of chromatin regulators[4]. Consistent with this important role for chromatin state in gene regulation, distinct histone modifications have been shown to demarcate various functional elements (promoters, enhancers, transcribed regions, etc.) and are used to define their activity state (active, inactive, repressed) and regulatory potential (poised/primed for activation)[5].

Understanding how the interplay between chromatin state and regulatory protein binding gives rise to cell type-specific gene expression has been a central goal of molecular biology for many decades[4], yet much remains unknown. The key challenge is that the large number of distinct regulatory proteins and histone modifications involved makes it difficult to comprehensively map their locations[6,7]. Current methods to map the genome-wide binding of specific proteins rely on chromatin immunoprecipitation followed by sequencing (ChIP-Seq)[8] or more recently, antibody-directed transposase assays (e.g. CUT&Tag)[9,10]. While these approaches provide comprehensive maps of individual proteins, they are generally limited to mapping a single protein at a time.

To address this challenge and comprehensively explore regulatory protein binding and histone modification patterns, various international consortia have been formed to generate reference maps of hundreds of proteins within a small number of cell types (ENCODE[11], PsychENCODE[12], ImmGen[13], etc.). Because of the large numbers of cells required to map many proteins[11], consortium efforts have focused primarily on cell lines that can be easily grown in cell culture[7]. Although these efforts have provided many critical insights[14–16], because protein binding maps and gene expression programs are intrinsically cell type-specific[17,18], it is not possible to study cell type-specific regulation using maps generated from reference cell lines[19]. Generating additional cell type-specific regulatory maps

currently requires consortium-level effort (dozens of labs across the world), time (many years), and resources (>$100 million) for each biological system. Moreover, for many biological systems of interest (e.g., primary and/or rare cell populations, experimental perturbations, or patient-derived samples) obtaining the numbers of cells required to generate such maps is not feasible. To address this challenge and comprehensively explore regulatory protein binding and histone modification patterns, various international consortia have been formed to generate reference maps of hundreds of proteins within a small number of cell types (ENCODE[11], PsychENCODE[12], ImmGen[13], etc.). Because of the large numbers of cells required to map many proteins[11], consortium efforts have focused primarily on cell lines that can be easily grown in cell culture[7]. Although these efforts have provided many critical insights[14–16], because protein binding maps and gene expression programs are intrinsically cell type-specific[17,18], it is not possible to study cell type-specific regulation using maps generated from reference cell lines[19]. Generating additional cell type-specific regulatory maps currently requires consortium-level effort (dozens of labs across the world), time (many years), and resources (>$100 million) for each biological system. Moreover, for many biological systems of interest (e.g., primary and/or rare cell populations, experimental perturbations, or patient-derived samples) obtaining the numbers of cells required to generate such maps is not feasible.

To enable the generation of comprehensive, context-specific protein localization maps within any experimental system and in any molecular biology lab, we developed a method called ChIP-DIP (ChIP Done In Parallel). ChIP-DIP enables simultaneous, genome-wide mapping of hundreds of diverse regulatory proteins within a single experiment. Here, we show that ChIP-DIP generates highly accurate genome-wide maps, equivalent to those generated by traditional approaches, and that data quality is not impacted by the number or precise composition of distinct proteins mapped in a single experiment. Because ChIP-DIP can generate hundreds of protein maps from the same cell lysate, it enables the generation of highly accurate maps using several orders of magnitude fewer cells per protein (~30,000-fold) than consortium efforts. In addition, we show that ChIP-DIP enables accurate mapping of all classes of DNA-associated proteins, including histone modifications,

chromatin regulators, transcription factors and other sequence-specific DNA binding proteins, and RNA Polymerases. Using these data, we explore quantitative combinations of histone modifications and integrate these signatures with RNA Polymerase activity, chromatin regulatory protein binding and transcription factor binding to define distinct regulatory features in the genome, their functional activity, and their regulatory potential. Together, our results demonstrate that ChIP-DIP generates consortium level data within a single lab without the need for specialized equipment and highlights the importance of this approach for studying mechanisms of gene regulation in a context-specific manner.

2.3. RESULTS

ChIP-DIP: A highly multiplexed method for mapping DNA-associated proteins

To enable highly multiplexed, genome-wide mapping of hundreds of DNA-associated proteins in a single experiment, we developed ChIP-DIP (ChIP Done In Parallel) (**Figure 1A**). ChIP-DIP works by (i) coupling individual antibodies to beads that contain a unique oligonucleotide tag and combining sets of different antibody-bead-oligo conjugates to create an antibody-bead pool (**Figure S1A**), (ii) performing ChIP using this pool, (iii) conducting split-and-pool barcoding followed by DNA sequencing[20–22], and (iv) computationally matching split-pool barcodes between DNA and the oligonucleotide tag corresponding to a specific antibody. We refer to all unique reads containing the same split-pool barcode as a cluster. We combine DNA reads from all clusters corresponding to the same antibody to generate a protein localization map for each individual protein. The output of a ChIP-DIP experiment is analogous to the data generated in a traditional ChIP-Seq experiment, however instead of a single map, ChIP-DIP provides a set of distinct maps — one for each antibody utilized (**Figure 1B**).

To ensure that chromatin-antibody-bead-oligo conjugates remain intact throughout the ChIP-DIP procedure (rather than dissociating and reforming new complexes), we designed a series of experiments to measure dissociation between (i) the oligo and bead, (ii) antibody and bead, or (iii) the antibody and chromatin. We observed minimal dissociation for any

of these cases; most beads contain a single oligo type (>95%), beads without a coupled antibody are associated with minimal chromatin (<0.5%) and most chromatin originates from the initial capture (> 94%, **Figure S1B-E, Supplemental Note 1-2**).

ChIP-DIP accurately maps hundreds of diverse DNA-associated proteins

To test whether ChIP-DIP can accurately map genome-wide protein localization, we performed a ChIP-DIP experiment in human K562 cells using four well-studied target proteins: (1) the CTCF sequence-specific DNA binding protein that binds to insulator sequences[23], (2) the H3K4me3 histone modification that localizes at the promoters of active genes[24,25], (3) the RNA Polymerase II enzyme that transcribes RNA[26], and (4) the H3K27me3 histone modification that accumulates over broad genomic regions that are transcriptionally repressed[24,25]. We compared ChIP-DIP binding profiles to ChIP-Seq profiles previously generated by the ENCODE consortium and found that the localization patterns are comparable at specific genomic sites (**Figure 1B-C**) and highly correlated genome-wide (r=0.837-0.956, **Figure 1D**). These results establish that the data generated by ChIP-DIP are qualitatively and quantitatively comparable to data generated by ChIP-Seq.

We next sought to determine whether ChIP-DIP can generate accurate maps regardless of the antibody pool size (number of distinct antibodies) or composition (what other antibodies are contained within the pool). To test whether increasing pool size might increase background and decrease data quality, we mapped the localization of the same four proteins within four distinct panels containing different antibody numbers (10, 35, 50 or 52 antibodies per pool) (**Figure 2A**). The localization maps generated for each of these four proteins were highly comparable regardless of the pool size and matched those for each protein mapped individually using ChIP-Seq (**Figure 2B, C**). Next, we considered the possibility that the antibody composition of a pool might impact data quality; if multiple antibodies within the pool bind to the same protein or to distinct proteins that bind similar sites, we might observe a loss of signal at these overlapping sites. To test this, we included multiple independent antibodies targeting the same protein (CTCF) or multiple antibodies

that recognize distinct proteins within a complex (e.g., members of the PRC1/2 complex) in our antibody pools. Across pools, we observe highly consistent binding profiles regardless of the number of antibodies targeting a protein (**Figure 2D**), and, within a single pool, we successfully map multiple components of a complex (**Figure S2**). Together, these results indicate that neither pool size nor composition impact the quality of the data generated.

We next explored whether ChIP-DIP can generate accurate maps using limited amounts of cell lysate, an important requirement for studying many biological systems where it is challenging to obtain large numbers of cells. We performed ChIP-DIP using 35 different antibodies targeting 29 distinct protein epitopes across a ~1,000-fold decreasing range of input cell lysate (amounts equivalent to 45 million (~10^7) to 50 thousand (~10^4) human K562 cells) (**Figure 2E, Figure S3, S4, Supplemental Note 3**). To assess the quality of the maps produced, we focused on the four well-studied proteins described above and compared their binding patterns across the range of cell lysate amounts. We observed that the localization patterns remained highly similar for all four proteins as the amount of lysate decreased **(Figure 2F-H, Figure S4**). This suggests that ChIP-DIP can generate high quality protein-DNA interaction maps for multiple protein targets from input amounts equivalent to as few as 50,000 cells. Because ChIP-DIP can generate dozens of individual maps from a single preparation of lysate, this further reduces the effective number of cells required per protein target. In this example, we utilized 35 antibodies which correspond to ~1,400 cell equivalents for an individual protein target, an ~30,000-fold reduction relative to material amounts used in previous consortium efforts. In this way, we expect that ChIP-DIP will be a critical tool for generating comprehensive maps in rare cell populations.

ChIP-DIP maps histone modifications, chromatin regulators, transcription factors, and RNA polymerases

Gene regulation involves many different types of DNA-associated proteins including post-translationally modified histone proteins that are organized into nucleosomes (histone modifications)[27], the enzymes that read, write, and erase histone modifications (chromatin

regulators)[28], sequence specific DNA binding proteins (e.g., insulators and transcription factors)[29], and enzymes that transcribe DNA into RNA (RNA polymerases)[30]. Since some of these classes have been traditionally easier to map (e.g., histone modifications) than others (e.g. transcription factors), we explored whether ChIP-DIP can simultaneously map large numbers of proteins from distinct protein categories. To do this, we performed ChIP-DIP on >60 distinct proteins in human K562 cells and >160 distinct proteins in mouse embryonic stem cells (mESCs) across six experiments (**Figure S5**, **Supplemental Table 1**). These included 39 histone modifications (HMs), 67 chromatin regulators (CRs), 51 transcription factors (TFs), and all three RNA Polymerases (RNAPs) and four of their modified forms.

Histone modifications. Histone modifications have proven incredibly useful for annotating cell type-specific regulatory elements[31]. We mapped 39 histone modifications — including 18 acetylation, 17 methylation, 3 ubiquitination, and 1 phosphorylation marks — in either mESCs or K562s (**Figure 3A**). We confirmed the localization of five histone modifications commonly used to demarcate five functional chromatin states[5], as well as additional modifications associated with each state (**Figure S6A-F**): enhancer regions[32] (H3K4me1, H3K4me2, H3K27ac, **Figure 3B**), transcribed regions[24,33,34] (H3K36me3, H3K79me1/2, **Figure 3C**), promoter regions[24,25,35] (H3K4me3, H3K9ac, **Figure 3D**), polycomb-repressed regions[36] (H3K27me3, H2AK119ub, **Figure 3E**), and constitutive heterochromatin regions[37] (H3K9me3, H4K20me3, **Figure 3F)**. These data indicate that ChIP-DIP accurately maps histone modifications with distinct genome-wide patterns (broad and focal localization) that represent distinct activity states (active and repressive), and that localize at distinct functional elements (promoters, enhancers, gene bodies, and intergenic regions).

Chromatin regulators. Chromatin regulators (CRs) are responsible for reading, writing, and erasing specific histone modifications and are critical for the establishment, maintenance, and transition between chromatin states[38,39]. We measured 67 CRs associated with various histone methylation, acetylation, and ubiquitination marks, as well as with DNA methylation, in either mouse ES or human K562 cells (**Figure 3A**). As expected, we

observe that an eraser (JARID1A)[40] and a writer (RBBP5-containing complex)[41] of H3K4me3 localize at H3K4me3-modified promoter sites (**Fig3G, Figure S6G**). Additionally, we observed that components of the PRC1 (RING1B, CBX8)[42] and PRC2 complex (EED, SUZ12, EZH2)[43] co-localize and are enriched over genomic regions containing their respective histone modifications (H2AK119ub and H3K27me3, **Figure 3H, Figure S6H**). Similarly, we observed co-localization of two members of the Heterochromatin Protein 1 (HP1) family, HP1α and HP1ß, at genomic DNA regions containing their associated heterochromatin marks, H3K9me3 and H4K20me3[44] (**Figure 3I, Figure S6I**). These data indicate that ChIP-DIP accurately maps chromatin regulators from diverse complexes and with distinct functional properties (i.e., modification recognition, enzymatic activity, chromatin packaging).

Transcription factors. Transcription factors (TFs) bind *cis*-regulatory elements in combinatorial patterns to control gene expression. Generating comprehensive maps of TF localization has proven difficult because there are large numbers of distinct TFs, most are cell type-specific, and they are challenging to map by ChIP-Seq because they tend to be lower in abundance and only transiently associated with DNA[45,46]. To explore whether ChIP-DIP can map large sets of TFs, we measured 15 TFs in K562 and 43 TFs in mESC, including constitutive (e.g. SP1 and USF2)[47,48], stimulus-dependent (e.g. p53 and NRF1)[49–52], and developmental/cell type-specific (e.g., Nanog and RFX1)[53,54] DNA binding proteins[55] **(Figure 4A)**. We obtained high-resolution binding maps for TFs in both cell types, with individual TFs localizing to their well-characterized targets at regions containing their known motifs[47,50,52,56–58] (**Figure 4A-B**). Using the genome-wide localization data, we can accurately identify the expected DNA binding motifs, including the 20bp dimer motif of p53[59] and the 21bp RE-1 consensus sequence of REST[60] (**Figure 4C**). Together, these data indicate that ChIP-DIP generates accurate, high-resolution binding maps of diverse TFs in multiple cell types.

RNA Polymerases (RNAPs). Different types of RNA are transcribed by distinct RNA polymerases: RNA Polymerase I (RNAP I) transcribes ribosomal RNA; RNAP II transcribes messenger RNAs and various non-coding RNAs, including snRNAs, snoRNAs

and lncRNAs; and RNAP III transcribes many classes of small RNAs, including tRNAs, the U6 snRNA, and SINE elements[61]. We leveraged the power of ChIP-DIP to simultaneously map all three RNAPs and the post-translationally modified forms of RNAP II. We observed that each RNAP localizes with high selectivity to its corresponding classes of genes; RNAP I binds at rDNA, RNAP II at mRNA and snRNA genes, and RNAP III at tRNA genes (**Figure 5A, Figure S7A)**. Moreover, we observed distinct localization patterns of different RNAP II phosphorylation states: serine 5 phosphorylated RNAP II localizes at promoters, while serine 2 phosphorylated RNAP II accumulates over the gene body and past the 3' end of the gene (**Figure S7B-C**). These data indicate that ChIP-DIP accurately maps the localization of the three RNA polymerases — including multiple functional phosphorylation states of RNAP II — at distinct gene classes and gene features.

Together, these results establish ChIP-DIP as a modular, highly multiplexed method that generates high-quality maps for a wide range of DNA-associated proteins spanning diverse biological functions.

ChIP-DIP enables integrated analysis of proteins and identifies regulatory features, activity, and potential.

Because of the large number of distinct regulatory proteins, previous integrative analyses of multiple protein targets have been limited to datasets generated by consortium efforts in a small number of human cell lines[62]. These integrated analyses have identified unique regulatory states and have highlighted that combinations of multiple histone modifications can demarcate distinct genomic elements (e.g. promoters, enhancers, transcribed regions, etc.)[63], their activity state (active, inactive, repressed), and regulatory potential (poised/primed for activation)[64]. Despite the importance of these combinations, mapping large numbers of modifications is technically challenging. Accordingly, many efforts to profile chromatin states have focused on mapping only five histone modifications that demarcate specific features and regulatory states (i.e., H3K4me3, H3K4me1, H3K36me3, H3K9me3, and H3K27me3 marking promoters, enhancers, elongated transcripts, heterochromatin, and polycomb-mediated silencing, respectively)[5]. Because ChIP-DIP can

map large numbers of diverse proteins, it facilitates comprehensive profiling and integrative analyses of histone modifications and other regulatory proteins within each specific cell state. To explore this, we asked whether combinations of histone modifications can provide additional information about distinct types, activity states, and regulatory potentials of *cis*-regulatory elements (promoters or enhancers) beyond those captured by the five commonly studied individual histone modifications.

Promoter type and activity state are defined by combinations of histone modifications

H3K4me3 is generally thought to mark the promoters of actively transcribed RNAP II transcripts [24,25,65]. While we find H3K4me3 over the promoters of actively transcribed RNAP II genes, we also observe this modification near RNAP I promoters (ribosomal RNA) and many active RNAP III genes (tRNAs) (**Figure 5B-C**). Similarly, other histone modifications that are associated with active RNAP II promoters, including H3K4me2, H3K9Ac, H3K27Ac, and H3K56Ac, are also enriched at RNAP I, II, and III genes (**Figure 5B-C, Figure S7C-D**).

Although the presence of these histone modifications does not appear to distinguish between genes transcribed by different polymerases, we observed that both their position relative to the transcriptional start site (TSS) and their relative levels vary by polymerase: for RNAP I genes, these modifications localize prior to the TSS; for RNAP II, they flank the promoter and are enriched downstream of the TSS; and for RNAP III, they flank the gene body, localizing both upstream of the TSS and downstream of the transcriptional termination site **(Figure 5B-C, Figure S7C)**. In addition, the three RNAPs have different relative levels of these histone modifications near their respective gene promoters. Focusing on H3K4me3, H3K4me2, and H3K56Ac, we found that RNAP I and II have a stronger acetylation component and RNAP I and III have a stronger H3K4me2 component **(Figure 5D)**. In this way, both combinations of modifications and their position relative to the promoter define distinct transcriptional programs (**Figure 5E**).

Next, we considered whether other histone modifications may distinguish types and activity states of RNAP II promoters. To explore this, we quantified the levels of ten additional histone modifications at each genomic region containing H3K4me3 and grouped them using hierarchical clustering. We identified five sets of H3K4me3 enriched genomic regions; four are enriched with other histone modification (sets 1-4) and one is not (set 5). The four co-occurring sets correspond to H3K4me3 along with: H3K27me3/H2AK119ub (set 1), H3K36me3/H3K79me (set 2), H3K9me3/H4K20me3 (set 3), or H3K4me1/H3K27ac (set 4) (**Figure 6A**). These correspond to sets of promoters that exhibit distinct transcriptional activity (e.g., high versus low expression) (**Figure 6B, Figure S8)** and are enriched for distinct classes of RNAP II-transcribed genes, such as ribosomal protein and cell cycle genes (set 2), zinc finger protein (set 3), and long intergenic ncRNAs genes (sets 3 and 4) (**Figure 6C-D**). Beyond these, there are smaller subsets of promoters that display additional histone modification patterns that correspond to specific gene classes and transcriptional states (**Figure S8**). Consistent with the fact that promoters of functionally distinct genes have unique chromatin profiles, we found that different readers, writers, and erasers of H3K4me3 localize at distinct sets of K4me3 modified promoters (**Figure S6G**).

Taken together, these results demonstrate that combinations of histone modifications can distinguish promoter features including polymerase, gene type, and activity level (**Figures 5E, 6I**).

Enhancer type, activity and potential are defined by combinations of histone modifications

There are >40 different histone acetylation marks[66], many of which have been associated with enhancers and active transcription. We mapped 15 of these, including marks on all four core histones and histone variants, in mESCs, and observed that they co-localize at similar sites genome-wide (Pearson r = 0.86-0.97)[67] (**Figure S9**). We wondered whether these strong correlations indicate that these marks are redundant or whether there is additional regulatory information encoded by the relative levels of each acetylation mark

at specific genomic sites. To explore this, we used a matrix factorization algorithm to define five weighted combinations of acetylation marks at highly acetylated genomic regions (quantitative combinations C1-C5; see **Methods, Figure 7A-B, Supplemental Note 4, Figure S10)**. These quantitative combinations correspond to genomic regions that contain distinct transcription factor and chromatin regulator binding profiles **(Figure 7C-F, Figure S11)**.

Active promoter-proximal elements. The first group of regions (C1) is defined by H3K9Ac and several other H3 acetylation marks (H3K14ac, H3K18ac, H3K36ac, H3K56ac, and H3K79ac) (**Figure 7B)**. Genomic regions containing this signature tend to be localized near the promoter region of transcribed genes and are enriched for RNAP II, general TFs (e.g. TFIIB), and other CpG-island associated factors (e.g. E2F1, CXX1) along with their sequence motifs (e.g. ETS, SP and NRF families) **(Figure 7C,E-F, Figure S11)**.

Poised promoter-proximal elements. The second group of regions (C2) contains high levels of H3K9Ac and acetylation of the histone variant H2AZ (H2AZAc) (**Figure 7B**). Genomic regions containing this signature tend to have lower levels of RNAP II (relative to C1) and are strongly enriched for polycomb (JARID2, SUZ12, RING1B) and other repressive chromatin regulators (KDM2B, HDAC2) **(Figure 7E-F, Figure S11)**.

Stress and signaling response elements. The third group of regions (C3) contains high levels of H2AZAc and H4Ac (**Figure 7B)**. Genomic regions containing this signature are also enriched for RNAP II but are bound by p53 and contain other stress response motifs (e.g., BACH1, NRF2) or signaling response motifs (e.g. CRE) (**Figure 7C, Figure 7E-F, Figure S11**). Consistent with these observations, H2AZ has been proposed as a facilitator of inducible transcription (e.g. signaling pathway responses and p53 regulation)[68–71]. Yet, because H2AZ is also a component of C2, our results suggest that this behavior is not solely a property of H2AZAc but of this unique C3 signature.

Active pluripotency distal regulatory elements. The fourth group of regions (C4) is defined by H2BK20Ac and H3K27Ac (**Figure 7B)**. Genomic regions containing this signature tend

to be promoter-distal (**Figure S11B**) and associated with actively transcribed embryonic and stem cell specific genes (**Figure 7D)**. These regions are enriched for binding of the pluripotency TFs, including Nanog, Oct4, and Sox2, as well as the P300 acetyltransferase and components of mediator **(Figure 7F)**.

Poised differentiation distal regulatory elements. The fifth group of regions (C5) is defined by high-levels of H2BK20Ac (similar to C4) and H3K14Ac (distinct from C4) (**Figure 7B**). Interestingly, these regions displayed similar TF and CR occupancy (e.g. Oct4, Sox2, Nanog, P300 and mediator) to C4 regions **(Figure 7F-G)**. However, in contrast to C4 regions, which contain a high-density of pluripotency TFs corresponding to enhancers of active genes involved in embryo and stem cell function, high-density of these TFs in C5 regions is associated with enhancers of genes involved in post-embryonic development, particularly muscle structure development **(Figure S12)**. Consistent with this, C5 regions are enriched for the sequence motifs of TFs involved in lineage specification and cardiac muscle morphogenesis (e.g. TEAD family)[72] **(Figure 7D, Figure S11)**. Because the heart muscle is one of the earliest organ systems to develop[73], this suggests that C5 enhancers might be important in establishing the gene expression program needed upon differentiation (regulatory potential). Interestingly, we identified a third set of genomic regions that also contain a high-density of pluripotency TFs but lack the C4 or C5 acetylation signatures; these are associated with genes involved in later stages of organogenesis (e.g. kidney and sensory systems) (**Figure S12**).

These analyses indicate that histone acetylation is not a redundant marker of enhancers, but that combinations of acetylation modifications can define unique classes of *cis* regulatory elements (promoter-proximal versus distal enhancers) that act in distinct ways (stimulus-responsive versus developmentally regulated) and that exhibit different activity (e.g. active gene expression versus poised for activation upon differentiation) (**Figure 7H**).

Overall, these observations highlight the importance of multi-component analyses and demonstrate why ChIP-DIP provides a powerful approach that will be critical for defining unique regulatory features within distinct cell states.

2.4. DISCUSSION

We demonstrated that ChIP-DIP enables highly multiplexed mapping of hundreds of regulatory proteins to genomic DNA in a single experiment, increasing the throughput of existing methods by >100-fold. Although the largest experiment in this study contained hundreds of proteins in a single experiment, these numbers were primarily selected because of the availability of high-quality antibodies; we expect that ChIP-DIP could be used to map even larger numbers of proteins. Because this approach employs standard molecular biology techniques, we expect that it will be readily accessible to any lab without the need for specialized training or equipment. As such, we anticipate that ChIP-DIP will enable a fundamental shift from reference maps generated by large consortia for a limited number of cell types to cell-type-specific maps generated by individual labs within any specific experimental system of interest.

We used ChIP-DIP to identify combinatorial regulatory information from a large panel of histone modifications and other DNA-associated factors. While distinct regulatory states have long been proposed to be encoded through diverse combinations of histone modifications[4], the actual number of such states has remained largely unexplored. For example, genomic regions containing H3K4me3 (a mark of active promoters) and H3K27me3 (a polycomb-mediated repressive mark) have been shown to represent developmentally poised promoters[64], yet the diversity of marks co-occurring with H3K4me3 and their functional implications has not been well characterized. Moreover, the possibility that histone combinations at various regulatory elements might represent diverse regulatory states has not previously well explored. Capitalizing on the scale and diversity provided by ChIP-DIP, we identified unique quantitative combinations of histone modifications that define classes of promoters corresponding to different polymerase activity, transcriptional levels, and gene types as well as classes of enhancers that display distinct activity states, induction potential, and regulatory potential. Importantly, these regulatory elements are occupied by distinct chromatin regulators and transcription factors, suggesting that combinations of histone modifications may be deposited or recognized by

unique networks of regulatory proteins. Consistent with this observation, various chromatin regulators have been shown to recognize unique histone combinations (e.g. SWI/SNF)[74], suggesting that gene regulation involves coordinated interplay between regulatory proteins and diverse sets histone modifications.

Given the important information encoded within quantitative combinations of histone modifications, chromatin regulators, and transcription factors, comprehensively mapping these factors across cell-types will be critical for studying gene regulation. The limitations of previous methods combined with the large numbers of histone modifications and regulatory proteins has necessitated a tradeoff between mapping many marks in a few cell-types or a few marks in many cell-types. ChIP-DIP overcomes this by mapping hundreds of proteins in a single experiment. Moreover, due to the nature of split-pool barcoding used in ChIP-DIP and because there is negligible antibody-bead-chromatin dissociation during the procedure, ChIP-DIP can also be used to map protein binding within multiple samples simultaneously using distinct sets of antibody-oligo-labeled beads or multiplexing during split-pool barcoding steps. While we did not directly emphasize this capability in this paper, several ChIP-DIP experiments described here were performed simultaneously using multiple sample conditions (e.g., crosslinking conditions, IP conditions, cell lysate amounts). In this way, ChIP-DIP will enable large scale mapping of proteins across many experimental conditions or at multiple timepoints.

In addition to increased scale, ChIP-DIP also provides important technical advantages that enable the study of complex protein binding relationships within individual cell-types. Specifically, by mapping multiple proteins within a single sample of crosslinked and sonicated material, ChIP-DIP reduces many sources of technical and biological variability associated with processing individual proteins and samples individually, enabling direct comparison of positions and levels between proteins. The ability to measure regulatory proteins at scale, in multiple cell conditions and with reduced sources of variability is ideally suited for use-cases requiring mapping dynamic protein localization changes across time and, more generally, will enable the construction of large-scale models to comprehensively understand gene regulation.

Beyond the applications highlighted in this work, ChIP-DIP can be directly integrated into existing split-pool approaches to create additional capabilities that are not currently possible. For example, we previously showed that we can map the 3D genome structure surrounding individual protein binding sites (SIP)[75]; by integrating this approach with ChIP-DIP, we can map the 3D structures that occur at hundreds of distinct protein binding sites simultaneously. Moreover, we previously developed a method to map 3D genome contacts for thousands of individual single cells using this same split-pool approach[76]. This single cell approach can be directly integrated into ChIP-DIP to enable comprehensive mapping of hundreds of regulatory protein binding sites within thousands of individual cells. Finally, we previously showed that split-and-pool barcoding can be used to simultaneously map DNA and RNA and measure the levels of nascent RNA transcription at individual DNA sites[77]. Accordingly, this approach can be combined with ChIP-DIP to enable the direct measurements of protein binding and transcriptional activity at individual genomic locations, providing a directly link between binding events and the associated transcription levels within the same cell. For these reasons, we expect that ChIP-DIP will represent a transformative new tool for dissecting gene regulation.

2.5. MAIN FIGURES

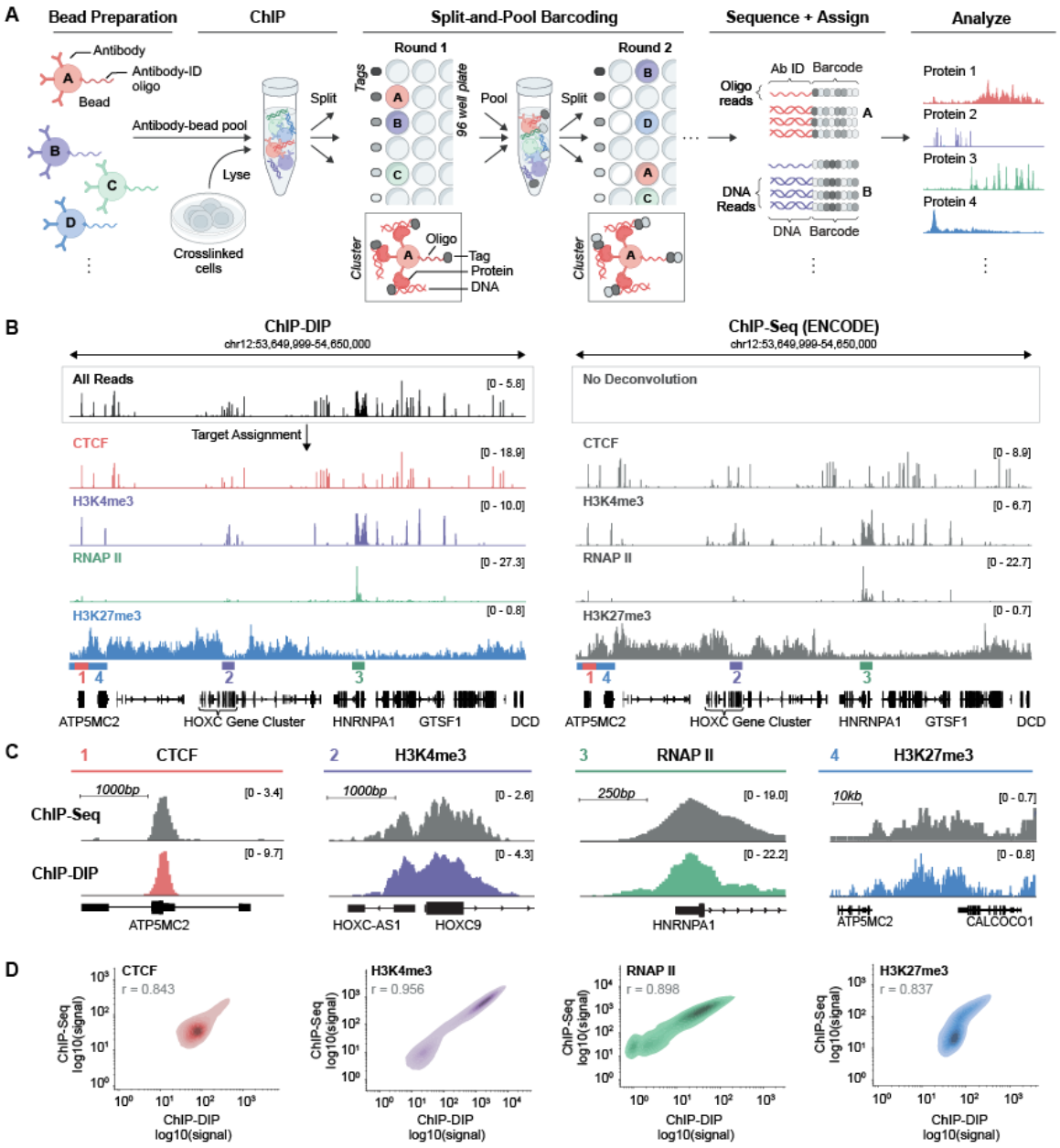


Figure 1: ChIP-DIP: A highly multiplexed method for mapping proteins to genomic DNA.

(A) Schematic of the ChIP-DIP method. Beads are coupled with an antibody and associated oligonucleotide (antibody-ID). Sets of beads are then mixed (antibody-bead pool, left) and

used to perform ChIP. Multiple rounds of split-and-pool barcoding are performed to identify molecules bound by the same Protein G bead (middle). DNA is sequenced and genomic DNA and antibody oligos containing the same split-and-pool barcode are grouped into a cluster, which are used to assign genomic DNA regions to their linked antibodies (right). All DNA reads corresponding to the same antibody are used to generate protein-localization maps. **(B)** Protein localization maps over a specific human genomic region (hg38, chr12:53,649,999-54,650,000) for four proteins targets: CTCF, H3K4me3, RNAP II and H3K27me3. Left panel: Protein localization generated by ChIP-DIP in K562. Top track shows read coverage prior to protein assignment and bottom tracks correspond to read coverage after assignment to individual proteins. Right panel: ChIP-Seq data generated by ENCODE within K562 for these same 4 proteins are shown for the same region. To enable direct comparison of scales between datasets, we normalized the scale to coverage per million aligned reads. Scale is shown from 0 to maximum coverage within each region. **(C)** Comparison of ChIP-DIP and ChIP-Seq maps over specific regions corresponding to zoom-ins of the larger region shown in (B). The locations presented are demarcated by colored bars at the bottom of (B). Scale shown similar to (B) **(D)** Genome-wide comparison (density plots of signal correlation) between the localization of each individual protein measured by ChIP-DIP (x-axis) or ChIP-Seq (y-axis). Points are measured genome-wide across 10kb windows (CTCF, H3K27me3) or all promoter intervals (H3K4me3, RNAP II).

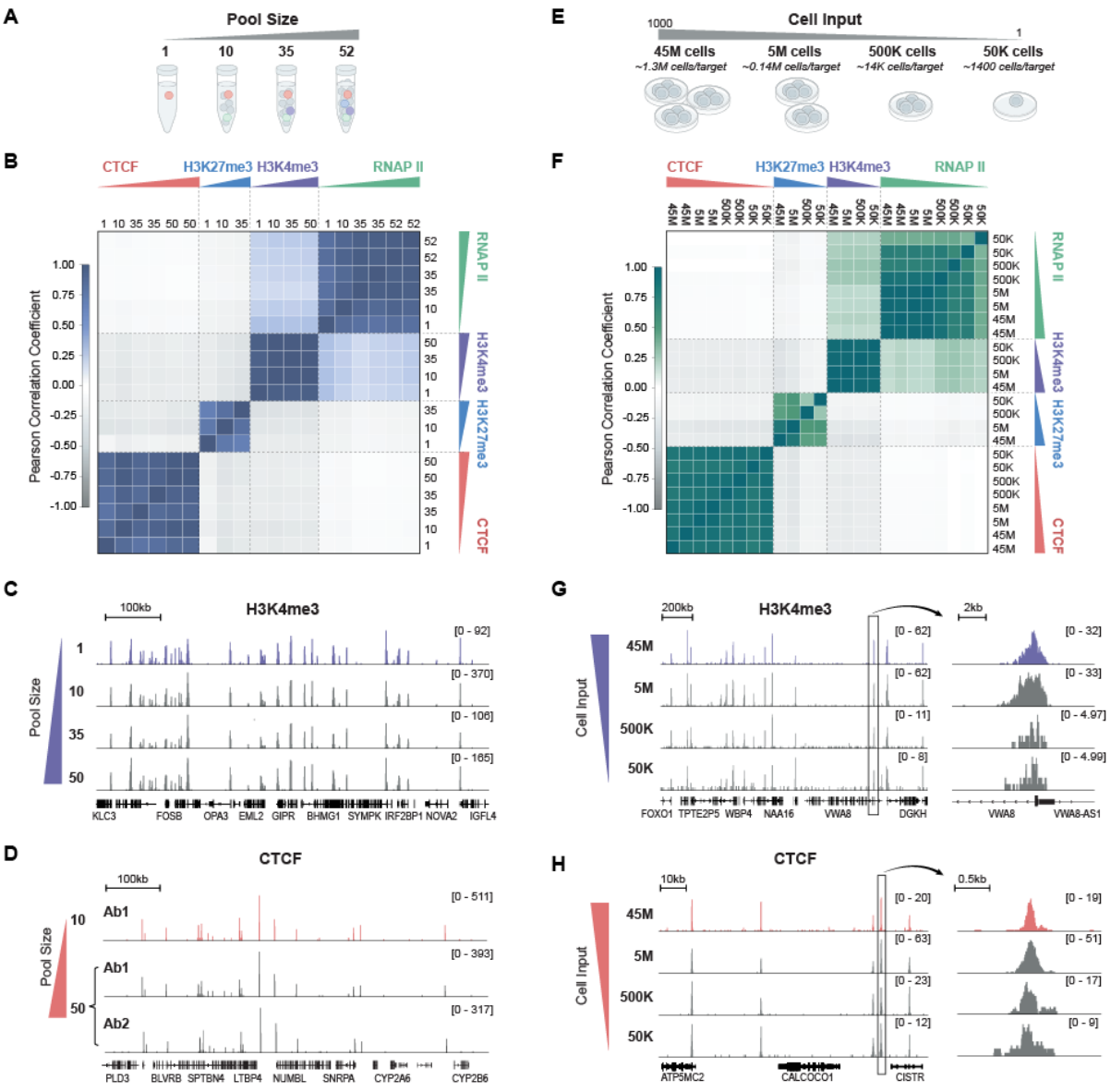


Figure 2: ChIP-DIP accurately maps large sets of proteins using low-levels of cell lysate.

(A) Schematic of experimental design to test scalability of antibody-bead pool size and composition. **(B)** Correlation heatmap for protein localization maps of four proteins — CTCF, H3K4me3, RNAP II and H3K27me3 — generated using antibody pools of four different sizes and compositions (see **Methods**). Pool sizes are listed along top and left axis. **(C)** Comparison of H3K4me3 localization over a specific genomic region (hg38,

chr19:45,345,500-46,045,500) when measured within various antibody pool sizes and compositions. **(D)** Comparison of CTCF localization over a specific genomic region (hg38, chr19:40,349,999-41,050,000) when measured within a pool containing a single CTCF-targeting antibody (top) or multiple CTCF-targeting antibodies in the same antibody pool (bottom). **(E)** Schematic of experimental design to test the amount of cell input required for ChIP-DIP. **(F)** Correlation heatmap for protein localization maps of four targets — CTCF, H3K4me3, RNAP II and H3K27me3 — generated using various amounts of input cell lysate (see **Methods**). Amounts of input cell lysate are listed along top and left axis. **(G)** Comparison of H3K4me3 localization over a specific genomic region (hg38, chr13:40,600,000-42,300,000) when measured using various amounts of input cell lysate. **(H)** Comparison of CTCF localization over a specific genomic region (hg38, chr12:53,664,000-53,764,000) when measured using various amounts of input cell lysate.

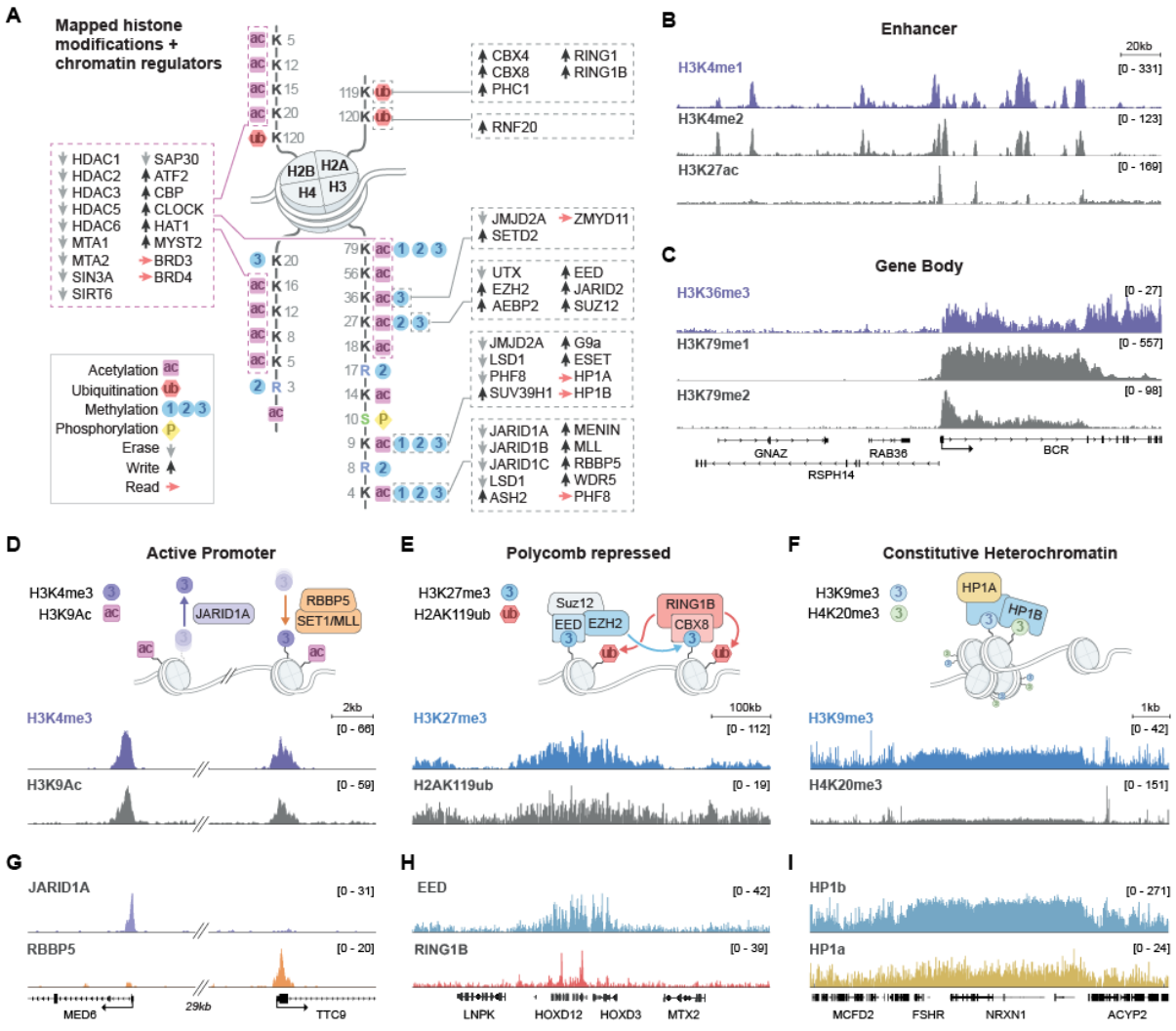


Figure 3: ChIP-DIP accurately maps dozens of functionally diverse histone modifications and chromatin regulators.

(A) Illustration of the diverse histone modifications and chromatin regulatory proteins mapped in K562 or mESC using ChIP-DIP. **(B-C)** Visualization of multiple histone modifications across a genomic region (hg38, chr22:23,050,000-23,290,000) in K562 corresponding to multiple histone modifications associated with **(B)** enhancers — H3K4me1, H3K4me2 and H3K27Ac and **(C)** active gene bodies — H3K36me3, H3K79me1, and H3K79me2. **(D)** Top: Schematic of histone modifications and chromatin regulators associated with active promoters. Bottom: Visualization of multiple histone

modifications associated with active promoters — H3K4me3 and H3K9Ac — across a genomic region (mm10, chr12:81,590,000-81,636,000) in mouse ESCs. Hashmarks indicate an intervening 29kb region that is not shown. **(E)** Top: Schematic of histone modifications and chromatin regulators associated with polycomb-mediated repression. Bottom: Visualization of multiple histone modifications associated with polycomb-mediated repression — H3K27me3 and H2A119ub — across a genomic region (hg38, chr2:175,846,000-176,446,000) containing the silenced HOXD cluster in K562. **(F)** Top: Schematic of histone modifications and chromatin regulators associated with constitutive heterochromatin. Bottom: Visualization of multiple histone modifications associated with constitutive heterochromatin — H3K9me3 and H4K20me3 — across a genomic region (hg38, chr2:46,200,000-55,700,000) in K562. **(G)** Visualization of an H3K4me3-associated eraser (JARID1A) and writer component (RBBP5) across the same genomic region as (D). **(I)** Visualization of PRC2 (EED) and PRC1 (RING1B) components across the same genomic region as (E). **(J)** Visualization of HP1b and HP1a across the same genomic region as (F).

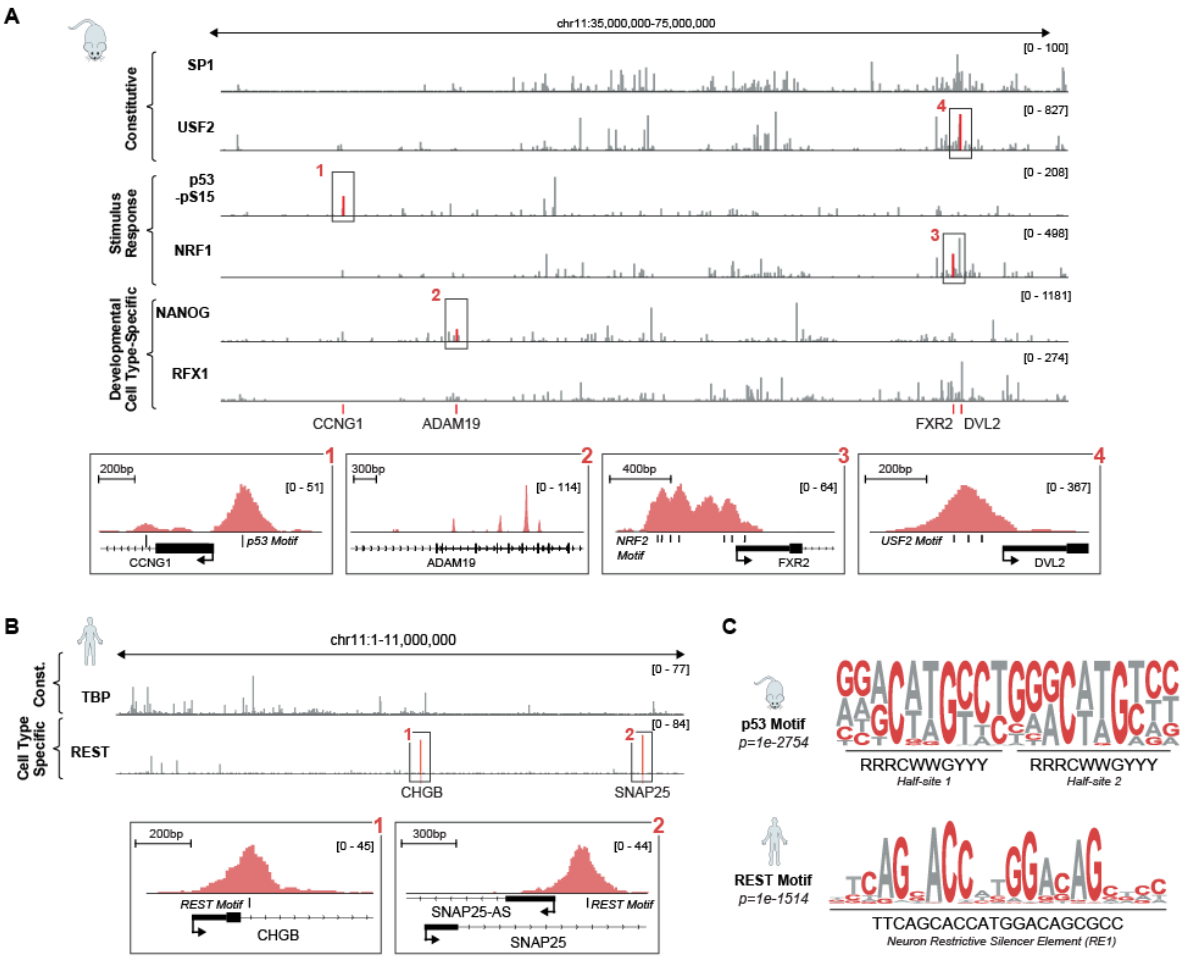


Figure 4: ChIP-DIP accurately maps dozens of transcription factors representing diverse functional classes.

(A) Top: Visualization of six transcription factors (SP1, USF2, p53-pSer15, NRF1, NANOG, RFX1) representing three broad function classes (constitutive, stimulus-response, development/cell type-specific) across a genomic region (mm10, chr11:35,000,000-75,000,000) in mESC. Bottom: Higher-resolution zoom-ins showing individual TF binding patterns at selected targets. (1) p53 binding the p53 response element on the Cyclin G1 gene promoter. (2) Nanog binding a cluster of sites internal to developmental gene. (3) Nuclear Respiratory Factor 1 (NRF1) binding multiple copies of its motif at the promoter of FXR2. (4) The constitutively active USF2 binding its triplicate

E-box motif. **(B)** Visualization of TBP (constitutive) and REST (cell type-specific) across a genomic region (hg38, chr11:1-11,000,000) in K562 cells. Bottom: Higher-resolution zoom-ins highlight two individual peaks of RE-1 Silencing Transcription Factor/ Neuron-Restrictive Silencer Factor (REST/NRSF) at motif sites near promoters of known neuronal gene targets. **(C)** *de novo* generated motifs for p53 (top) in mESCs and REST (bottom) in K562 cells using binding sites identified using ChIP-DIP.

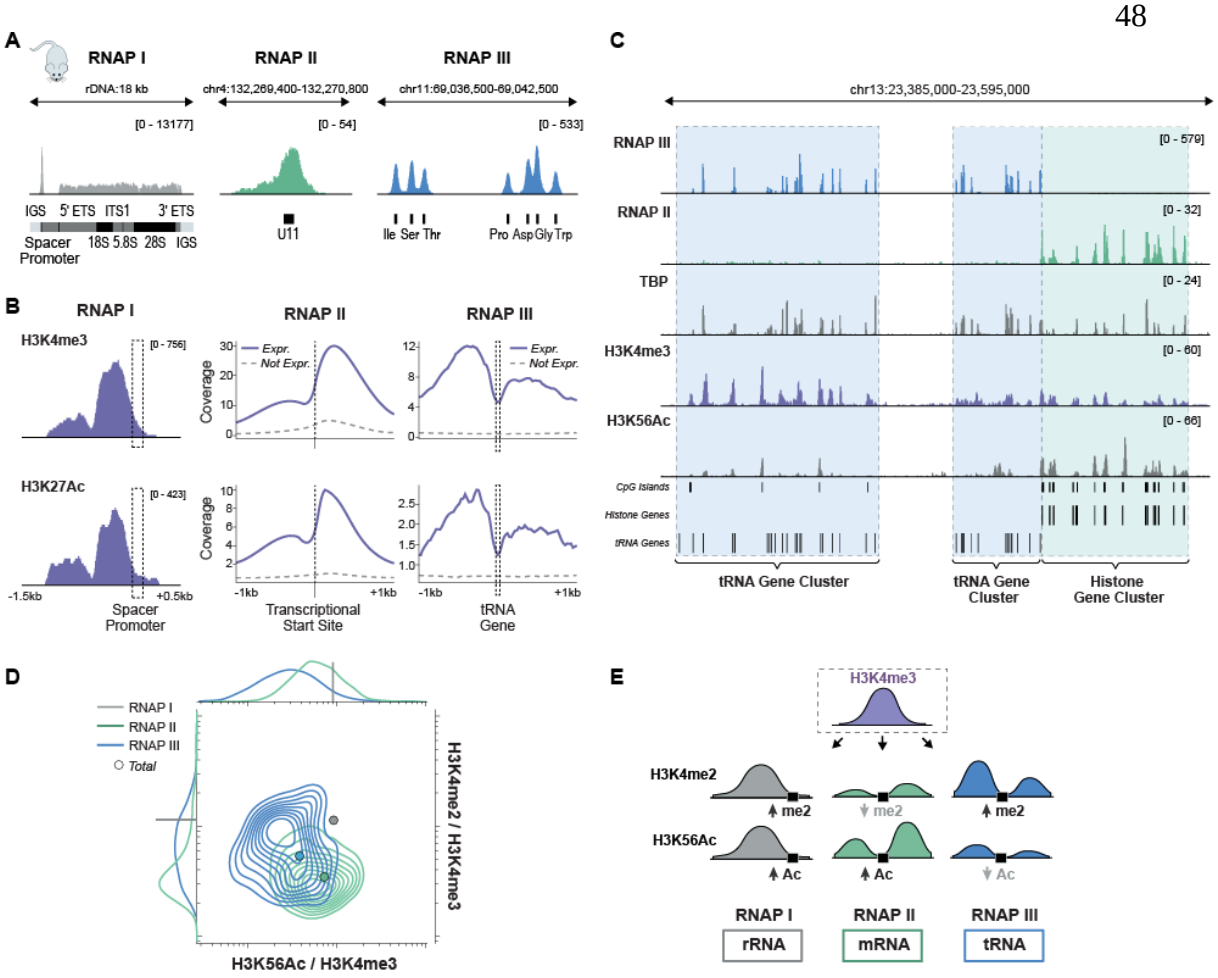


Figure 5: Distinct chromatin signatures define the promoters of each RNA Polymerase.

(A) Visualization of RNAP I at the promoter and along the gene body of rDNA (left), RNAP II at a snRNA gene (middle), and RNAP III at a cluster of tRNA genes (right) in mouse ESCs. **(B)** Visualization of RNAP II and RNAP III along with the shared transcription factor TBP, and histone modifications H3K4me3 and H3K56Ac across a genomic region (mm10, chr13:23,385,000-23,595,000) containing a cluster of RNAP III genes adjacent to a cluster of RNAP II genes (separated by dashed line). **(C)** Comparison of H3K4me3 and H3K27Ac profiles at the promoters of RNAP I, II, and III genes. RNAP I is displayed over the rDNA spacer promoter (left) while RNAP II and III are displayed as metaplots across active (blue) and inactive (dashed gray) promoters. **(D)** Density

distribution of H3K4me2/H3K4me3 versus H3K56Ac/H3K4me3 ratios at RNAP I, active RNAP II and active RNAP III promoters. Points show ratios when computed using the total sum of histone coverage over all promoters. Marginal distributions are shown for RNAP II and III along x and y-axis. Axes are log10 scaled. **(E)** Schematic showing relative levels of histone modifications H3K4me2 and H3K56Ac at H3K4me3 enriched regions and the associated RNAP promoter.

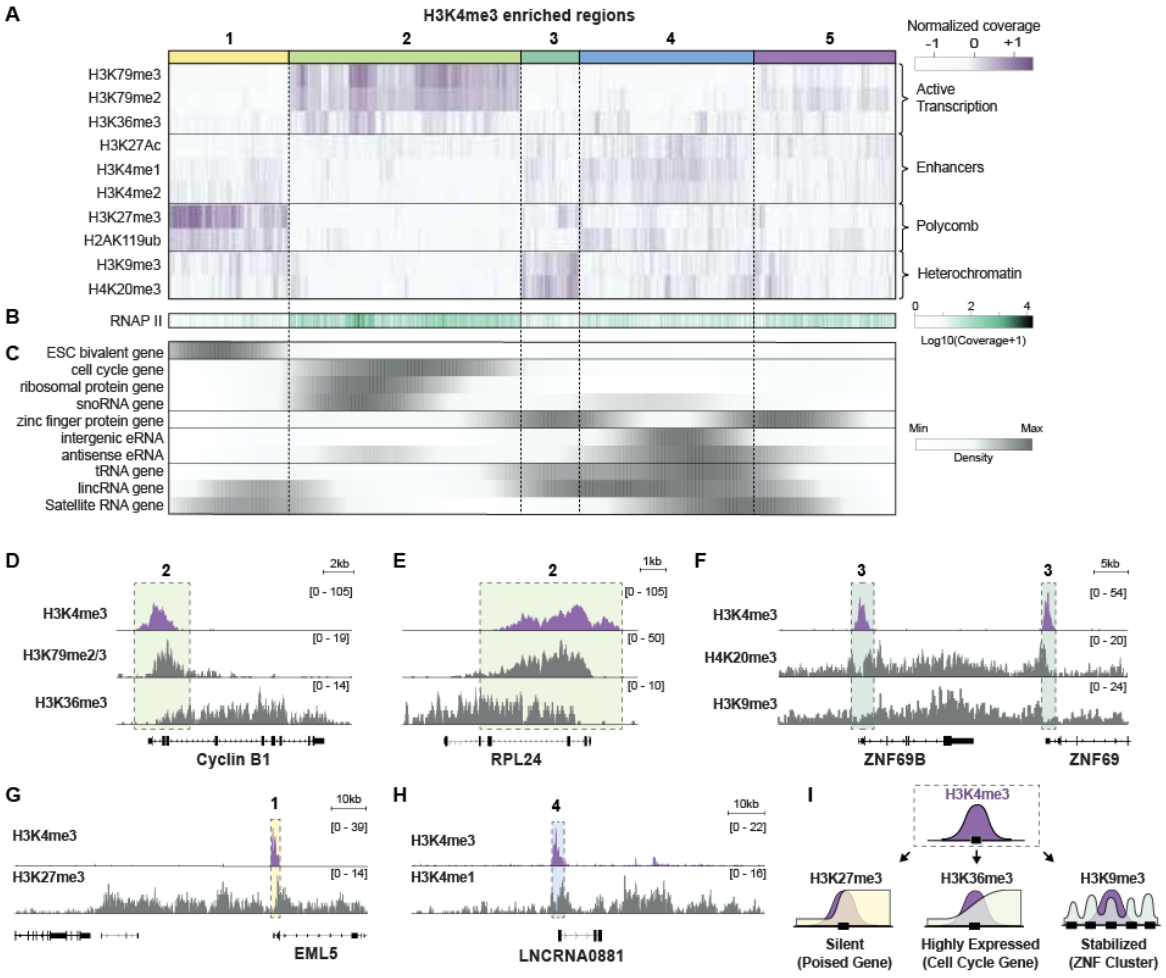


Figure 6: Combinations of histone modifications distinguish RNAP II promoter type, activity, and potential.

(A) Hierarchically clustered heatmap of coverage levels of 10 different histone modifications (y-axis) at individual H3K4me3 enriched genomic regions (x-axis). Five distinct sets of regions are indicated by colored bars along top-axis. **(B)** RNAP II coverage at H3K4me3-enriched regions, sorted as in (A). **(C)** Gene density of 10 different gene classes at H3K4me3 enriched regions, sorted as in (A). **(D/E)** Visualization of H3K4me3, H3K79me2/3, and H3K36me3 histone modifications (associated with set 2) across the cell-cycle associated gene Cyclin B1 and ribosomal protein gene RPL24 in K562. **(F)** Visualization of H3K4me3, H4K20me3, and H3K9me3 (associated with set 3) across

neighboring zinc finger genes in K562. **(G)** Visualization of H3K4me3 and H3K27me3 (associated with set 1) across the EML5 gene in K562. **(H)** Visualization of H3K4me3 and H3K4me1 (associated with set 4) across the long intergenic noncoding RNA gene LNCRNA0881. **(I)** Illustration summarizing the co-occurring promoter-associated histone modifications and their associated gene groups.

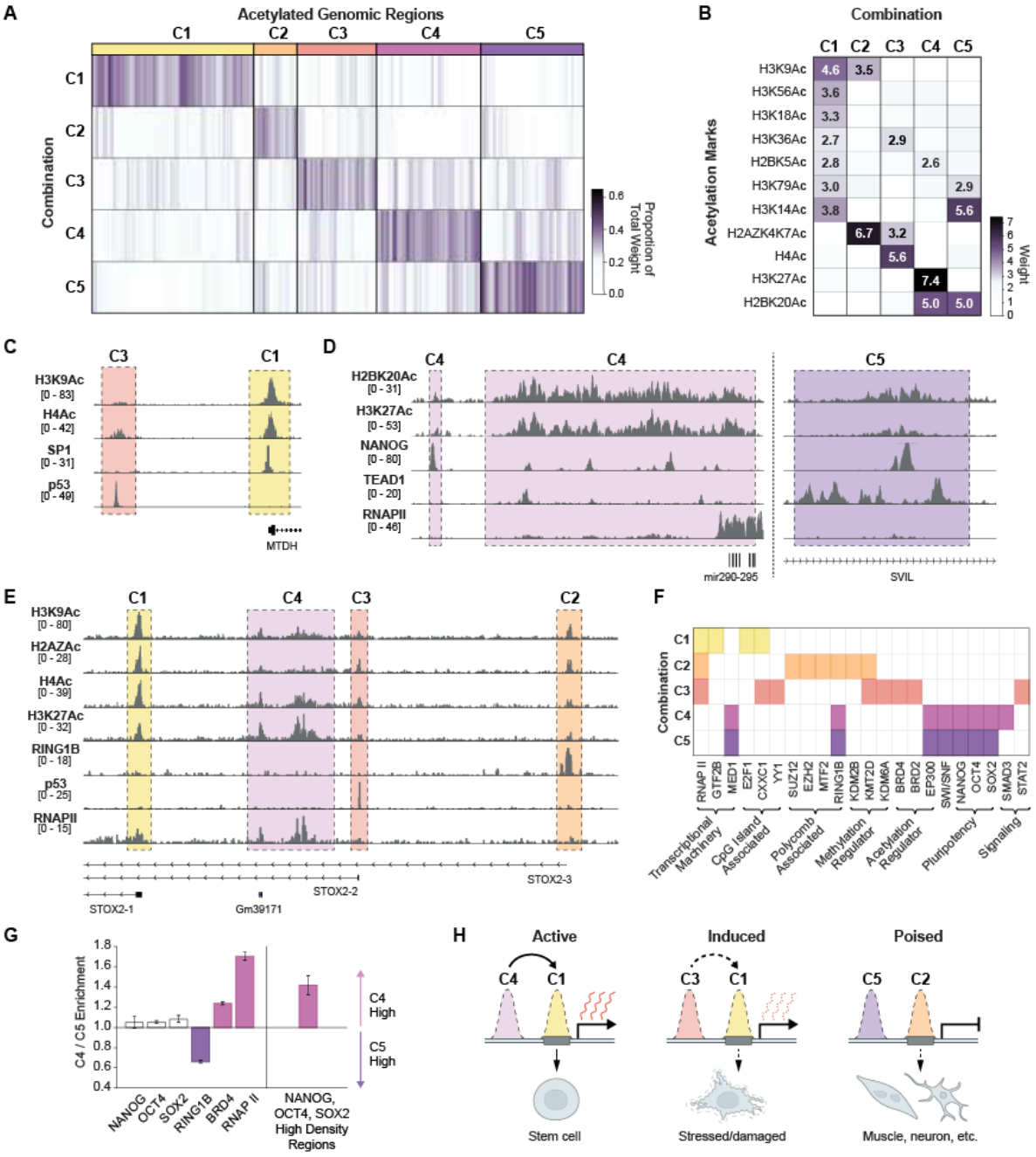


Figure 7: Distinct combinations of histone acetylation marks define unique enhancer types that differ in their activity and developmental potential.

(A) The relative weights of five different combinations of histone acetylation marks (C1-C5, y-axis) for each acetylated genomic region (x-axis). Regions are grouped according to

the combination that received the greatest weight, as indicated along top-axis. **(B)** The relative weights of each histone acetylation mark (y-axis) within each combination (x-axis). Only weights greater than 2.5 are labeled. **(C)** Visualization of H3K9ac and H4ac along with SP1 and P53 across a genomic region (mm10, chr15:34,065,000-34,086,000) containing enhancers assigned to the C1 (yellow) and C3 (red) state. **(D)** Visualization of H2BK20Ac and H3K27Ac along with Nanog, Tead1, and RNAP II across a genomic region (left: mm10, chr7:3,191,500-3,221,500, right: mm10, chr18:5,006,500-5,016,500) containing enhancers assigned to C4 (left) and a distinct region assigned to C5 (right). (Scale of the Nanog track is capped to the maximum of the left region; Tead1 data is from published ChIP-Seq data in fetal cardiomyocytes[78]). **(E)** Visualization of H3K9Ac, H2AZAc, and H4Ac along with RING1B, P53, and RNAP II over a genomic region (mm10, chr8:47,272,800-47,427,000) containing enhancers assigned to all four states (C1-C4). **(F)** DNA-associated proteins (x-axis, ordered by function) with significant binding at genomic regions defined by each combination (y-axis) are indicated in color (see **Methods**). (**F**) Enrichment bargraph of selected transcription-associated factors or regions with high density of pluripotency TFs (see **Methods, Figure S13**) in C4 vs C5 associated-regions. Error bars correspond to the enrichment range from bootstrap resampling. **(G)** Schematic of C1-C5 associated regions and their corresponding functions.

2.6. SUPPLEMENTAL FIGURES

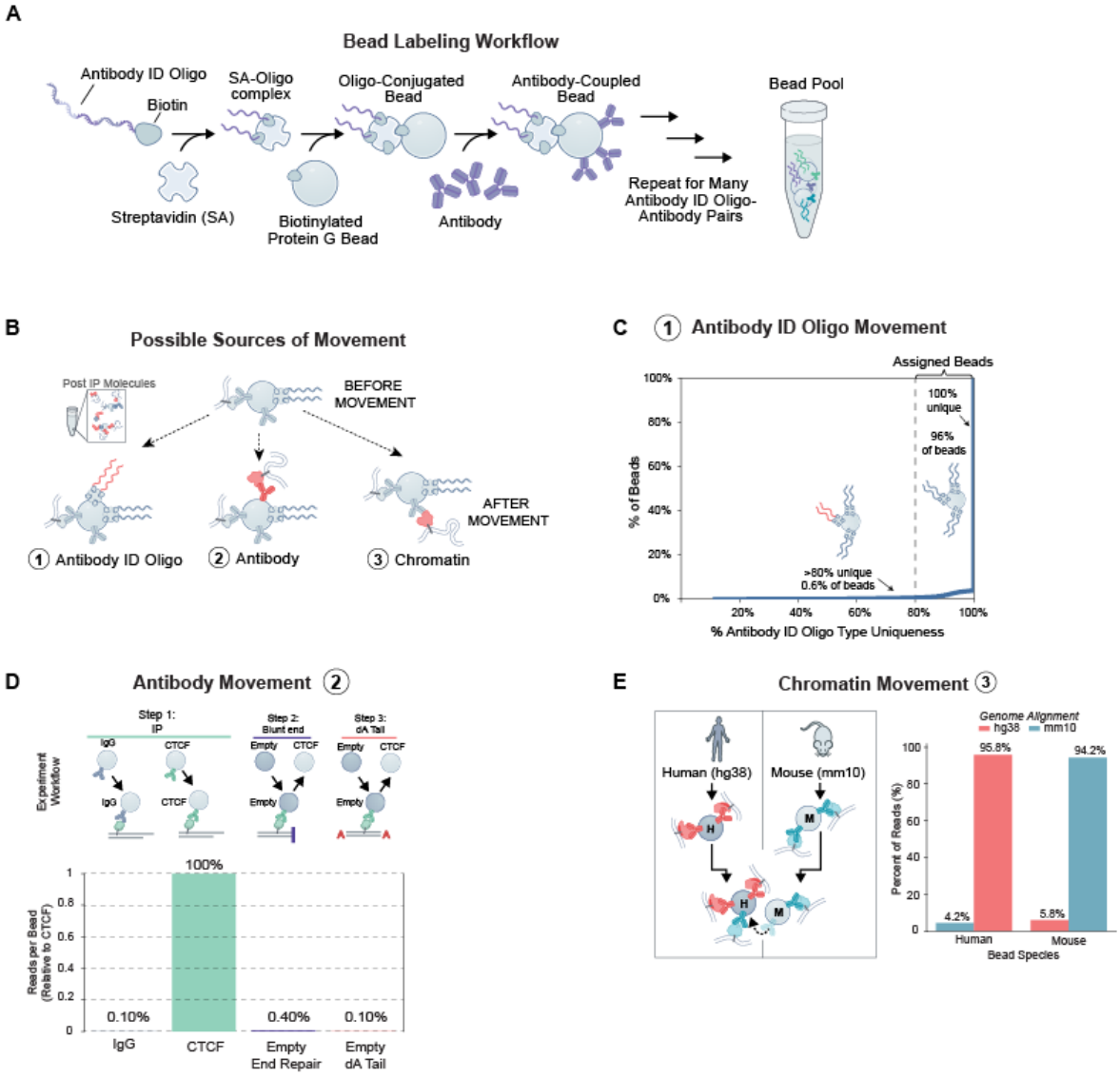


Figure S1: Potential sources of mixing in ChIP-DIP, related to Figure 1. (A) Schematic of labeling strategy to generate Protein G beads coupled with a unique antibody-identifying oligonucleotide and a matched antibody. Protein G beads are covalently modified with a biotin; oligonucleotides containing a 3' biotin are conjugated to streptavidin; oligo-streptavidin complexes are mixed with beads and beads are mixed with antibodies. This process is repeated for each unique oligonucleotide-antibody pair and pooled together. **(B)** Schematic of three potential sources of mixing during ChIP-DIP. **(C)** Cumulative

distribution plot representing the uniqueness of antibody-ID oligos type (x-axis) within individual clusters. **(D)** Schematic of experimental design to test for antibody movement between beads and quantification of relative reads per bead assigned to true targets (CTCF) or empty beads added during experimental processing steps. **(E)** Schematic of human-mouse experimental design to test for chromatin movement and quantification of species-specific reads assigned to human or mouse beads.

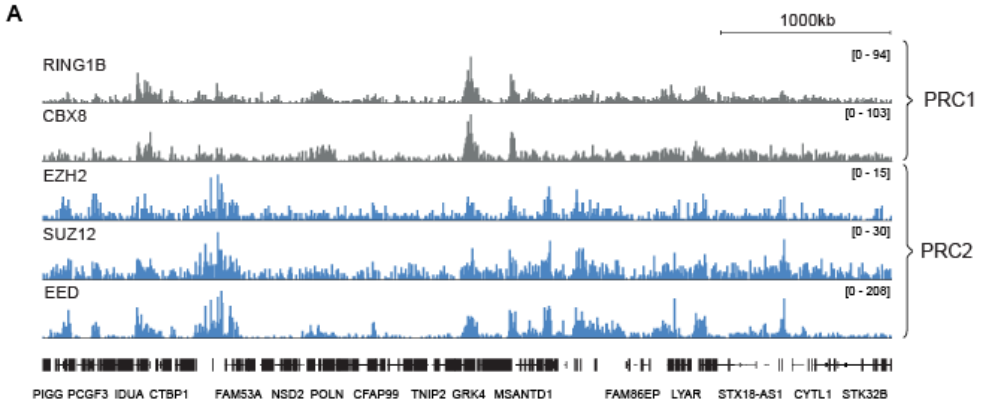


Figure S2: Simultaneous mapping of multiple components of a single protein complex using ChIP-DIP, related to Figure 2. (A) Visualization of various components of the PRC1 (RING1B, CBX8) and PRC2 (EZH2, SUZ12, EED) complexes that were mapped within the same ChIP-DIP pool (K562 52 Antibody Pool) along a genomic region (hg38, chr4:500,000-5,500,000).

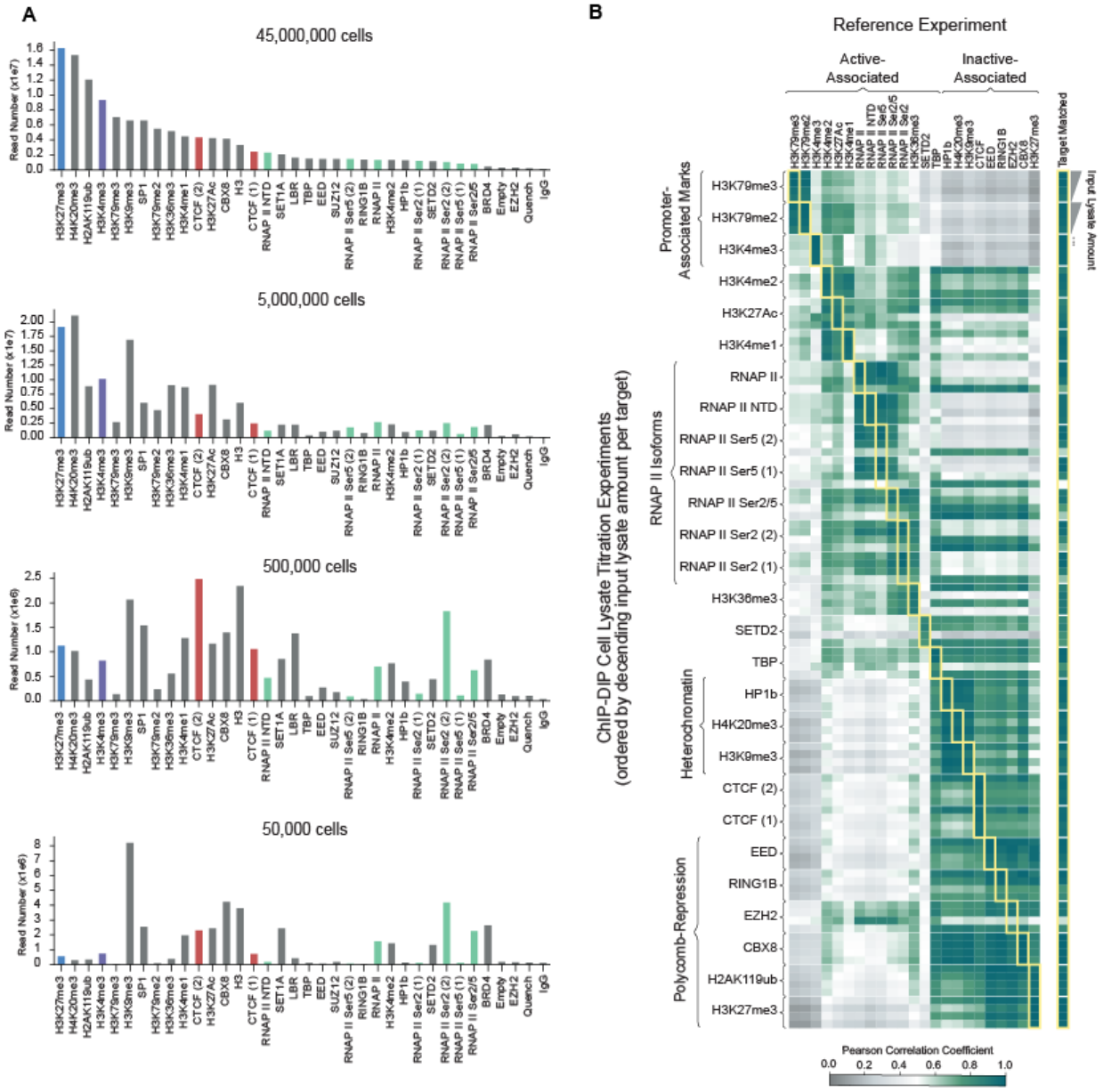


Figure S3: Comparison of ChIP-DIP experiments run with different amounts of input cell lysate, related to Figure 2. (A) Comparison of read counts per target across four ChIP-DIP experiments run using different amounts of input cell lysate (top to bottom: 45M, 5M, 500K, 50K cell equivalents)**. (B)** Correlation heatmap for protein localization maps of various targets measured across different input cell lysate conditions (y-axis) relative to an

independent reference sample (independent high-cell input ChIP-DIP experiment with matched antibodies or in the few cases where this was not available published ChIP-Seq data in K562). For each antibody target, squares are ordered top-to-bottom in descending order of input lysate amount. Squares with matched targets on x and y axes are highlighted in yellow and shown in isolation on the right (“target matched”).

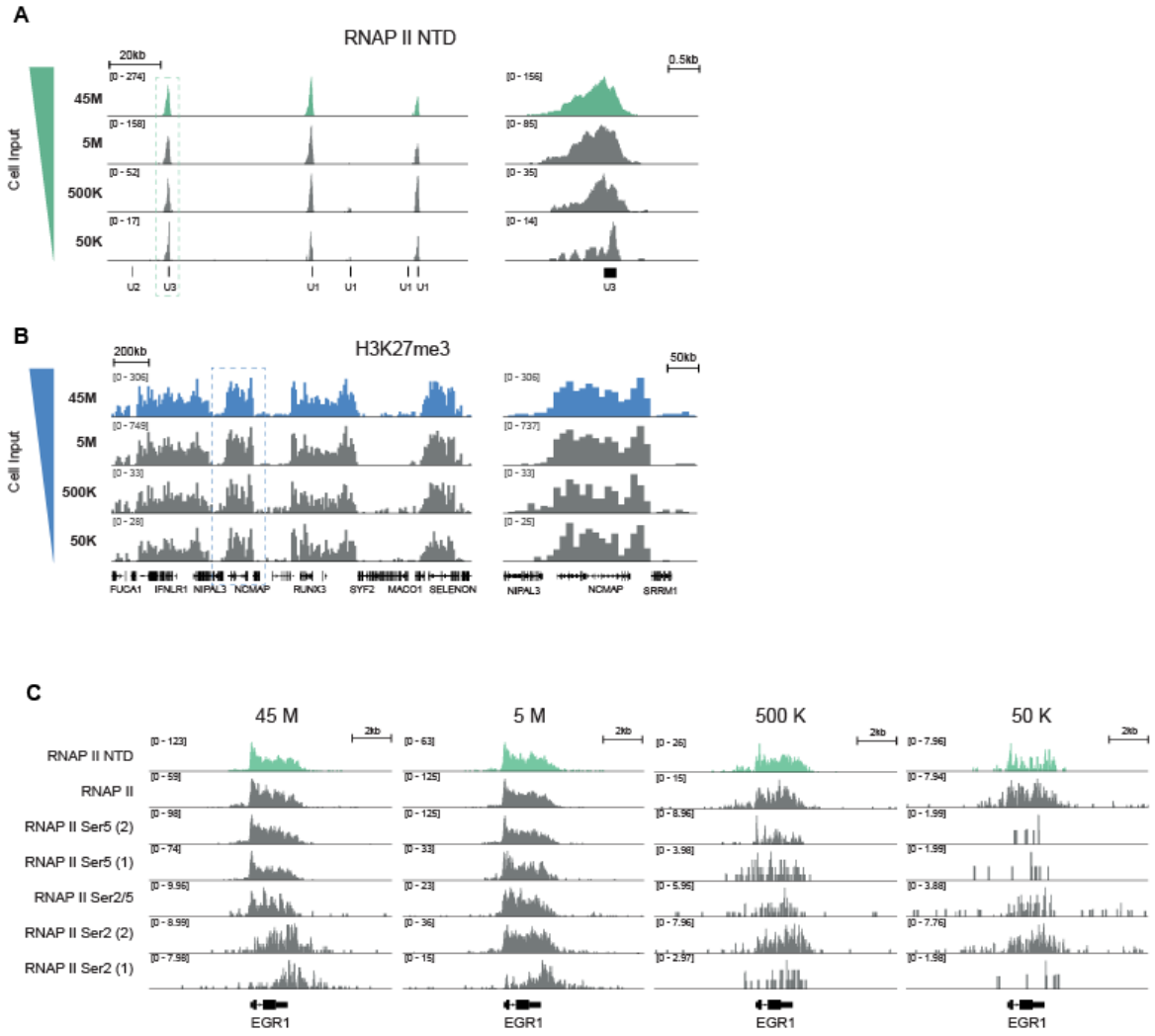


Figure S4. Comparison of protein localization across different amounts of cell lysate, related to Figure 2. (A) Comparison of RNAP II NTD localization across a snRNA gene cluster (hg38, chr17:58,620,000-58,689,000) when generated using various amounts of input K562 cell lysate. **(B)** Comparison of H3K27me3 localization across a genomic region (hg38, chr1:23,850,000-25,850,000) generated using various amounts of input K562 cell lysate. **(C)** Comparison of various isoforms of RNAP II at the EGR1 locus (hg38, chr5:138,455,000-138,480,000) generated using various amounts of input K562 cell lysate.

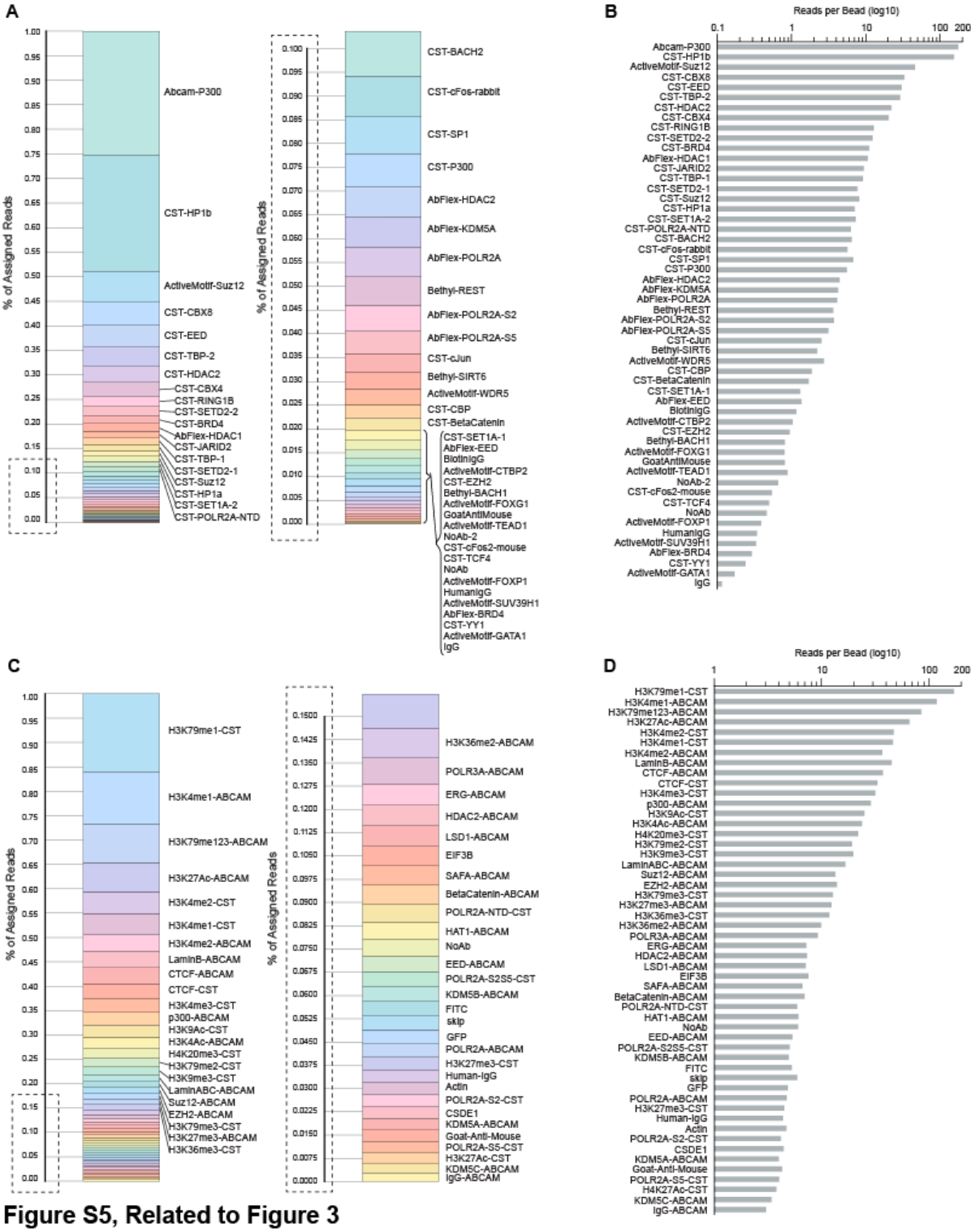


Figure S5, Related to Figure 3

Figure S5: Read statistics for two ChIP-DIP experiments in K562, related to Figure 3. (A) Distribution of reads assigned to each protein in the K562 50 Antibody Pool. **(B)** Reads per bead for each protein target in the K562 50 Antibody Pool. **(C)** Distribution of reads assigned to each target in the K562 52 Antibody Pool. **(D)** Reads per bead for each target in the K562 52 Antibody Pool.

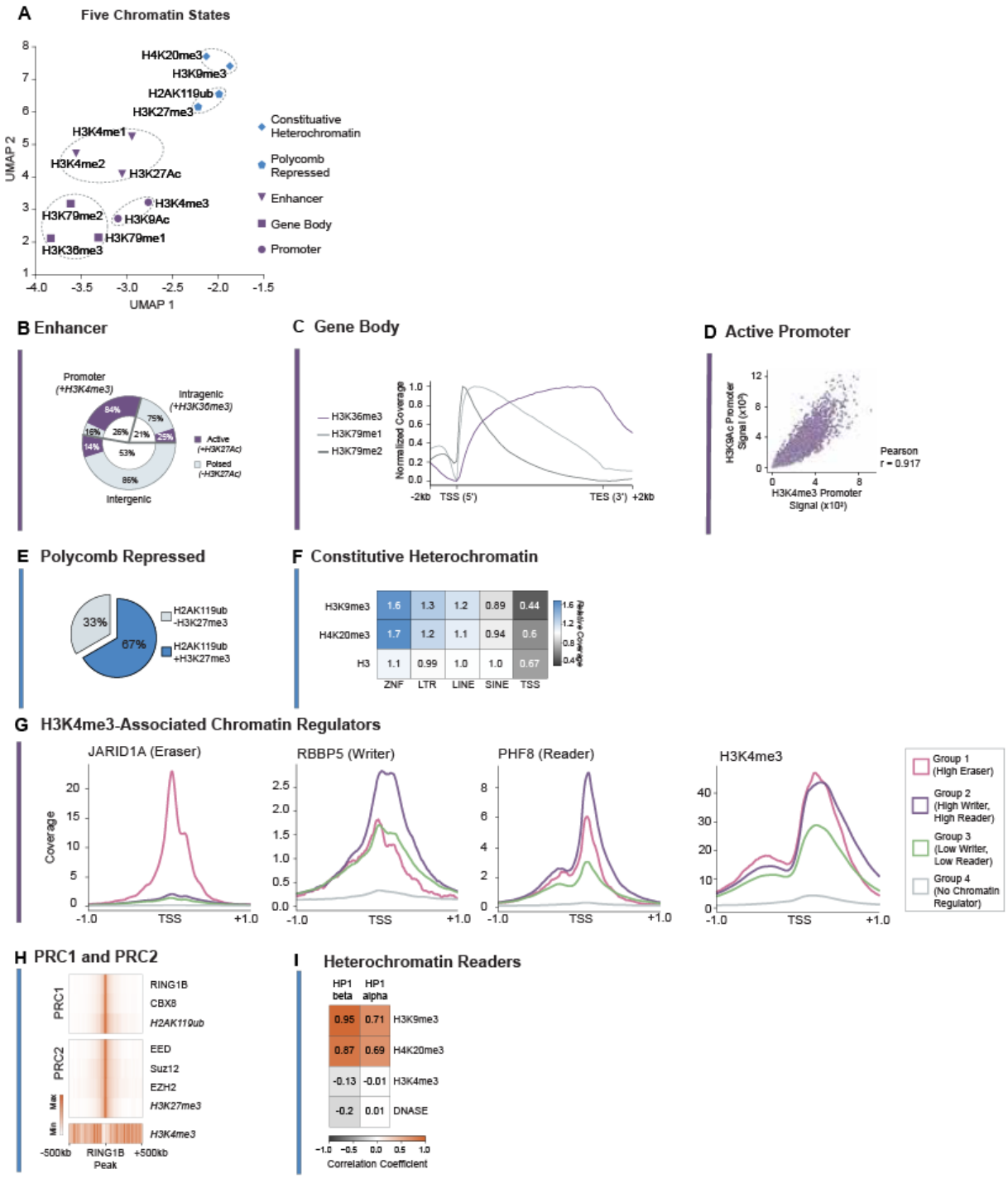


Figure S6: Histone modifications associated with five chromatin states, related to Figure 3. (A) UMAP embedding of 12 histone modifications measured in K562

correspond to five chromatin states. **(B)** Pie chart showing proportion of H3K4me1 peaks in K562 at various genomic position categories — promoters (co-localizing H3K4me3), intragenic (co-localizing H3K36me3) or intergenic — in K562. Proportion of peak regions overlapping with H3K27Ac within each genomic position category are shown in purple. **(C)** Metaplot of signal distribution of H3K36me3, H3K79me1, and H3K79me2 across the gene body of protein coding genes in K562. **(D)** Correlation scatterplot of H3K9Ac and H3K4me3 signals at promoter sites in mESC. **(E)** Pie chart showing overlap of H2AK119ub and H3K27me3 sites in K562. **(F)** Enrichment heatmap of H3K9me3 and H4K20me3 at various associated (ZNF genes, LTRs, LINES) and unassociated (SINES, TSS) genomic elements in K562. H3 is shown as reference. **(G)** Metaplots of read coverage for three H3K4me3-associated chromatin regulators (JARID1A, RBBP5, PHF8) and H3K4me3 at four promoter groups in mESC. Promoter groups were identified using k-means clustering of CR signal (see **Methods**). **(H)** Metaplot showing colocalization of multiple PRC1 and PRC2 members and their respective histone modifications at RING1B sites in K562. **(I)** Genome-wide correlation matrix of multiple HP1 proteins versus heterochromatin and euchromatin markers in K562.

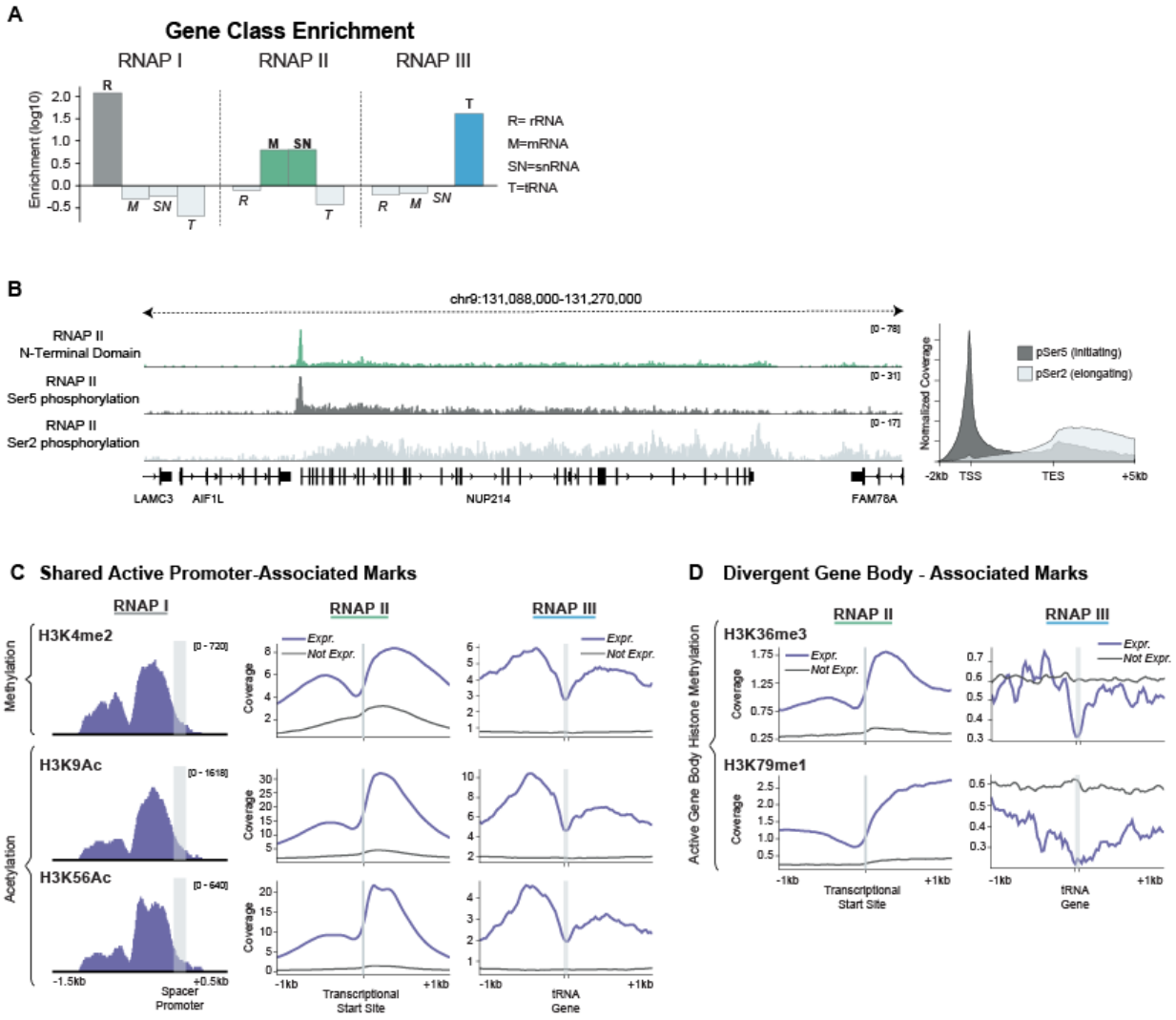


Figure S7: Chromatin states corresponding to distinct RNA polymerases and isoforms, related to Figure 5. **(A)** Bar graph showing enrichment of gene class coverage (rRNA, mRNA, snRNA or tRNA) for RNAP I, II, and III in mESC. For each RNAP, the bar of its associated class (or classes) is highlighted. **(B)** Visualization of RNAP II phosphorylation isoforms across the NUP214 gene in K562. **(C)** Metaplot of signal distribution of RNAP II phosphorylation isoforms across the gene body of protein coding genes in K562. **(D)** Comparison of histone profiles for H3K4me2, H3K9Ac, and H3K56Ac at the promoters of RNAP I, II, and III, similar to **Figure 5B**. (Left) Histone modification

over the RNAP I-transcribed rDNA spacer promoter. (Middle/Right) Metaplot of histone profiles at active (blue) and inactive (gray) promoters for RNAP II (middle) and RNAP III (right). **(E)** Metaplot of H3K36me3 and H3K79me3 at the promoters of RNAP II, and RNAP III.

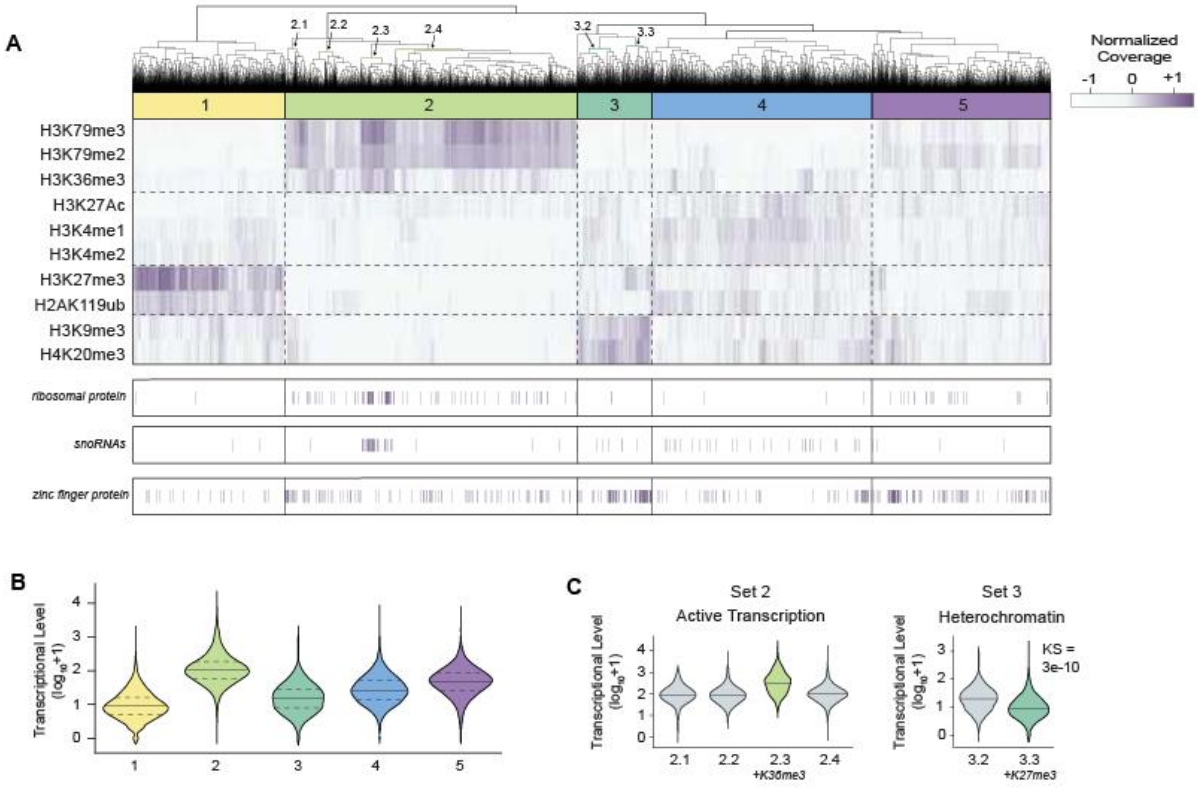


Figure S8: Transcription levels of specific subsets of H3K4me3 enriched regions, related to Figure 6. (A) (Top) Hierarchically clustered heatmap of coverage levels of 10 different histone modifications (y-axis) at individual H3K4me3 enriched regions (x-axis), identical to **Figure 6A**, with dendrogram. (Bottom) Individual H3K4me3 regions corresponding to selected gene classes (Ribosomal protein genes, snoRNA genes or zinc finger genes) are annotated with tick marks. **(B)** Violin plot of transcriptional levels of five major sets of H3K4me3 regions identified in (A). **(C)** (Left) Violin plot of transcriptional levels of subsets of Set 2 (left) and Set 3 (right). Tree levels of subsets are indicated in the dendrogram of (A).

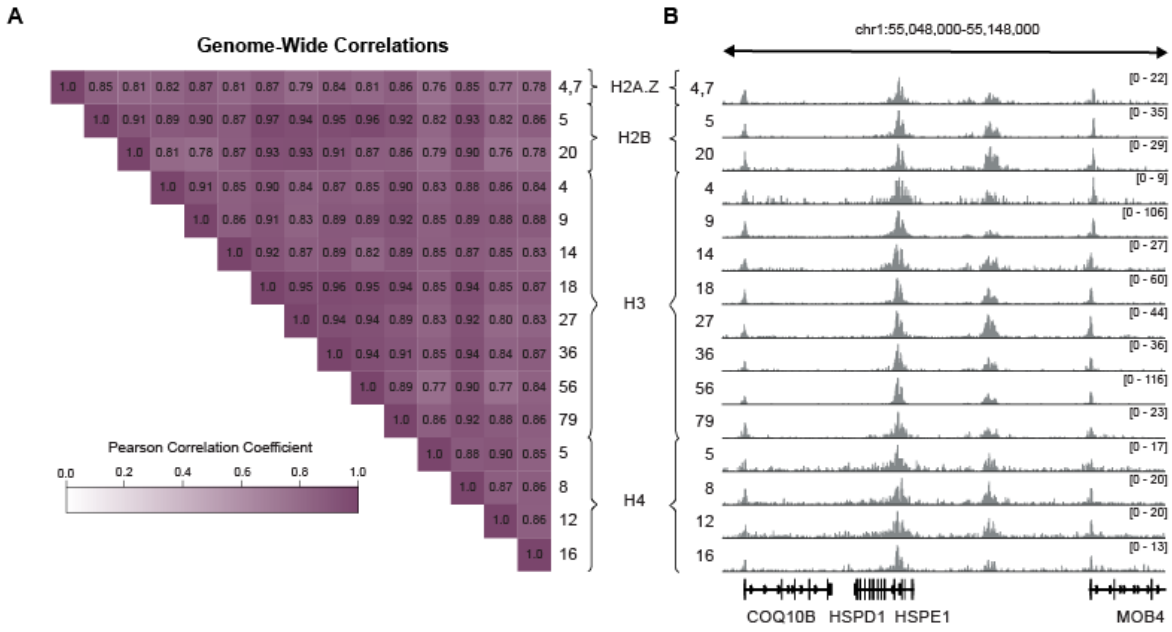


Figure S9: Histone acetylation marks are highly correlated genome-wide, related to Figure 7. (A) Genome-wide Pearson correlation of 15 different histone acetylation marks in mESC. Correlations are based on coverage computed in 10kB windows. **(B)** Comparison of 15 different histone acetylation marks across a genomic region (mm10, chr1:55,048,000-55,148,000) in mESC.

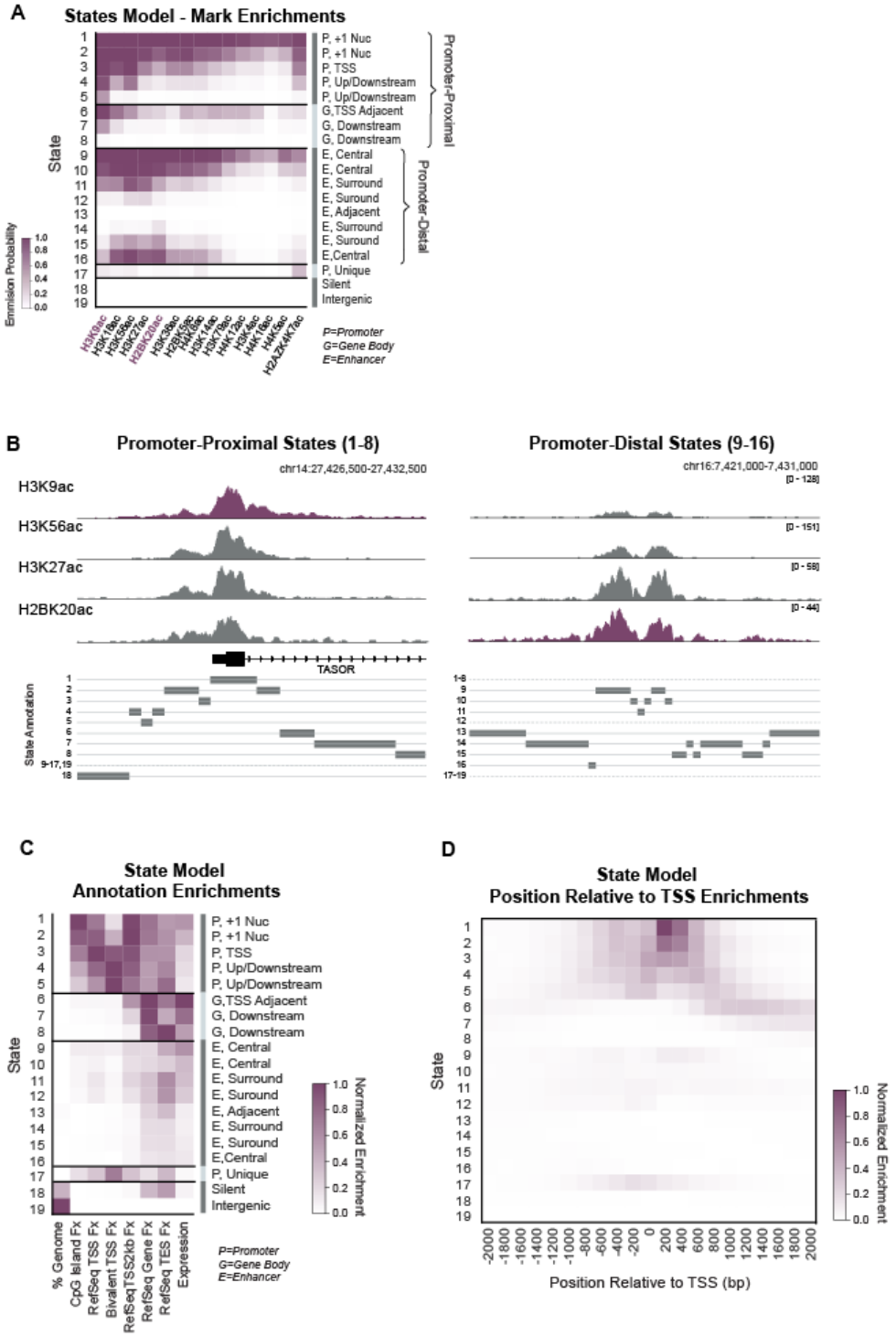


Figure S10: ChromHMM model using histone acetylation marks, related to Figure 7.

(A) Histone acetylation mark emission probability matrix for 19-state ChromHMM model.

State annotations (right) were assigned manually based on genomic position enrichments of states. **(B)** Track visualization of histone acetylation marks (top) and chromatin state annotations (bottom) at example promoter region (left) versus example intergenic region (right). Histone acetylation marks are scaled to the same maximum values at both regions. At each region, the chromatin states that are present are shown with solid lines and a box indicating the exact position; chromatin states that are absent are listed next to dotted lines. **(C)** Heatmap of genome annotation enrichment of chromatin states. Enrichment scores are normalized to the maximum and minimum of each column. **(D)** Heatmap of genomic position enrichment relative to the TSS of chromatin states. Enrichment scores are normalized to the maximum and the minimum of the heatmap.

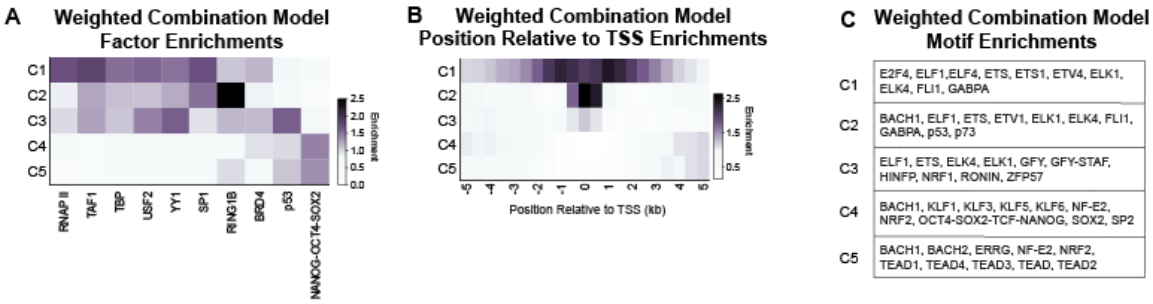


Figure S11: Enrichment profiles for NMF generated combinations (C1-C5) of histone acetylation marks, related to Figure 7. (A) RNAP II, TF, and CR enrichment matrix for regions assigned to combinations (C1-C5) from NMF decomposition of highly acetylated regions using histone acetylation marks. **(B)** Heatmap of genome position enrichments relative to TSS for regions assigned to combinations. **(C)** Transcription factors of top 10 most significant sequence motifs for regions assigned to each combination are listed.

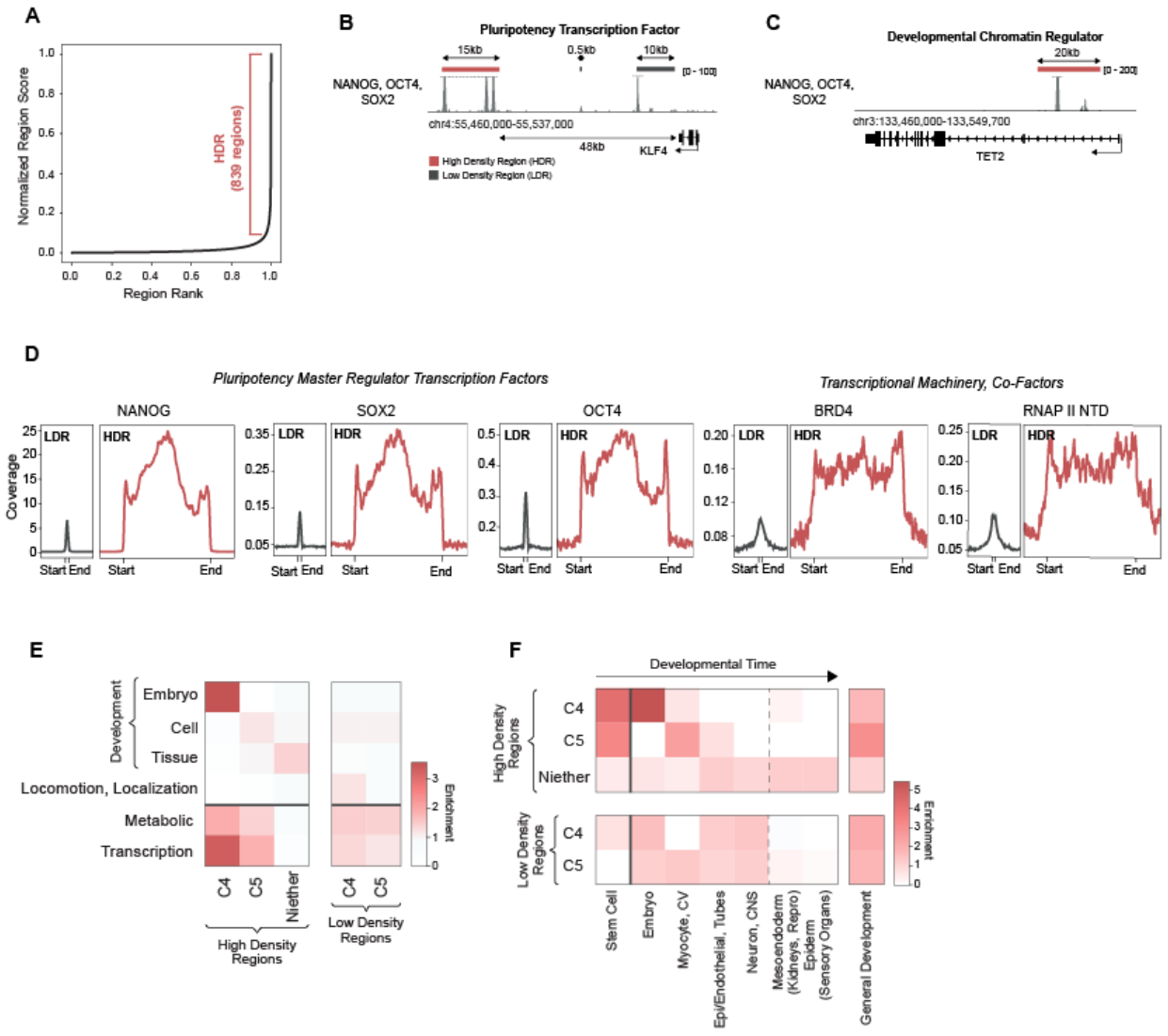


Figure S12: Profiles for high density regions of NANOG-OCT4-SOX2, related to Figure 7. (A) Plot showing normalized region scores (x-axis) for peak regions of NANOG-OCT4-SOX2, ordered by rank (y-axis). High density regions are defined as regions past the point where the slope = 1. **(B)** Track visualization of NANOG-OCT4-SOX2 upstream of the gene for KLF4, a pluripotency transcription factor, in mESC. A high-density region is indicated with a red bar; low-density regions are indicated with grey bars. **(C)** Visualization of NANOG-OCT4-SOX2 near the TET2 gene, a developmentally associated chromatin regulator, in mESC. A high-density region internal to the gene is indicated with

a red bar. **(D)** Coverage metaplots over low density regions (LDR) vs high density regions (HDR) for pluripotency transcription factors and other transcriptional-related factors. Metagenes are centered on the region and the lengths represent the approximate difference in mean lengths (500 for LDRs and 14500 for HDRs). An additional 4kb surrounding each region is shown. **(E)** Enrichment heatmap for GO terms of genes associated with HDRs or LDRs containing C4, C5 or neither C4/C5 chromatin signatures. **(F)** Enrichment heatmap for development-associated GO terms of genes associated with HDRs or LDRs containing C4, C5 or neither C4/C5 chromatin signatures.

2.7. SUPPLEMENTAL TABLE LEGENDS

Table S1. Antibody Pools and Read Counts for ChIP-DIP Experiments

Table S2. Antibody-ID Oligonucleotide Sequences

2.8. SUPPLEMENTAL NOTES

Note S1. Minimal inter-bead mixing during ChIP-DIP ensures accurate protein-DNA assignment. We considered three potential issues in our ChIP-DIP experimental system that could lead to misassignment between antibodies and chromatin: (1) antibody-ID oligo movement, (2) antibody movement, and (3) chromatin movement (**Figure S1B**). To evaluate each of these possibilities, we designed experiments to estimate their frequency.

1. ***Antibody-ID oligo movement***: If antibody ID oligos were to dissociate from their original bead and bind to other beads during the ChIP-DIP procedure, then multiple antibody ID oligo types would share a single split-pool barcode, leading to errors in assigning a chromatin region to the correct protein. If this were the case, we would expect to observe clusters that contain multiple distinct antibody-ID oligos. To explore this, for each split-pool barcode we calculated the proportion of antibody ID oligo reads corresponding to the maximumly represented antibody ID oligo type. In a representative experiment, we observed that 96% of clusters had only a single type of antibody ID oligo (100% maximum representation) and 99.4% of barcodes had at least 80% maximum representation (**Figure S1C**). Since most split-pool barcodes have unique representation, this suggests that antibody ID oligo movement occurs infrequently and should not significantly impact the accuracy of antibody-to-chromatin assignments.

2. ***Antibody movement***: If antibodies for protein X dissociated from their bead and reassociated with a distinct protein G bead containing a label for an antibody recognizing protein Y, then chromatin captured by that antibody would be improperly assigned to protein Y. To quantify the frequency of such events, we performed a ChIP-DIP experiment where we added in oligo-labeled protein G beads that were not conjugated to an antibody or to an IgG antibody and measured the amount of chromatin that was assigned to these beads. In all cases, we observed minimal detection of chromatin (<0.5%). Specifically, we performed a ChIP-DIP

experiment with oligo-labeled beads containing a CTCF antibody, an isotype control IgG antibody, or no antibody (empty). Only beads containing an antibody (i.e., CTCF, IgG) were present during the IP stages and empty beads were added in various post-IP, pre-split pool processing steps to determine the frequency of mixing at each step. If antibodies moved between beads during post-IP processing, we would expect to find chromatin associated with empty beads. When we measured the amount of chromatin assigned to each bead, we observed that beads with the IgG control antibody received 0.10% ($1/1000^{th}$) the amount of chromatin compared to the CTCF antibody (**Figure S1E**). Empty beads added during the end repair reaction and dA-tailing reaction received 0.40% and 0.10% the amount of chromatin of the CTCF antibody beads, respectively. We note that these estimates are likely to be an overestimate of the true mixing rate because this experimental design does not distinguish between chromatin associating due to antibody movement from chromatin that may non-specifically bind to IgG or empty protein G beads. Nonetheless, these results indicate that the impact of antibody movement on chromatin assignment is minimal, representing no more than 0.5% of all chromatin captured.

3. ***Chromatin movement***: If proteins dissociate from their epitope-specific antibodies post IP, they may specifically bind to other beads containing the same epitope-specific antibodies. If this movement occurs during the split-pool process, then these chromatin fragments would not be assigned because they would lack a paired antibody ID oligo with the same barcode. We estimated the frequency of chromatin movement using a human-mouse mixing experiment in which we separately IP'd a human chromatin sample and a mouse chromatin sample using identical pools of labeled beads. After the IP, we mixed the samples and performed the remainder of the standard ChIP-DIP protocol (DNA processing and split-pool). If chromatin dissociates and rebinds prior to split-pool then we would expect that it could rebind to beads from the other species containing the same antibody type as its original bead. We observed that only ~5% of reads were assigned to beads of the other

species (4.2% and 5.8%, **Figure S1E**). This suggests that disassociation and reassociation of chromatin-crosslinked proteins from their epitope-specific antibodies during ChIP-DIP processing steps is minimal. Nonetheless, in the cases where this does occur it would impact the assignability of these chromatin fragments (sensitivity of detection) rather than resulting in incorrect assignment (specificity of detection).

Note S2. Unique sources of background in ChIP-DIP data and normalization approaches to address them. Because beads from many antibodies are processed together, ChIP-DIP has sources of potential background that are distinct from a traditional ChIP-Seq experiment. In any ChIP experiment, the antibody used will immunoprecipitate its specific protein (and the associated chromatin) but will also non-specifically purify other proteins (and their associated chromatin) at some lower frequency. This non-specific chromatin (background) is generally proportional to the overall distribution of genomic DNA present in the starting material ("input"). In ChIP-DIP, the same is true during the IP; any given antibody will preferentially capture its specific protein but will, at some lower frequency, non-specifically capture other proteins. However, because ChIP-DIP entails purification with many antibodies, the source of proteins and chromatin for this non-specific binding is no longer the entire cellular input but rather the material present within the pooled IP (e.g., the proteins and chromatin that were pulled down by the pool of antibodies). Indeed, we observed that some antibodies displayed background signal that was distinct from the input library. For example, antibodies targeting CTCF displayed higher background over promoter regions, likely reflecting the presence of various promoter-enriched histone modifications present in the same experimental pool. To account for this in our analysis, we used the pool of all genomic DNA reads captured in a ChIP-DIP experiment as the background control (see **Methods**). We found that normalization using this empirically defined background led to a more conservative enrichment calculation for ChIP-DIP data. For example, in the CTCF example noted above, this normalization approach successfully removed the background promoter-associated signal while retaining signal at known CTCF binding sites.

Note S3. ChIP-DIP requires low amounts of input cell material. One of the major challenges with mapping DNA binding proteins in primary cell types, disease models, and other rare cell populations is the large number of cells required for traditional ChIP-Seq experiments. Because ChIP-DIP enables simultaneous mapping of many proteins within the same experiment, we reasoned that it may dramatically reduce the total number of cells required in two ways: (i) the number of cells required to map any individual protein is instead distributed across all protein targets in a pool, and (ii) the total chromatin purified from multiple proteins may enable purification of lower DNA concentrations associated with a single/low abundance proteins that might otherwise be lost due to experimental handling.

To test whether ChIP-DIP can generate high-quality data from lower amounts of material, we performed a series of ChIP-DIP experiments using the same antibody pool and differing amounts of cell lysate. The goals of these experiments were to 1) determine whether experimental results changed when lower amounts of input material were used and 2) determine which targets could be successfully mapped using low levels of input material.

Specifically, to enable direct comparison of ChIP-DIP data produced as a function of input material amounts, we crosslinked >100M cells in a single batch and then performed four ChIP-DIP experiments from this same crosslinked lysate in two pairs. For the first pair (45M and 5M conditions), we lysed and sonicated 50 million cells and then split the lysate into 45M and 5M cell equivalents. For the second pair (500K and 50K conditions), we lysed and sonicated 1 million cells and split the lysate into 500K and 50K cell equivalents. We performed ChIP-DIP using an antibody pool containing 35 antibodies targeting 29 distinct proteins from diverse classes (e.g., histone modification, chromatin regulators, transcription factors, and RNA Polymerase) and several distinct targets that are known to colocalize (i.e., multiple components of PRC1/2). This pool also contained multiple antibodies to the same protein (i.e., CTCF, RNAP II) to test for variability of antibody dependence on input amounts and saturation effects.

To examine how various targets behaved in different input conditions, we first visually compared tracks. A subset of targets showed clear enrichments in all conditions (**Figure 2G-H, S4**). These included histone modifications (H3K4me3, H3K79me2, H3K79me3), RNAP II, and CTCF. Notably, H3K79me2/me3 were the two histone modifications that received the lowest read coverage in the 50K condition (91,000 and 39,000 reads respectively; <2% the number of reads received in the 45M condition) but showed high specificity, with most reads near the TSS. However, other targets did not show strong peaks at 50K/500K due to lower overall complexity of the material captured. Next, we calculated correlation coefficients between each ChIP-DIP dataset and the corresponding reference dataset (independent high-cell input ChIP-DIP experiment with matched antibodies or in the few cases where this was not available published ChIP-Seq data in K562) across all genomic regions with high coverage of histone modifications (see **Methods, Figure S3B**). This approach allowed us to test whether global patterns of enrichments were maintained at different input conditions. We found that most proteins mapped displayed high correlation values and similar patterns across all four cell conditions. The exceptions largely reflected proteins present at lower levels or those for which the experimental pool contained multiple targeting antibodies (e.g., RNAP II pSer2), which tended to generate reduced chromatin yields overall (**Table S1**).

Importantly, not all targets showed a consistent trend related to input amounts, suggesting that some of the observed variability is not due to input amounts but other experiment-to-experiment variability. For instance, the 5M condition had higher signal-to-noise for broad histone marks (e.g., H3K9me3, H3K36me3) than the 45M condition, and the 500K condition had the strongest tracks for SETD2 (which localizes at focal binding sites) but the weakest tracks for EZH2 (which localizes across broader regions) (**Figure S3B**).

Note S4. ChromHMM Model of histone acetylation states. To investigate the spatial relationships between histone acetylation marks, we generated a 19-state genome segmentation model using ChromHMM and 15 different histone acetylation marks (**Figure**

S10A). Based on the transition probabilities, we grouped these 19 states and found two large sets that differed in their genome positioning: set 1 (States 1-8) was promoter proximal, with the individual states identifying the relationship to the TSS, while set 2 (States 9-16) was promoter distal, with the individual states demarcating the acetylation peaks and surroundings between them (**Figure S10B-D**). State 17 was also promoter proximal but was grouped separately because of a unique signature with H2AZAc (**Figure S10A, D**). Finally, States 18 and 19 corresponded to silent genic and intergenic regions. Remarkably, this model found that histone acetylation marks were sufficient to define multiple functional elements (e.g., promoters, enhancers, gene bodies, silent, intergenic).

Overall, our state-model found that the acetylation marks cover similar genomic loci; there exist multiple states that have nearly all the marks (i.e., State 1, 2, 9, and 10) and multiple states that appear alike in composition but differ in relationship to genomic annotations (i.e., State 1 vs State 9) (**Figure S10A, D**). Comparing sets 1 and 2 (i.e., promoter proximal vs promoter distal), we found that all acetylation marks are present in both sets, however, some marks are more enriched in one set over the other. For example, H3K9Ac was strongest in set 1 (the promoter proximal set), while H3K18Ac, H3K27Ac, and H2BK20Ac were enriched throughout set 2 (the promoter distal set) (**Figure S10A**). Notably H3K18Ac, H3K27Ac and H2BK20Ac are all targets of the CBP/p300 acetyltransferase[79], which is strongly associated with activity at enhancers. In contrast, the GCN5/PCAF subfamily preferentially acetylates H3K9, H3K14, and H4[80] — all marks we see preferentially in set 1. Comparing all 19 states individually, we found that a small subset of states had greater selective enrichment for specific histone modifications. For example, State 5 (promoter up/down stream) and State 7 (gene body) both had greater selectivity for H3K9Ac, State 14 was defined by H2BK20Ac and State 17 was defined by H2AZAc. Such states may indicate subtle positional shifts or locations unique to these marks. Correspondingly, by visual comparison, we saw that H3K9Ac appears more enriched downstream of the TSS and into the gene body than other histone acetylation marks.

Our 19 state ChromHMM genome segmentation results corresponded well with the findings of our weighted combinatoric NMF model (**Figure 7**). In our NMF model, we

found that TSS associated combinations (C1 and C2) are defined by H3K9Ac; similarly, in our ChromHMM model, we found that H3K9Ac preferred the promoter-associated set of states (set 1). In our NMF model, we predicted a unique role for H2AZAc in defining multiple promoter-associated combinations (C2 and C3); in our ChromHMM model, we found H2AZAc was selectively enriched in State 17, a promoter-associated state. In our NMF model, we found that promoter distal combinations are defined by H2BK20Ac and H3K27Ac (C4 and C5); in our ChromHMM model, we found that these two marks prefer the promoter distal set of states (set 2).

2.9. METHODS

Cell Lines, Cell Culture and Crosslinking

Cell lines used in this study. We used the following cell lines in this study: (i) Female mouse ES cells (pSM44 mES cell line) derived from a 129 × castaneous F1 mouse cross and (ii) K562, a female human lymphoblastic cell line (ATCC, Cat # CCL-243).

Cell Culture Conditions.

(i) pSM44 mES cells were grown at 37C under 7% CO_2 on plates coated with 0.2% gelatin (Sigma, G1393-100ML) and 1.75 mg/mL laminin (Life Technologies Corporation, #23017015) in serum-free 2i/LIF media composed as follows: 1:1 mix of DMEM/F-12 (GIBCO) and Neurobasal (GIBCO) supplemented with 1x N2 (GIBCO), 0.5x B-27 (GIBCO 17504-044), 2 mg/mL bovine insulin (Sigma), 1.37 mg/mL progesterone (Sigma), 5 mg/mL BSA Fraction V (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), 5 ng/mL murine LIF (GlobalStem), 0.125 mM PD0325901 (SelleckChem) and 0.375 mM CHIR99021 (SelleckChem). 2i inhibitors were added fresh with each medium change. Fresh medium was replaced every 24-48 hours depending on culture density, and cells were passaged every 72 hours using 0.025% Trypsin (Life Technologies) supplemented with 1mM EDTA and chicken serum (1/100 diluted; Sigma), rinsing dissociated cells from the plates with DMEM/F12 containing 0.038% BSA Fraction V.

(ii) K562 cells were purchased from ATCC and cultured in 1x DMEM (Life Technologies, # 11965118), 10% FBS (VWR, #97068-091), 100U/mL Penicillin/Streptomycin (Life Technologies, # 15140122), 1mM Sodium Pyruvate (Thermofisher, #11360070), 2mM L-Glutamine (Life Technologies # 25030081) at 37C and 5% CO in 15cm plates (USA Scientific # 5663-9160Q).

Cell Harvest

(i) For harvesting pSM44 mESCs, cells were trypsinized by adding 5 mL of TVP (1 mM EDTA, 0.025% Trypsin, 1% Sigma Chicken Serum; pre-warmed at 37C) to each 15 cm

plate and rocking gently for 3-4 min until cells start to detach. 25 mL of wash solution (DMEM/F-12 supplemented with 0.03% GIBCO BSA Fraction V, pre-warmed at 37C) was added to each plate to inactivate the trypsin. Detached cells were transferred into 50 mL conical tubes, pelleted at 330 g for 3 min, washed in 4 mL of 1X PBS per 10 million cells and then pelleted in 1X PBS in preparation for crosslinking. (ii) For harvesting K562s, the cell suspension was transferred to 50mL conical tubes, pelleted at 330 g for 3 min, washed with 4 mL of 1X PBS per 10 million cells, and then pelleted in 1X PBS in preparation for crosslinking.

Cell Crosslinking

Cells were crosslinked in suspension with 1% formaldehyde for 10 min at room temperature. For both cell lines, during crosslinking steps and subsequent washes, volumes were maintained at 4 mL of buffer or crosslinking solution per 10 million cells. Pelleted cells were resuspended in 1ml of 1X PBS per 10 million cells and pipetted to disrupt clumps of cells. Next, cells were crosslinked in suspension in a final volume of 4 mL of 1% formaldehyde (FA Ampules, Pierce 28906) diluted in 1X PBS per 10 million cells and rocked gently for 10 min at room temperature. Formaldehyde was immediately quenched with addition of 200 ml of 2.5 M glycine (Sigma G7403-250G) per 1 mL of 1% FA solution and incubated with gentle rocking for 5 min at room temperature. Cells were then washed three times with 0.5% BSA in 1X PBS that was kept at 4C. Finally, aliquots of 10 million cells were prepared in 1.7 mL tubes; these cell aliquots were pelleted, flash frozen in liquid nitrogen, and stored in -80C until lysis.

Nuclear Isolation and Chromatin Preparation

Nuclear Isolation.

Crosslinked cell pellets (10 million cells) were lysed using the following nuclear isolation procedure: cells were incubated in 0.7 mL of Nuclear Isolation Buffer 1 (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, 0.5% NP-40, 10% Glycerol, 1X PIC) for 10 min on ice. Cells were pelleted at 850 g for 10 min

at 4C. Supernatant was removed, 0.7 mL of Lysis Buffer 2 (50 mM HEPES pH 7.4, 1.5 mM EDTA, 1.5 mM EGTA, 200 mM NaCl, 1X PIC) was added and the sample was incubated for 10 min on ice. Nuclei were obtained after pelleting and supernatant was removed (as above). Then, 550 uL of Lysis Buffer 3 (50 mM HEPES pH 7.4, 1.5 mM EDTA, 1.5 mM EGTA, 100 mM NaCl, 0.1% sodium deoxycholate, 0.5% NLS, 1X PIC) was added and the sample was incubated for 10 min on ice prior to sonication.

Chromatin fragmentation and size analysis.

Chromatin was fragmented via sonication of the nuclear pellet using a Branson needle-tip sonicator (3 mm diameter (1/8'' Doublestep), Branson Ultrasonics 101-148-063) at 4C for a total of 2.5 min at 4-5 W (pulses of 0.7 s on, followed by 3.3 s off). To check the resulting DNA size distribution, a small aliquot of 20uL of sonicated lysate was then added to 80uL of Proteinase K buffer ((20 mM Tris pH 7.5, 100 mM NaCl, 10 mM EDTA, 10 mM EGTA, 0.5% Triton-X, 0.2% SDS) and reverse crosslinked at 80C for 30 minutes. DNA was isolated using Zymo IC DNA Clean and Concentrator columns and eluted in water. 10uL of purified DNA was then run for 10 minutes on a 1% e-gel (Invitrogen™ E-Gel™ EX Agarose Gels, 1%, Cat.No. G402021). Fragments were found to be 150-700 bp with an average size of roughly 350 bp. The remaining chromatin prep was stored at 4C overnight to be used for the immunoprecipitation the next day.

ChIP-DIP: Bead Preparation

Antibody-ID oligo design

Antibody-ID oligos were designed and ordered from IDT (**Table S2**). The sequence is as follows:
/5phos/TGACTTGNNNNNNNNTATTATGGTAGATCGGAAGAGCGTCGTGTACACAGAGTC/3Bio/. This corresponds to a sticky end that ligates Odd barcodes, UMI, antibody barcode, Illumina primer binding site (i5 primer binding site), spacer sequence. The oligo contains a 5' phosphate to enable ligation and a 3' biotin to enable coupling to beads.

Protein G Bead biotinylation

1 mL of Protein G Dynabeads (ThermoFisher, #10003D) were washed once with 1X PBSt (1X PBS + 0.1% Tween-20), separately keeping the original storage buffer, and resuspended in 1mL PBSt. Beads were then incubated with 20 µL of 5 mM EZ-Link Sulfo-NHS-Biotin (Thermo, #21217) on a HulaMixer for 30 minutes at room temperature. To quench the NHS reaction, beads were placed on a magnet, 500 µL of buffer was removed and replaced with 500 µL of 1M Tris pH 7.4, and beads were incubated on the HulaMixer for an additional 30 minutes at room temperature. Beads were then washed twice with 1 mL PBSt and resuspended in their original storage buffer until use.

Preparation of streptavidin-coupled oligonucleotides

Biotinylated antibody-ID oligonucleotides were coupled to streptavidin (BioLegend, #280302) in a 96-well PCR plate. In each well, 20 µL of 10 µM oligo was added to 75 µL 1X PBS and 5 µL 1 mg/mL streptavidin to make a 909 nM (calculated from the molarity of streptavidin molecules) stock. The 96-well plate was incubated with shaking at 1600 rpm on a ThermoMixer for 30 minutes at room temperature. Each well was diluted 1:4 in 1X PBS for a final concentration of 227 nM.

Preparation of oligonucleotide coupled Protein G beads

For each antibody in the experiment, 10uL of oligonucleotide-coupled Protein G beads were prepared. All biotinylated Protein G beads that would be needed for the entire experiment were first pooled into a tube, washed in 1mL of PBSt and resuspended in 200uL of 1x oligo binding buffer (0.5X PBST, 5 mM Tris pH 8.0, 0.5 mM EDTA, 1M NaCl) per 10uL of beads. 200 µL of bead suspension was aliquoted into individual wells of a deep well 96-well plate (Nunc 96-Well DeepWell Plates with Shared-Wall Technology, Thermo Scientific, Cat. No. 260251) and 14 µL of 5.675nM (1:40 from 227nM working stock made fresh) of streptavidin-coupled oligo was added to each well. The 96-well plate was then sealed with a Nunc 96-well cap mat (Thermo Scientific, Cat. No. 276000) and shaken at 1200 rpm on a ThermoMixer for 30 minutes at room temperature. Beads in each well were

washed twice with M2 buffer (20 mM Tris 7.5, 50 mM NaCl, 0.2% Triton X-100, 0.2% Na-Deoxycholate, 0.2% NP-40), twice with 1X PBSt, and finally resuspended in 200 μL of 1X PBSt.

Estimating number of oligos per bead

After oligo-coupling, a QC step was performed to estimate the number of oligos bound to each bead. 20% of a representative well for each row of the 96-well plate of oligo-coupled beads was isolated and the "Terminal" tag from split-and-pool barcoding was ligated onto the oligos in these aliquots. Then, half of the ligated product was PCR amplified for 10 cycles and purified using 1x SPRI beads. The purified DNA product was run on an Agilent Tapestation using a D1000 tape to estimate molarity and this molarity was used to calculate the total number of molecules post PCR. Using this post-PCR number and the number of cycles of PCR, the number of unique molecules pre-PCR was estimated[22]. Finally, the number of unique molecules was divided by the number of beads put into the PCR reaction (2.7*10^6 beads per 1uL of stock biotintylated protein G beads) to calculate the estimated oligos per bead.

Antibody Coupling

2.5 μg of each antibody was added to each well of the 96-well plate containing oligonucleotide labeled beads resuspended in 1X PBSt. The plate was incubated on a ThermoMixer overnight at 4C with 30 seconds of shaking at 1200 RPM every 15 minutes. The following morning, beads were washed twice with 1X PBSt (Sigma, #B4639-5G), resuspended in 200 μL of 1x PBSt + 4mM biotin + 2.5ug Human IgG Fc and left shaking at 1200 rpm for 15 minutes at room temperature to quench free Protein G or streptavidin binding sites.

Preparation of bead pool

All wells containing oligo labeled, antibody coupled beads were washed 2X with 200uL 1X PBSt + 2 mM biotin, taking care to remove all supernatant after the final wash.

Afterwards, one of two protocols were followed for bead pooling: 1) Equal bead pooling – Beads were pooled using equal amounts of prepared beads for each antibody (10uL of Protein G beads per antibody); 2) Titrated bead pooling – Beads were pooled using unequal amounts of prepared beads for each antibody. The relative number of beads for each antibody was determined based on the chromatin pull-down efficiency (chromatin reads per bead) measured in QC experiments. Fewer beads were used for antibodies with higher pull-down efficiencies and greater beads were used for antibodies with lower pull-down efficiencies or negative controls. This strategy was intended to generate a more uniform distribution for the number of chromatin reads assigned to each antibody in the final experiment.

ChIP-DIP: Immunoprecipitation, Split-and-pool and Library Preparation

Pooled immunoprecipitation

Fragmented lysate was diluted with PBSt +10mM biotin + 1x PIC + 2.5ug of human IgG Fc per 10ul of beads. The pool of labeled beads was added to the lysate and rotated on a HulaMixer for 1 hour at room temperature. Beads were washed 2X with 1mL IP Wash Buffer I (20mM TrispH8.0, 0.05% SDS, 1% Triton X 100, 2mM EDTA, 150mM NaCl in water), 2X with 1mL of IP Wash Buffer II (20mM TrispH8.0, 0.05% SDS, 1% Triton X 100, 2mM EDTA, 500mM NaCl in water) and 2X with 1mL of M2 buffer (20mM Tris pH7.5, 0.2% Triton X100, 0.2% NP-40, 0.2% DOC, and 50mM NaCl).

Chromatin End Repair and dA-tailing

To blunt end and phosphorylate double stranded DNA, the NEB End Repair Module (E6050L; containing T4 DNA Polymerase and T4 PNK) was used. Beads were incubated in 1X NEBNext End Repair Enzyme cocktail + 1X NEBNext End Repair Reaction Buffer + 4mM biotin + 1ug human IgG Fc per 10uL beads at 20C for 15 minutes. The reaction was quenched with 3X volume of PBSt + 100uM EDTA and beads were washed 2X with 1mL PBSt. Next, to dA-tail DNA, the NEBNext dA-tailing Module (Klenow fragment (50 -30 exo-, NEBNext dA-tailing Module, E6053L) was used. Beads were incubated in 1X

NEBNext dA-tailing Reaction Buffer + 1X Klenow Fragment (exo-) + 4mM biotin + 1ug human IgG Fc per 10uL beads at 37C for 15 minutes. The reaction was quenched with 3X volume of PBSt + 100uM EDTA and beads were washed 2X with 1mL PBSt.

Split-and-pool barcoding

Split-and-pool barcoding was performed as previously described[20,22] with modifications. Specifically, beads were first split-and-pool ligated by DPM to attach a common sticky end to all DNA molecules. Then, beads were split-and-pool ligated for ≥ 6 rounds with sets of "Odd," "Even," and "Terminal" tags. The number of barcoding rounds and number of tags used for each round was determined based on the number of beads that needed to be resolved. These parameters were selected to ensure that virtually all barcode clusters (>95%) represented molecules belonging to unique, individual beads. In most cases, 6 rounds of barcoding with 24-36 tags per round were performed. Each individual tag sequence was used in only a single round of barcoding. All split-and-pool ligation steps were performed for 5 minutes at room temperature and supplemented with 2mM biotin and 5.4uM Protein G. After split-and-pool barcoding was complete, beads were resuspended in 1mL of MyRNK buffer [20 mM Tris pH 7.5, 100 mM NaCl, 10 mM EDTA, 10 mM EGTA, 0.5% Triton-X, 0.2% SDS]. Aliquots of various sizes (0.05% to 4% of total beads) were prepared, ensuring that the number of beads within each aliquot was resolvable by the number of possible unique split-and-pool barcodes. Each aliquot was then digested with 8ul of Proteinase K (NEB) for 2 hrs at 55C, 1200RPM shaking and reverse crosslinking at 65C, 1600 RPM shaking overnight.

Library Preparation

DNA from each reverse crosslinked aliquot was isolated with a Zymo IC column using a 6X volume of the DNA binding buffer (Zymo Cat. No. D4014) and eluted in 21ul of H20. Libraries were amplified for 9-12 cycles using Q5 Hot-Start Mastermix (NEB Cat No M0294L) and primers that added the full Illumina adaptor sequences. The following PCR mixture was used: 21uL DNA in H2O, 2uL of i5 primer (12.5uM), 2uL of i7 primer

(12.5uM), 25uL 2X Q5 MM. After amplification, libraries were cleaned with 1.2x SPRI (Bulldog Bio CNGS500) and eluted in 20uL. Prior to sequencing, libraries were gel purified to remove unused primers using a 2% agarose gel [Invitrogen Cat No. G401002].

Sequencing

Sequencing was performed on Illumina NovaSeq S4 (300 cycle) and NextSeq (200 cycle or 300 cycle) paired-end runs, Read lengths were asymmetrical in order to capture the full split-and-pool barcode sequence on read 2 (R2) and the chromatin sequence on read 1 (R1). For 300 cycle kit – 100 cycles for R1, 200 cycles for R2; For 200 cycle kit – 50 cycles for R1 and 150 cycles for R2.

For each experiment, multiple different libraries were generated and sequenced. Each library corresponds to a distinct aliquot which is amplified with a unique pair of primers, providing an additional round of barcoding.

Data Processing Pipeline

Read Processing. Paired-end sequencing reads were trimmed with Trim Galore! V0.6.2 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove adaptor sequences and quality assessed with FastQC v0.11.8. Split-and-pool barcodes were identified from Read 2 using Barcode ID v1.2.0 (https://github.com/GuttmanLab/chipdip-pipeline). Reads missing split-and-pool tags or with tags in the incorrect position given the split-and-pool round they correspond to were discarded. Subsequently, reads were split into two files, one for antibody ID reads and one for DNA reads, based on the presence of "BPM" (bead tag) or "DPM" (DNA tag), respectively, on Read 1.

For DNA reads, the DPM sequence was trimmed from Read 1 using Cutadapt v3.4[81]. The remaining sequence was aligned to mm10 or hg38 using Bowtie2 (v2.3.5)[82] with default parameters. Only primary alignments with a mapq score of 20 or greater were kept for further analysis. Finally, reads were masked using the repeat genome obtained from ENCODE[83].

For antibody ID reads, the BPM sequence, which contains the antibody-ID information, was trimmed from Read 1 using Cutadapt v3.4 and the UMI extracted from the remaining sequence.

MultiQC v1.6[84] was used to aggregate metrics from all steps.

Cluster Generation. A "cluster file" was generated by aggregating all reads (i.e., aligned, masked DNA reads and antibody ID reads) that share the same split-and-pool barcode sequence. During this step, reads in each cluster were deduplicated by alignment position for DNA reads or UMI for antibody ID reads.

Antibody ID Oligo Movement Quality Control Check. To assess the frequency of antibody ID oligo movement between beads, the proportion of antibody ID reads corresponding to the maximum representation in each cluster was calculated. Only clusters with >1 antibody ID read were considered. For each experiment, these values were plotted as an empirical cumulative distribution function (ECDF) using the python plotting package seaborn[85].

Cluster Filtering and Assignment. Individual clusters in the "cluster file" were assigned to a specific antibody based on antibody ID reads within the cluster. First, clusters in the "cluster file" without antibody ID reads or clusters with >10,000 genomic DNA reads (which likely represent undersonicated material or clumps of beads) were filtered out. Next, each remaining cluster was assigned to the antibody ID that had maximum representation within the cluster if 1) there were greater than two unique reads corresponding to the antibody ID and 2) the antibody ID represented >80% of all antibody ID reads within the cluster. These criteria were selected empirically to ensure high confidence assignments of antibody IDs to each cluster. Clusters that did not meet these criteria were removed from further analysis.

Antibody-specific protein maps. Genomic DNA alignments were split into separate bam files such that each file corresponded to all alignments associated with an individual antibody based on the antibody ID assignments within each cluster. DNA reads from clusters that did not have antibody ID reads, were too large, or could not be uniquely

assigned to a single antibody ID were filtered out. DNA reads were deduplicated such that only one read per alignment position per cluster was retained.

Visualization and Peak Calling

Bigwig Generation. Bigwigs were generated from each antibody-specific BAM file using the 'bamCoverage' function from deeptools v3.1.3[86] and were visualized with IGV[87]. Track visualizations are scaled to the maximum over the region and scales indicate reads per bin, unless indicated otherwise.

Background Normalization. To correct for nonspecific background, a background model was generated for each individual antibody using the total pool of assigned sequencing reads. The background for an antibody contained all reads except those assigned to it, or other antibodies targeting the same or related proteins. For example, for an antibody targeting RNAPII-NTD, reads from all antibodies targeting RNAP II were excluded from this background set. To calculate a scaling factor for this background: 1) the total experiment coverage was calculated in 10kB bins, 2) the high coverage bins (80%) were selected, 3) a per-bin enrichment quotient of the target compared to the background coverage was calculated, 4) a kernel density plot of the enrichment quotient was generated, 5) a threshold was calculated based on the position of the smallest peak, and 6) the ratio of total coverage in target versus background bins below the threshold was determined. The goal of this procedure was to locate regions that represented background noise in the target and calculate the target-to-background ratio using only those regions. The kernel density plot was frequently bimodal or with a long tail, with the higher peak or tail representing signal bins and the lower peak representing background bins. Background normalized peaks were called using the scaled background as a substitute for input. Background normalized bigwigs were generated using the 'bamCompare' function from deeptools v3.1.3 by subtracting the scaled background and, subsequently, removing negative value bins.

Peak Calling. Peaks were called using the HOMER v4.11[88] program 'findPeaks' on tag directories generated for target datasets using '-style histone' for histone modification targets and '-style factor' for other targets. Background normalized peaks were generated using the scaled background distribution (described above). Specific parameter settings, such as '-minDist' (distance between adjacent peaks), '-size' (width of peaks) or filtering thresholds were tuned according to the nature of the target. For instance, peaks for focal histone modification H3K4me3 were generated using '-F 2 -P 0.001 -L 0' while enriched regions for broad histone modification H3K36me3 were calculated using '-F 2 -P 0.001 -L 0 -size 1000 -minDist 7500 -region'.

Motif Prediction. Transcription factor motifs were predicted using the HOMER program 'findMotifsGenome' on peaks generated using HOMER, as described above.

Ribosomal DNA Alignments

To analyze reads aligning to genomic DNA encoding ribosomal RNA (rDNA), we aligned reads directly to an rDNA reference. We generated a modified reference of the mouse rDNA sequence (NCBI Genbank BK000964.3)[89]. Because the original mouse BK000964.3 sequence begins with the TSS and ends with the Pol I promoter, we transposed a portion at the end of the rDNA reference to the beginning, as previously described[90], to enable a continuous visualization of the promoter-TSS region. Specifically, the rDNA sequence was cut at the 36,000 nt position and the sequence downstream of the cut site were moved upstream of the TSS, such that the resulting rDNA sequence begins with ~10kb of IGS, then the promoters and then transcribed regions. Processing steps prior to sequence alignment followed the standard ChIP-DIP pipeline. After barcode identification, DNA sequence was aligned to the custom rDNA genome using Bowtie2 (v2.3.5) with default parameters. Only primary alignments with a mapq score of 20 or greater were kept for final analysis. The subsequent cluster generation and read assignment steps followed the standard ChIP-DIP pipeline.

ChIP-DIP Experiments

We performed 8 ChIP-DIP experiments in this paper, each of which, along with the associated antibodies, proteins, and statistics, are described in **Supplemental Table 1**. Briefly, these experiments were:

1. *Chromatin Movement Experiment*: A quality control human and mouse mixing experiment used to quantify possible chromatin movement during the procedure.
2. *Antibody Movement Experiment:* A quality control human and mouse mixing experiment used to quantify possible antibody movement during the procedure.
3. *K562 10 Antibody Pool Experiment:* An initial data-generation experiment performed in human K562 to measure a small number of well-defined targets.
4. *K562 50 Antibody Pool Experiment:* A data-generation experiment performed in human K562 measuring 50 antibodies.
5. *K562 52 Antibody Pool Experiment:* A data-generation experiment performed in human K562 measuring 52 antibodies.
6. *K562 35 Antibody Pool Experiment:* A data-generation experiment in human K562 measuring 35 antibodies as a function of different cell input amounts.
7. *mESC 67 Antibody Pool Experiment:* A data-generation experiment performed in mouse ES cells measuring 67 antibodies.
8. *mESC 165 Antibody Pool Experiment:* A data-generation experiment performed in mouse ES cells measuring 165 antibodies.

All ChIP-DIP experiments were performed using the same general protocol with the following experiment-specific modifications:

1. Chromatin Movement Experiment: To test whether chromatin dissociates during the ChIP-DIP procedure and binds to other beads, we designed a human-mouse mixing experiment. Cell lysate from 20M mESC cells, cell lysate from 10M K562 cells and two sets of antibody-coupled, oligonucleotide-labeled beads were prepared according to standard protocol. Prior to IP, lysate yields were quantified using TapeStation and equal

amounts of mouse and human chromatin preparations were used for the subsequent, separate IPs. One set of antibody-ID labeled beads was used for the human IP and the other set of antibody-ID labeled beads was used for the mouse IP. After IP, the two species-specific IPs were mixed and split into three conditions using different quenchers: (i) 10% BSA quencher, (ii) 1X Blocking Buffer quencher and (iii) No quencher. For the 10% BSA quencher condition, end-repair, dA tailing and DPM reactions were performed in buffer supplemented with 10% BSA. For the 1X blocking buffer quencher condition, end-repair, dA tailing and DPM reactions were performed in buffer supplemented with 1X protein blocking buffer (Abcam ab126587). The three conditions were combined for split-and-pool barcoding.

For alignment of human-mouse mixing experiments, DNA reads were aligned to a custom combination genome including both mm10 and hg38 genomes using Bowtie2 (v2.3.5) with default parameters. Only primary alignments with a mapq score of 20 or greater were kept for further analysis. Reads were then masked using a merged version of mm10 and hg38 blacklist regions defined by ENCODE. Reads were then uniquely assigned to human beads (beads present only in the human IP condition) or mouse beads (beads present only in the mouse IP condition) using the standard assignment pipeline. Total reads aligned to mm10 or hg38 for each bead set were quantified and the relative proportions were plotted as a bar graph.

2. Antibody Movement Experiment: To test whether antibodies dissociate from their labeled beads during the ChIP-DIP procedure and bind to other beads, we designed an experiment that involved the addition of labeled antibody-free beads at various steps. Following a similar set-up as the chromatin mixing experiment, cell lysate from 20M mESC cells, cell lysate from 10M K562 cells and two sets of antibody-coupled, oligonucleotide-labeled beads were prepared using the standard protocol. One set of beads was used for the human IP and the other set of beads was used for the mouse IP. After IP, half of each species-specific IP was mixed together, and half was left separate. For this mixed condition only, oligonucleotide-labeled beads without a coupled antibody were added prior to the end repair and the dA-tailing reactions. These empty beads were added

to capture antibodies that dissociated from other IP'd beads. Finally, the three conditions (mouse only, human only, mixed) were ligated with unique sets of DPM adaptors and combined for split-and-pool barcoding. To calculate the frequency of antibody movement, total reads and total beads assigned to human CTCF beads, human IgG beads, empty beads added prior to end repair and empty beads added prior to dA tailing were quantified. Reads per bead for each group were normalized to the mean value for human CTCF beads. These normalized values were plotted as a bar graph with 99% CI using the python plotting package seaborn.

3. K562 10 Antibody Pool Experiment: We performed an initial small scale proof-of-concept (POC) experiment in K562 using 10 different antibodies. The POC experiment was performed using lysate from 50M K562 cells per IP. Standard protocol with equal bead pooling was used with the exception of IP conditions. Two identical sets of antibody coupled beads were prepared using different biotinylated oligonucleotides; one set was used for an overnight immunoprecipitation at 4C and one set was used for 1-hr immunoprecipitation at room temperature. DNA processing steps and DPM ligation reactions were performed separately for the two IP conditions and then the two samples were pooled for the remaining rounds of split-and-pool barcoding. See **Supplemental Table 1** for full list of antibodies under the "K562 10 Antibody Pool" tab. For data processing, the standard pipeline generated individual clusters corresponding to antibody-IP condition pairs and individual bam files for each target in each IP condition. Data from both IP conditions were merged for each target, resulting in a single file per antibody.

Comparison to ENCODE data: All ChIP-DIP comparisons to ENCODE-generated ChIP-Seq data was performed using this 10 pool experiment in K562. Visual comparisons were performed using IGV and the raw ENCODE datasets: ENCFF656DMV (H3K4me3), ENCFF785OCU (POLR2A), ENCFF800GVR (CTCF) and ENCFF508LLH (H3K27me3). Genome-wide coverage comparisons were calculated across all RefSeq TSS for H3K4me3 and POLR2A or across 10kB bins for CTCF and H3K27me3. Calculations were performed using the 'multiBigwigSummary' function of the python package deeptools v3.1.3 and plotted as 2-D kernel density plots using the python library seaborn.

4. K562 50 Antibody Pool Experiment: The K562 50 Antibody Pool Experiment was performed using lysate from 50M K562 cells. The standard protocol with equal bead pooling was used. See **Supplemental Table 1** for full list of antibodies under the "K562 50 Antibody Pool" tab.

5. K562 52 Antibody Pool Experiment: The K562 52 Antibody Pool Experiment was performed using lysate from K562 cells. To test the efficiency of different crosslinking strategies, two parallel IPs were performed using the same pool of prepared beads. One IP utilized 60M K562 cells crosslinked with 1% FA and the other IP utilized 60M K562 cells crosslinked with 1% FA + DSG. Cells for the 1% FA condition were prepared as described above. Cells for the 1% FA + DSG condition were prepared as follows: After harvest and pelleting, K562 cells were crosslinked in 4 mL of 2 mM disuccinimidyl glutarate (DSG, Pierce) dissolved in 1X PBS per 10 million cells for 45 minutes at room temperature. Cells were then pelleted, washed with 1X PBS and crosslinked with 1% FA, as described above.

For antibody ID oligonucleotide-labeling of beads, beads were labeled in two sequential rounds. First, beads were labeled according to the standard protocol. Then, beads were labeled again using another 2.5uL of 5.67nM streptavidin-coupled oligo in 200uL of 1x oligo binding buffer for 30 minutes at room temperature. During the first round of labeling, all wells received a unique streptavidin-coupled oligonucleotide. During the second round of labeling, most wells received the same streptavidin-coupled oligonucleotide as the first round, with the exception of eleven wells. Eleven pairs of wells received the same streptavidin-coupled oligonucleotide in the second round; one well of each pair was labeled with the same oligonucleotide in both rounds while the other well was labeled with different oligonucleotides. The result was that most beads were labeled with a single, unique oligonucleotide label, eleven beads were labeled with a pair of oligonucleotide labels, and eleven beads were labeled with a single oligonucleotide label that can also be found on other beads. This labeling strategy was designed to test combinatorial labeling of beads. After antibody coupling, beads were pooled in equal amounts and half of the bead pool was used for IP of each crosslinking condition. Following IP, each condition was processed separately and DPM-ligated with unique, condition-identifying sets of adaptors.

Conditions were kept separate for the first round of split-and-pool barcoding and then combined for the remaining rounds of split-and-pool. See **Supplemental Table 1** for full list of antibodies under the "K562 52 Antibody Pool" tab.

Sequenced data was processed using the standard ChIP-DIP pipeline up until the clustering assignment step. To account for the dual oligo labeling of selected antibodies, prior to assignment of unique antibodies to each cluster, clusters with multiple labels (clusters containing both oligo types from a known co-occurring pair) were isolated and antibody-ID oligos in these clusters corresponded to the second round of labeling were reassigned to the matched antibody-ID oligo from the first round of labeling. The result is that all antibodies now corresponded to a unique antibody-ID oligo; for the eleven combinatorial pairs, this is the first round of labeling. Afterwards, the remaining steps in the standard ChIP-DIP pipeline (cluster assignment, etc.) were performed as described above.

6. K562 35 Antibody Pool Experiment for input cell number titration: The K562 35 Antibody Pool Experiment was designed to measure the amount of cell input material required for ChIP-DIP. This experiment involved four separate ChIP-DIP experiments, performed in pairs of two. For the first pair, the 45M and 5M conditions, a 50M cell aliquot was lysed and sonicated and then split into 45M and 5M cell equivalents of lysate. For the second pair, the 500K and 50K conditions, a 1M cell aliquot was lysed and sonicated and then split into 500k and 50k cell equivalents of lysate. Each pair of experiments used a single preparation of antibody-coupled, antibody ID oligonucleotide labeled beads that was split in half. See **Supplemental Table 1** for full list of antibodies under the "K562 35 Antibody Pool" tab.

First, read coverage profiles of four targets — H3K4me3, H3K27me3, CTCF, and RNAP II — were compared. For both RNAP II and CTCF, two different antibodies were included (RNAP II: CST 91151 and 14958S; CTCF: CST 3418S and ABCAM ab128873). Coverage of normalized bigwig files across the set of all peak regions from the 10 Antibody Pool experiment, the same set of regions used for the pool size comparison correlations, was calculated using the 'multiBigwigSummary' function of the python package deeptools

v.3.1.3. Pearson correlation coefficients for all pairs were calculated using the 'plotCorrelation' function of deeptools v.3.1.3 and the plotted as a heatmap, manually ordering the rows/columns from lowest to highest amount of input lysate for each target.

Second, read coverage profiles for a larger subset of targets were correlated with reference profiles over a set of high coverage regions. Only the targets that showed visible signal (i.e., peaks or enriched regions) in the 45M or 5M condition were analyzed. High coverage regions consisted of the set of all peaks from H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K79me4, H3K36me3, H3K9me3, H3K27me3, H4K20me3, and CTCF in either the 45M or the 5M condition. Coverage for all targets over these regions was calculated from raw bam files using the 'multiBamSummary' function and all pairwise Pearson correlation coefficients were calculated using the 'plotCorrelation' function from the python package deeptools.

Reference tracks were defined as follows:

Target	**Reference Experiment**
H3K27me3	K562 10 Antibody Pool Experiment
H3K4me3	K562 10 Antibody Pool Experiment
CTCF	K562 10 Antibody Pool Experiment
RNAP II NTD	K562 10 Antibody Pool Experiment
H3K4me1	K562 50 Antibody Pool Experiment
H3K4me2	K562 50 Antibody Pool Experiment
H3K79me2	K562 50 Antibody Pool Experiment
H3K79me3	K562 50 Antibody Pool Experiment
H3K27Ac	K562 50 Antibody Pool Experiment
RNAP II	K562 52 Antibody Pool Experiment
RNAP II Ser2	K562 52 Antibody Pool Experiment
RNAP II Ser5	K562 52 Antibody Pool Experiment
RNAP II Ser2/5	K562 52 Antibody Pool Experiment
TBP	K562 52 Antibody Pool Experiment
SETD2	K562 52 Antibody Pool Experiment
HP1b	K562 52 Antibody Pool Experiment
RING1B	K562 52 Antibody Pool Experiment
CBX8	K562 52 Antibody Pool Experiment
EZH2	K562 52 Antibody Pool Experiment
EED	K562 52 Antibody Pool Experiment
H3K9me3	ENCODE accession: ENCFF155UQU

H3K36me3	ENCODE accession: ENCFF035SOZ, ENCFF272JVI, ENCFF816ECC, ENCFF880HKV
H4K20me3	SRA accession: SRR8838489, SRR8838492, SRR8838490, SRR8838493, SRR8838491

7. mESC 67 Antibody Pool Experiment: The mESC 67 Antibody Pool experiment was performed using lysate from 80M mESC cells. The standard protocol with titrated bead pooling was used. This experiment contained 67 different antibodies. See **Supplemental Table 1** for full list of antibodies under the "mESC 67 Antibody Pool" tab.

8. mESC 165 Antibody Pool Experiment: The mESC 165 Antibody pool experiment was performed using lysate from 60M mESC cells. Because the number of antibodies exceeded the number of unique antibody-ID oligonucleotides, two experiments were performed in parallel. Two plates of antibody-coupled, oligonucleotide-labeled beads were prepared separately, and pooled using the titrated bead pooling strategy and used to IP half of the prepared cell lysate. After IP, the two samples were processed separately up until the third round of split-and-pool barcoding and then combined for the remaining rounds of split-and-pool. This experiment contained 165 different antibodies. See **Supplemental Table 1** for full list of antibodies under the "mESC 165 Antibody Pool" tab. Sequencing data was processed through the standard pipeline, using a concatenated string of antibody names (i.e., LSD1-CST_SAP30-Bethyl) to match individual antibody-ID sequences during barcode identification. After cluster generation and prior to cluster assignment, each read antibody-ID read was assigned to only one antibody based on its first-round split-and-pool tag. Cluster assignment and BAM file generation then proceeded using the standard pipeline.

Protein Target Classification

Antibody targets were assigned to one of five categories: histone modification (HM), transcription factor (TF), chromatin regulator (CR), RNA polymerase (RNAP), and other

DNA associated protein. Transcription factors were defined as proteins with a DNA-binding domain and were manually subclassified into constitutive, stimulus response or cell type specific/developmental manually curated based on functional descriptions from GeneCards. Chromatin regulators contained proteins or members of complexes that read, write, or erase histone modifications or DNA methylation. Proteins that were part of chromatin regulator complexes and contained a DNA binding domain were considered part of the chromatin regulatory category. Proteins involved in chromatin remodeling (e.g., BRG1) or other structural proteins that interact with chromatin (e.g., Lamin A) were also considered chromatin regulators. Dual function proteins (e.g., transcription factors with intrinsic acetyltransferase capabilities) were assigned to a single category (e.g., transcription factor) but were included in chromatin regulator schematics. Other DNA associated proteins included a mixture of targets, such as RNAP elongation factors (e.g., ELL), RNA binding proteins (e.g., NONO) and antibodies that detected DNA methylation.

Pool Size Comparison Analysis

To measure the influence of the number of antibodies contained within an individual pool, read coverage profiles of four targets — H3K4me3, H3K27me3, CTCF, and RNAP II — generated in four different ChIP-DIP experiments in K562 cells were compared. ChIP-DIP experiments included the10 Antibody Pool, the 45M condition from the 35 Antibody Pool, the 50 Antibody Pool, and the 52 Antibody Pool in K562. For both RNAP II and CTCF, two different antibodies were included (RNAP II: CST 91151 and 14958S; CTCF: CST 3418S and ABCAM ab128873). Coverage of normalized bigwig files across the set of all peak regions from the 10 Antibody Pool experiment was calculated using the 'multiBigwigSummary' function of the python package deeptools v.3.1.3. Pearson correlation coefficients for all pairs were calculated using the 'plotCorrelation' function of deeptools v.3.1.3 and plotted as a heatmap, manually ordering the rows/columns from smallest to largest pool size for each target.

Histone Modification Diversity Analysis

Chromatin-State: Genome-wide coverage for 10kb windows for 12 histone marks (H3K27me3, H2AK119ub, H3K9me3, H4K20me3, and H3K9me3 from the 5M condition in the 35 Antibody Pool Experiment in K562; H3K79me2, H3K79me1, H3K4me3, H3K4me2, H3K4me1, H3K9Ac, and H3K27Ac from the histone panel in K562) was calculated using the 'multiBamCoverage' function from deeptools v3.1.3. These values were standardized for each mark by transforming into z-score values. The UMAP reduction was generated using the UMAP[91] python package and parameters n_components=2 and n_neighbors=3.

Polycomb-Associated Histone Modifications: Validation of polycomb-associated histone modifications used the 5M condition in the K562 35 Antibody Pool Experiment. H3K27me3 and H2AK119ub bam alignment files were converted into binary signal files using the 'BinarizeBam' script from the ChromHMM[92] package with standard settings. The number of bins with only H2AK119ub signal or with both H2AK119ub and H3K27me3 signal were computed and plotted as a pie chart.

Heterochromatin-Associated Histone Modifications: Validation of heterochromatin-associated histone modifications used the 5M condition in the K562 35 Antibody Pool Experiment. Read coverage of H3K9me3, H4K20me3, and H3 were computed over annotation groups (ZNFs, LTRs, LINES, SINES, TSS+/-2kb) using the 'depth' function from samtools v1.9[93]. An enrichment score was calculated by normalizing for feature and target abundance. Specifically, let a = total base pairs within an annotation group, b = effective genome size, c = read coverage of a target over the annotation group, and d = total reads of the target. The enrichment score would be (c/d) / (a/b).

Promoter-Associated Histone Modifications: Validation of promoter-associated histone modifications used the ChIP-DIP histone dataset in mESC. Promoter coverage correlations were calculated across promoters from EPDNew[94], a database of non-redundant eukaryotic

RNAP II promoters, +/- 500bp using the ‘multiBamSummary’ and ‘plotCorrelations’ functions of the python package deeptools v.3.1.3.

Gene Body-Associated Histone Modifications: Validation of gene body-associated histone modifications used the 5M condition in the K562 35 Antibody Pool Experiment and the K562 50 Antibody Pool Experiment. Coverage metaplots over the gene bodies of all protein coding genes from GENCODE v38 basic annotation were calculated using ‘computeMatrix’ function of the python package deeptools v.3.1.3 and normalized to the maximum and minimum for each target.

Enhancer-Associated Histone Modifications: Validation of enhancer-associated histone modifications used the 5M condition in the K562 35 Antibody Pool Experiment and the K562 50 Antibody Pool Experiment. H3K4me1 peaks were assigned to three categories (promoter, gene or intergenic) based on overlap with H3K4me3 (promoter), H3K79me1 (gene) or H3K36me3 (gene). These categories were further sub-divided based on the co-occurrence of H3K27Ac peaks. The proportion of peaks in each category was computed and plotted as a pie chart.

Chromatin Regulator Diversity Analysis

Polycomb-Associated Chromatin Regulators: Validation of polycomb-associated chromatin regulators used the K562 50 Antibody Pool Experiment. Metaplots respective to RING1B peak sites were calculated using ‘computeMatrix’ function of the python package deeptools v.3.1.3 with the following settings: ‘reference-point -bs 10000 -a 500000 -b 500000’. The resulting read coverage profiles were normalized to the maximum and minimum for each target and plotted as a heatmap.

Heterochromatin-Associated Chromatin Regulators: Validation of heterochromatin-associated chromatin regulators used the K562 50 Antibody Pool Experiment. Genome-wide coverage for 10kB windows and Pearson correlation coefficients were calculated using the ‘multiBigwigSummary’ function and ‘plotCorrelation’ function, respectively, of the python package deeptools v3.1.3.

H3K4me3-Associated Chromatin Regulators: Analysis of H3K4me3-associated chromatin regulator used the mESC 165 Antibody Pool Experiments. Binding profiles of JARID1A, RBBP5 and PHF8 were measured +/- 1kB around the TSS of all representative promoters from EPDNew and were clustered using k-means clustering with k=4 by the 'plotCoverage' function of the python package deeptools v.3.1.3. H3K4me3 binding profiles from the mESC 67 Antibody Pool Experiment were measured over the same four promoter groups.

Polymerase Diversity Analysis

RNAP I, II and III Comparison: Validation of the various RNA polymerases used the mESC 165 Antibody Pool Experiment. First, read coverage within a +/- 100bp window surrounding the promoters/TSS of various gene groups were calculated. For tRNAs, the TSS of repeatmasker[95] tRNAs were used. For snRNAs, the TSS of repeatmasker snRNAs (excluding U6 which is transcribed by RNAP III) were used. For mRNAs, EPDNew TSS annotations were used. For rDNA, the spacer promoter was used. Next, for each polymerase, coverage was normalized to the total reads aligned with any gene group. Finally, an enrichment score of the relative coverage compared to IgG was calculated and plotted as a bar graph.

RNAP II Phosphorylation State Comparison: Validation of the various RNA polymerases used the K562 52 Antibody Pool Experiment. Metaplots over the gene bodies of all protein coding genes from GENCODE v38 basic annotation were calculated using 'computeMatrix' function of the python package deeptools v.3.1.3.

Histone Combinatorial Analyses

Polymerase-Associated Histone Profiles

For RNAP I, track coverage profiles of various histone modifications 1.5kB upstream to 0.5kB downstream of the spacer promoter were visualized using IGV.

For RNAP II, metaplots of coverage profiles for various histone modifications were generated around active and inactive RNAP II promoters using the deeptools v.3.1.3 ‘computeMatrix’ (reference-point -a 1000 -b 1000) and ‘plotProfile’ functions. Promoters were defined as the TSS of all representative promoters from EPDNew and were grouped into active or inactive based on the read coverage of RNAP II in the surrounding +/-1kB window.

For RNAP III, metaplots of coverage profiles for various histone modifications were generated around active and inactive tRNA genes using the deeptools v.3.1.3 ‘computeMatrix’ (scale-regions -a 1000 -b 1000 -m 75 -bs 25) and ‘plotProfile’ functions. tRNA genes were grouped into active or inactive based on the read coverage of RNAP III.

For comparison of relative histone levels, total coverage for each histone mark was calculated in the -1.5kB to +0.5kB window surround the spacer promoter for rDNA, -0.5kB to +0.5kB window around active RNAP II promoters and -0.5kB to +0.5kB window around active RNAP III tRNA gene promoters. To account for differences in window size, the coverage of H3K56Ac and H3K4me2 was normalized to the level of H3K4me3. The density profiles of these ratios were plotted using the seaborn ‘jointplot’ function with the following kde parameters: “common_norm=False, thresh=0.2, log_scale=True, levels=10, cut=True”. For comparison to RNAP I, the total sum ratios (e.g., total H3K4me2 coverage across all active RNAP II promoter intervals divided by total H3K4me3 coverage across all active RNAP II promoter intervals) were also calculated and plotted for RNAP II and RNAP III.

H3K4me3 Enriched Regions Clustering

Combinatorial histone modification analysis for H3K4me3 regions used the 5M condition of the K562 35 Antibody Pool Experiment. Read coverage of ten histone targets (H3K79me3, H3K79me2, H3K36me3, H3K4me1, H3K4me2, H3K27Ac, H3K27me3, H2AK119ub, H3K9me3, and H4K20me3) was calculated over all H3K4me3 peak regions using the ‘multicov’ function of bedtools[96]. The resulting region vs histone data matrix (A)

was normalized using log normalization[97]: 1) The log of the data matrix was computed $L = \log(A)$. 2) The column mean ($\bar{L}_{i.}$), row mean ($\bar{L}_{.j}$), and overall mean ($\bar{L}_{..}$) of the log matrix were computed. 3) All individual cells of the final matrix were computed according to $K_{ij} = L_{ij} - (\bar{L}_{i.}) - (\bar{L}_{.j}) + (\bar{L}_{..})$. This method of normalization is intended to capture the "extra" coverage of histone modification *j* in region *i* that is not explained simply by the overall difference between region *i* and other regions or between histone modification *j* and other histone modifications. Instead, it is special to the combination of region *i* (a region with H3K4me3 enrichment) and coverage of histone modification *j*. The regions of the normalized data matrix were clustered using cluster.hierarchy.linkage function from scipy v.1.6.2[98] with a Euclidean distance metric and complete linkage method. The clustered matrix was visualized using the 'clustermap' function of python package seaborn.

Gene annotation of H3K4me3 regions was performed using the 'annotatePeaks.pl' function from HOMER v4.11. ZNF genes, RP genes, and lincRNA genes were defined as regions whose annotation gene description contained the terms 'zinc finger protein', 'ribosomal protein' and 'long intergenic', respectively, and had the nearest TSS within 2000bp. snoRNA genes were defined as all regions whose annotation gene type was snoRNA. Satellite RNA genes were defined as regions whose detailed annotation contained the term 'Satellite'. tRNA genes were defined as all regions that intersected with tRNA gene bodies or upstream by 500bp of the tRNA TSS from repeat masker. Cell cycle genes were defined as regions whose gene annotation belonged to the Kegg Cell Cycle Pathway[99]. Bivalent genes were defined as regions whose gene annotation belong to those identified by Court and Arnaud in human H1 cells[100]. Enhancer RNA regions (both antisense and intergenic) were defined as regions that intersected those identified by Lidschreiber et al.[101] and had the nearest TSS greater than 2000bp away. To visualize enrichments of gene annotations in sets and subsets of the hierarchically clustered heatmap, the kernel density estimate (KDE) was calculated for each annotation group based on their clustering-defined order.

RNAP II levels of individual H3K4me3 regions were measured as the summed coverage over each region from four antibodies targeting RNAP II (RNAP II, RNAP II NTD, RNAP II Ser5, RNAP II Ser2) from the K562 52 Antibody Pool Experiment. Transcriptional levels for sets and subsets of H3K4me3 regions were compared using violin plots generated by the python plotting package seaborn. P-values for comparison of transcriptional levels within subsets of H3K4me3-enriched regions were calculated using the Kolmogorov Smirnov test from scipy.stats.

ChromHMM Model of Acetylation

The ChromHMM genome segmentation model was built using 15 different histone acetylation modifications measured in the mESC 67 Antibody Pool Experiment. Bam files were binarized using the BinarizeBam function from ChromHMM with a Poisson threshold of 0.000001 and other default parameters. The signal threshold was increased from default to remove spurious noise. State models with 5-20 states were built using the LearnModel function with default parameters. States were manually reordered and grouped based on transition probabilities between states. 19 states were selected for the final model to retain state 17, a state with a distinctive enrichment and transition profile.

Non-Negative Matrix Factorization of Acetylated Regions

Non-negative matrix factorization analysis utilized the histone acetylation mark data from the mESC 67 Antibody Pool Experiment. NMF is a matrix factorization technique to reduce dimensionality and explain the observed data using a limited number of combinatorial components[97]. NMF decomposes the original data matrix (dimensions: N x M) into a basis matrix (dimensions: N x k) and a mixture coefficient matrix (dimensions: k x M). In this case, N represents genomic regions of interest, M represents individual histone acetylation marks and k represents the number of combinatorial histone acetylation states. High coverage regions were defined using the results of the ChromHMM Model. Specifically, the 200bp genomic bins corresponding to states with enrichment of multiple histone acetylation marks (states 1,2,3,4,6,9,10,11,12,15,16) were merged to form high

coverage regions. Then, to reduce the number of fragmented or spurious regions, bins with 400 base pairs (2 genomic windows) were merged and regions with size less than 400 base pairs (2 genomic windows) were filtered out. A initial normalized data matrix (N x M) was generated by computing the coverage of each histone modification over each region and normalizing for region size and histone abundance. Specifically, to account for differences in region size between regions, the total reads per region was scaled by region size and, to account for differences in total measured histone abundance between marks, sigmoidal scaling was used[102,103]. NMF was then performed using 'Nimfa'[104], a python library for nonnegative matrix factorization, with the nndsvd initialization method. The rank *k* was selected empirically, taking into account the biological assignability of the resulting states, the complexity of the model and the stability of the factorization (the number of iterations the algorithm required to coverage).

After factorization, the resulting basis matrix (N x k) contained the coefficient of each combination *i* for each genomic region. A sorted heatmap of the basis matrix was generated by grouping the regions according to the combination that contributed the greatest coefficient for each region. For visualization, this heatmap was normalized by dividing the coefficients for each region by the total coefficient sum of the region.

To profile and assign a biological interpretation to individual combinations, each region was assigned to the combination with the maximum coefficient. Identification of transcription factors with significant binding overlap to regions assigned to a single combination was performed using the Cistrome Data Browser, an interactive database of public ChIPseq[105]. For each combination, the top 100 scores were filtered for targets with at least 2 hits in any cell type. Motif enrichment was calculated using the HOMER function 'findMotifs' on all genomic regions assigned to each combination. For comparison of enrichment levels in C4 versus C5, enrichments were calculated using bedgraphs from the mESC 165 Antibody Pool Experiment and the ChromHMM program 'OverlapEnrichment' (java -jar ChromHMM.jar OverlapEnrichment -binres 1 -signal). Interval bars for these enrichments were generated by bootstrap resampling; enrichments were recalculated for 200 independent draws of 75% of the regions assigned to C4 or C5.

High Density Regions of NANOG-OCT4-SOX2

High density regions of pluripotency associated transcription factors were calculated using the NANOG, OCT4 and SOX2 data from the mESC 165 Antibody Pool Experiment. Specifically, high-density and low-density regions were defined using the super-enhancer setting of the 'callPeaks' function from HOMER on the merged tag directors of the three transcription factors. To remove nonspecific background peaks, the merged tag directories of the background models for these three factors was used as input. Briefly, the super enhancer setting with default parameters first identifies peaks, then stiches together individual peaks that are within 12.5kb of each other, calculates a 'super enhancer score' for each region based on input-normalized read coverage, generates a 'super enhancer plot' (regions sorted by score vs number of regions) and identifies the regions where the slope of the plot is greater than 1. These regions are labeled as putative 'super enhancers' while all remaining regions are labeled as 'typical enhancers'. We consider the 'super enhancer' regions as high-density regions (HDR) and the 'typical enhancer' regions as low-density regions (LDR).

TF and CR enrichments over HDRs versus LDRs were calculated using the 'computeMatrix' function with scale-regions setting from deeptools v.3.1.3. To account for the differences in typical region size between LDRs and HDRs, which tended to be much larger, the -m parameter was set to approximately the median region size for each group.

GO terms associated with the intersection of HDRs, LDRs and NMF-based acetylation combinations were calculated using the GO analysis function of 'annotatePeaks' from HOMER. To limit the number of terms under consideration, only terms assigned to the biological process category that received a cutoff $p<0.001$ were used. Terms were then manually grouped into larger categories (e.g., developmental, metabolic). Enrichment scores were calculated by normalizing for the total number of possible unique terms assigned the category and the total number of terms assigned to the intersection group.

Statistics

Pearson correlation coefficients for coverage comparisons versus ENCODE were calculated using pearsonr function of scipy.stats library[98]. Pearson correlation coefficients for heatmaps were generated using the 'plotCorrelation' function from deeptools v.3.1.3[86].

2.10. REFERENCES

1. Kim-Hellmuth, S. (2020). Cell type–specific genetic regulation of gene expression across human tissues. Science *369*, *eaaz8528*.

2. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

3. Xin, B., and Rohs, R. (2018). Relationship between histone modifications and transcription factor binding is protein family specific. Genome Res *28*, 321–333.

4. Jenuwein, T., and Allis, C.D. (2001). Translating the Histone Code. Science *293*, 1074–1080.

5. Kundaje, A. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

6. Dellaire, G., Farrall, R., and Bickmore, W.A. (2003). The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. Nucleic Acids Res *31*, 328–330.

7. Abascal, F. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710.

8. Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497–1502.

9. Kaya-Okur, H.S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun *10*, 1930.

10. Skene, P.J., Henikoff, J.G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat Protoc *13*, 1006–1019.

11. Dunham, I. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

12. Consortium, P. (2015). The PsychENCODE project. Nat Neurosci *18*, 1707–1712.

13. Consortium, T.I.G.P. (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol *9*, 1091–1094.

14. Partridge, E.C. (2020). Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature *583*, 720–728.

15. He, Y. (2020). Spatiotemporal DNA methylome dynamics of the developing mouse fetus. Nature *583*, 752–759.

16. Sisu, C. (2020). Transcriptional activity and strain-specific history of mouse pseudogenes. Nat Commun *11*, 3695.

17. Chasman, D., and Roy, S. (2017). Inference of cell type specific regulatory networks on mammalian lineages. Curr Opin Syst Biology *2*, 130–139.

18. Ota, M. (2021). Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell *184*, 3006–3021 17.

19. Madhani, H.D., Francis, N.J., Kingston, R.E., Kornberg, R.D., Moazed, D., Narlikar, G.J., Panning, B., and Struhl, K. (2008). Epigenomics: A Roadmap, But to Where? Science *322*, 43–44. 10.1126/science.322.5898.43b.

20. Quinodoz, S.A. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744–757 24.

21. Quinodoz, S.A. (2021). RNA promotes the formation of spatial compartments in the nucleus. Cell *184*, 5775–5790 30.

22. Quinodoz, S.A. (2022). SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. Nat Protoc *17*, 36–75.

23. Kim, S., Yu, N.-K., and Kaang, B.-K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. Exp Mol Medicine *47*, 166– 166.

24. Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837. 10.1016/j.cell.2007.05.009.

25. Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*, 693–705. 10.1016/j.cell.2007.02.005.

26. Girbig, M., Misiaszek, A.D., and Müller, C.W. (2022). Structural insights into nuclear transcription by eukaryotic DNA-dependent RNA polymerases. Nat Rev Mol Cell Bio *23*, 603–622. 10.1038/s41580-022-00476-9.

27. Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res *21*, 381–395.

28. Chen, T., and Dent, S.Y.R. (2014). Chromatin modifiers and remodellers: regulators of cellular differentiation. Nat Rev Genet *15*, 93–106.

29. Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat Rev Genet *13*, 613–626.

30. Barba-Aliaga, M., Alepuz, P., and Pérez-Ortín, J.E. (2021). Eukaryotic RNA Polymerases: The Many Ways to Transcribe a Gene. Frontiers Mol Biosci *8*, 663209.

31. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc *12*, 2478–2492. 10.1038/nprot.2017.124.

32. Spicuglia, Salvatore, and Vanhille, L. (2012). Chromatin signatures of active enhancers. Nucleus *3*, 126–131.

33. Steger, D.J., Lefterova, M.I., Ying, L., Stonestrom, A.J., Schupp, M., Zhuo, D., Vakoc, A.L., Kim, J.-E., Chen, J., Lazar, M.A., et al. (2008). DOT1L/KMT4 Recruitment and H3K79 Methylation Are Ubiquitously Coupled with Gene Transcription in Mammalian Cells. Mol Cell Biol *28*, 2825–2839. 10.1128/mcb.02076-07.

34. Gates, L.A., Foulds, C.E., and O'Malley, B.W. (2017). Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. Trends Biochem Sci *42*, 977–989. 10.1016/j.tibs.2017.10.004.

35. Karmodiya, K., Krebs, A.R., Oulad-Abdelghani, M., Kimura, H., and Tora, L. (2012). H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. Bmc Genomics *13*, 424.

36. Chen, Z., Djekidel, M.N., and Zhang, Y. (2021). Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. Nat Genet *53*, 551–563.

37. Saksouk, N., Simboeck, E., and Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. Epigenet Chromatin *8*, *3*.

38. Chen, T., and Dent, S.Y.R. (2014). Chromatin modifiers and remodellers: regulators of cellular differentiation. Nat Rev Genet *15*, 93–106. 10.1038/nrg3607.

39. Ho, L., and Crabtree, G.R. (2010). Chromatin remodelling during development. Nature *463*, 474–484. 10.1038/nature08911.

40. Kirtana, R., Manna, S., and Patra, S.K. (2020). Molecular mechanisms of KDM5A in cellular functions: Facets during development and disease. Exp Cell Res *396*, 112314. 10.1016/j.yexcr.2020.112314.

41. Shilatifard, A. (2008). Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation. Curr Opin Cell Biol *20*, 341–348. 10.1016/j.ceb.2008.03.019.

42. Geng, Z., and Gao, Z. (2020). Mammalian PRC1 Complexes: Compositional Complexity and Diverse Molecular Mechanisms. Int J Mol Sci *21*, 8594.

43. Mierlo, G. van, Veenstra, G.J.C., Vermeulen, M., and Marks, H. (2019). The Complexity of PRC2 Subcomplexes. Trends Cell Biol *29*, 660–671.

44. Bosch-Presegué, L. (2017). Mammalian HP1 Isoforms Have Specific Roles in Heterochromatin Structure and Organization. Cell Reports *21*, 2048–2057.

45. Mazzocca, M., Colombo, E., Callegari, A., and Mazza, D. (2021). Transcription factor binding kinetics and transcriptional bursting: What do we really know? Curr Opin Struc Biol *71*, 239–248. 10.1016/j.sbi.2021.08.002.

46. Bartman, C.R., Hamagami, N., Keller, C.A., Giardine, B., Hardison, R.C., Blobel, G.A., and Raj, A. (2019). Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. Mol Cell *73*, 519-532.e4. 10.1016/j.molcel.2018.11.004.

47. Rada-Iglesias, A. (2008). Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res *18*, 380–392.

48. O'Connor, L., Gilmour, J., and Bonifer, C. (2016). The Role of the Ubiquitously Expressed Transcription Factor Sp1 in Tissue-specific Transcriptional Regulation and in Disease. Yale J Biology Medicine *89*, 513–525.

49. Li, Z., Cogswell, M., Hixson, K., Brooks-Kayal, A.R., and Russek, S.J. (2018). Nuclear Respiratory Factor 1 (NRF-1) Controls the Activity Dependent Transcription of the GABA-A Receptor Beta 1 Subunit Gene in Neurons. Front Mol Neurosci *11*, 285.

50. Satoh, J., Kawana, N., and Yamamoto, Y. (2013). Pathway Analysis of ChIP-Seq-Based NRF1 Target Genes Suggests a Logical Hypothesis of their Involvement in the Pathogenesis of Neurodegenerative Diseases. Gene Regul Syst Biology *7*, GRSB.S13204. 10.4137/grsb.s13204.

51. Horn, H.F., and Vousden, K.H. (2007). Coping with stress: multiple ways to activate p53. Oncogene *26*, 1306–1316. 10.1038/sj.onc.1210263.

52. Fischer, M. (2017). Census and evaluation of p53 target genes. Oncogene *36*, 3943–3956. 10.1038/onc.2016.502.

53. Akberdin, I.R. (2018). Pluripotency gene network dynamics: System views from parametric analysis. Plos One *13*, *e0194464*.

54. Reith, W., Herrero-Sanchez, C., Kobr, M., Silacci, P., Berte, C., Barras, E., Fey, S., and Mach, B. (1990). MHC class II regulatory factor RFX has a novel DNA-binding domain and a functionally independent dimerization domain. Gene Dev *4*, 1528–1540. 10.1101/gad.4.9.1528.

55. Brivanlou, A.H., and Jr., J.E.D. (2002). Signal Transduction and the Control of Gene Expression. Science *295*, 813–818. 10.1126/science.1066355.

56. Qi, B., Sang, R.G.N.Q.-X.A., and Sang, Q.-X.A. (2009). ADAM19/Adamalysin 19 Structure, Function, and Role as a Putative Target in Tumors and Inflammatory Diseases. Current Pharmaceutical Design *20*, 2336–2348. 10.2174/138161209788682352.

57. Schoch, S., Cibelli, G., and Thiel, G. (1996). Neuron-specific Gene Expression of Synapsin I MAJOR ROLE OF A NEGATIVE REGULATORY MECHANISM (∗). J Biol Chem *271*, 3317–3323. 10.1074/jbc.271.6.3317.

58. Martin, D., and Grapin-Botton, A. (2017). The Importance of REST for Development and Function of Beta Cells. Frontiers Cell Dev Biology *5*, 12. 10.3389/fcell.2017.00012.

59. Bao, F., LoVerso, P.R., Fisk, J.N., Zhurkin, V.B., and Cui, F. (2017). p53 binding sites in normal and cancer cells are characterized by distinct chromatin context. Cell Cycle *16*, 2073–2085.

60. Otto, S.J. (2007). A New Binding Motif for the Transcriptional Repressor REST Uncovers Large Gene Networks Devoted to Neuronal Functions. J Neurosci *27*, 6729–6739.

61. Barba-Aliaga, M., Alepuz, P., and Pérez-Ortín, J.E. (2021). Eukaryotic RNA Polymerases: The Many Ways to Transcribe a Gene. Frontiers Mol Biosci *8*, 663209. 10.3389/fmolb.2021.663209.

62. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49. 10.1038/nature09906.

63. Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol *28*, 817–825. 10.1038/nbt.1662.

64. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326. 10.1016/j.cell.2006.02.041.

65. Wang, H. (2023). H3K4me3 regulates RNA polymerase II promoter-proximal pause-release. Nature *615*, 339–348.

66. Huang, H., Sabari, B.R., Garcia, B.A., Allis, C.D., and Zhao, Y.S. (2014). Histone Modifications. Cell *159*, 458–458 1.

67. Zentner, G.E., and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. Nat Struct Mol Biol *20*, 259–266. 10.1038/nsmb.2470.

68. Giaimo, B.D., Ferrante, F., Vallejo, D.M., Hein, K., Gutierrez-Perez, I., Nist, A., Stiewe, T., Mittler, G., Herold, S., Zimmermann, T., et al. (2018). Histone variant H2A.Z deposition and acetylation directs the canonical Notch signaling response. Nucleic Acids Res *46*, gky551-. 10.1093/nar/gky551.

69. Gévry, N., Chan, H.M., Laflamme, L., Livingston, D.M., and Gaudreau, L. (2007). p21 transcription is regulated by differential localization of histone H2A.Z. Gene Dev *21*, 1869–1881. 10.1101/gad.1545707.

70. Giaimo, B.D., Ferrante, F., Herchenröther, A., Hake, S.B., and Borggrefe, T. (2019). The histone variant H2A.Z in gene regulation. Epigenet Chromatin *12*, 37. 10.1186/s13072-019-0274-9.

71. Gévry, N., Hardy, S., Jacques, P.-É., Laflamme, L., Svotelis, A., Robert, F., and Gaudreau, L. (2009). Histone H2A.Z is essential for estrogen receptor signaling. Gene Dev *23*, 1522–1533. 10.1101/gad.1787109.

72. Currey, L., Thor, S., and Piper, M. (2021). TEAD family transcription factors in development and disease. Development *148*. 10.1242/dev.196675.

73. Später, D., Hansson, E.M., Zangi, L., and Chien, K.R. (2014). How to make a cardiomyocyte. Development *141*, 4418–4431. 10.1242/dev.091538.

74. Mashtalir, N., Dao, H.T., Sankar, A., Liu, H., Corin, A.J., Bagert, J.D., Ge, E.J., D'Avino, A.R., Filipovski, M., Michel, B.C., et al. (2021). Chromatin landscape signals differentially dictate the activities of mSWI/SNF family complexes. Science *373*, 306–315. 10.1126/science.abf8705.

75. Vangala, P., Murphy, R., Quinodoz, S.A., Gellatly, K., McDonel, P., Guttman, M., and Garber, M. (2020). High-Resolution Mapping of Multiway Enhancer-Promoter

Interactions Regulating Pathogen Detection. Mol Cell *80*, 359-373.e8. 10.1016/j.molcel.2020.09.005.

76. Arrastia, M.V., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A., Ismagilov, R.F., and Guttman, M. (2022). Single-cell measurement of higher-order 3D genome organization with scSPRITE. Nat Biotechnol *40*, 64–73. 10.1038/s41587-021-00998-1.

77. Goronzy, I.N., Quinodoz, S.A., Jachowicz, J.W., Ollikainen, N., Bhat, P., and Guttman, M. (2022). Simultaneous mapping of 3D structure and nascent RNAs argues against nuclear compartments that preclude transcription. Cell Reports *41*, 111730. 10.1016/j.celrep.2022.111730.

78. Akerberg, B.N., Gu, F., VanDusen, N.J., Zhang, X., Dong, R., Li, K., Zhang, B., Zhou, B., Sethi, I., Ma, Q., et al. (2019). A reference map of murine cardiac transcription factor chromatin occupancy identifies dynamic and conserved enhancers. Nat Commun *10*, 4907. 10.1038/s41467-019-12812-3.

79. Weinert, B.T., Narita, T., Satpathy, S., Srinivasan, B., Hansen, B.K., Schölz, C., Hamilton, W.B., Zucconi, B.E., Wang, W.W., Liu, W.R., et al. (2018). Time-Resolved Analysis Reveals Rapid Dynamics and Broad Scope of the CBP/p300 Acetylome. Cell *174*, 231-244.e12. 10.1016/j.cell.2018.04.033.

80. Fang, Z., Wang, X., Sun, X., Hu, W., and Miao, Q.R. (2021). The Role of Histone Protein Acetylation in Regulating Endothelial Function. Frontiers Cell Dev Biology *9*, 672447. 10.3389/fcell.2021.672447.

81. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EBNet Journal *10*.

82. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359. 10.1038/nmeth.1923.

83. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci Rep-uk *9*, 9354. 10.1038/s41598-019-45839-z.

84. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. 10.1093/bioinformatics/btw354.

85. Waskom, M. (2021). seaborn: statistical data visualization. J Open Source Softw *6*, 3021. 10.21105/joss.03021.

86. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160–W165. 10.1093/nar/gkw257.

87. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24–26. 10.1038/nbt.1754.

88. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and 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. 10.1016/j.molcel.2010.05.004.

89. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., et al. (2011). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *39*, D38–D51. 10.1093/nar/gkq1172.

90. George, S.S., Pimkin, M., and Paralkar, V.R. (2022). Customized genomes for human and mouse ribosomal DNA mapping. Biorxiv, 2022.11.10.514243. 10.1101/2022.11.10.514243.

91. McInnes, L., Healy, J., Saul, N., and Grossberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. The Journal of Open Source Software *3*, 861.

92. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat Methods *9*, 215–216. 10.1038/nmeth.1906.

93. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

94. Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R., and Bucher, P. (2017). The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic Acids Res *45*, D51–D55. 10.1093/nar/gkw1069.

95. Smit, A., Hubley, R., and Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org.

96. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.

97. Kluger, Y., Basri, R., Chang, J.T., and Gerstein, M. (2003). Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. Genome Res *13*, 703–716. 10.1101/gr.648603.

98. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods *17*, 261–272. 10.1038/s41592-019-0686-2.

99. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2022). KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res *51*, D587–D592. 10.1093/nar/gkac963.

100. Court, F., and Arnaud, P. (2016). An annotated list of bivalent chromatin regions in human ES cells: a new tool for cancer epigenetic research. Oncotarget *8*, 4110–4124. 10.18632/oncotarget.13746.

101. Lidschreiber, K., Jung, L.A., Emde, H., Dave, K., Taipale, J., Cramer, P., and Lidschreiber, M. (2021). Transcriptionally active enhancers in human cancer cells. Mol Syst Biol *17*, e9873. 10.15252/msb.20209873.

102. Jain, A., Nandakumar, K., and Ross, A. (2005). Score normalization in multimodal biometric systems. Pattern Recogn *38*, 2270–2285. 10.1016/j.patcog.2005.01.012.

103. Cieślik, M., and Bekiranov, S. (2014). Combinatorial epigenetic patterns as quantitative predictors of chromatin biology. Bmc Genomics *15*, 76. 10.1186/1471-2164-15-76.

104. Zitnik, M., and Zupan, B. (2018). NIMFA: A Python Library for Nonnegative Matrix Factorization. Arxiv. 10.48550/arxiv.1808.01743.

105. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.-H., Brown, M., Zhang, X., Meyer, C.A., et al. (2018). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res *47*, gky1094-. 10.1093/nar/gky1094.

C h a p t e r 3

SIMULTANEOUS MAPPING OF 3D STRUCTURE AND NASCENT RNAS ARGUES AGAINST NUCLEAR COMPARTMENTS THAT PRECLUDE TRANSCRIPTION

Isabel N. Goronzy Sofia A. Quinodoz, Joanna W. Jachowicz, Noah Ollikainen, Prashant Bhat, Mitchell Guttman

A modified version of this chapter was published as “Simultaneous mapping of 3D structure and nascent RNAs argues against nuclear compartments that preclude transcription” in *Cell Reports*. 41(9):111730. 2022. doi: 10.1016/j.celrep.2022.111730.

3.1. SUMMARY

Mammalian genomes are organized into three-dimensional DNA structures called A/B compartments which are associated with transcriptional activity/inactivity. However, whether these structures are simply correlated with gene expression or are permissive/impermissive to transcription has remained largely unknown because we lack methods to measure DNA organization and transcription simultaneously. Recently, we developed RNA&DNA (RD)-SPRITE, which enables genome-wide measurements of the spatial organization of RNA and DNA. Here we show that RD-SPRITE measures genomic structure surrounding nascent pre-mRNAs and maps their spatial contacts. We find that transcription occurs within B compartments — with multiple active genes simultaneously colocalizing within the same B compartment — and at genes proximal to nucleoli. These results suggest that localization near or within nuclear structures thought to be inactive does not preclude transcription and that active transcription can occur throughout the nucleus. In general, we anticipate RD-SPRITE will be a powerful tool for exploring relationships between genome structure and transcription.

3.2. INTRODUCTION

The three-dimensional (3D) arrangement of DNA in the nucleus is thought to be important for regulating critical nuclear processes such as DNA replication and transcription[1–3]. Accordingly, there have been significant efforts to map DNA structure across different cell types using proximity-ligation methods like 3C[4], Hi-C[5–7] and related variants[2,8]. These methods have identified several structural features including chromosome territories, A/B compartments, topologically associating domains (TADs), loops, and promoter-enhancer interactions. However, which of these are critical for gene regulation and other cellular functions remains unclear.

The main reason that structure-function relationships within the nucleus are poorly understood is that current methods cannot simultaneously measure transcriptional states and 3D genome organization[9,10]. Instead, analysis of the functional consequences of nuclear structure relies on correlations between distinct measurements of DNA organization and gene expression profiles generated from a combination of experimental methods (e.g. Hi-C and RNA-Seq) in different populations of cells. These measurements capture an ensemble of many individual cells, each of which may contain heterogenous functional states and structures, making the direct comparison between 3D structure and transcription challenging[5,11].

To highlight this limitation, consider A and B compartments, which refer to alternating sets of DNA regions that broadly partition chromosomes; DNA regions within one compartment preferentially interact with each other (e.g., A-A) rather than with neighboring regions of the other (e.g., A-B). Early studies found that A compartments are enriched for genomic DNA regions containing actively transcribed RNA Polymerase II (Pol II) genes, whereas B compartments are depleted for active Pol II genes and enriched for repressive chromatin marks[7,12]. As such, these compartments are generally thought to represent spatial organization of transcriptionally active (A) and inactive (B) Pol II genes within distinct regions of the nucleus[3,10,13–15].

In contrast to this general observation, there are specific genes located within B compartments that are actively transcribed[12]. This is predominantly explained by a model where actively transcribed DNA loci "loop out" of the inactive (B) compartment to localize within an active (A) compartment[1,16–19]. In this model, Pol II transcription does not occur within B compartments; actively transcribed genes may appear to be within them simply because of the ensemble nature of compartment (e.g. Hi-C, SPRITE) and gene expression measurements (e.g., RNA-seq). In support of this "looping out" model, single cell microscopy measurements have shown that individual active genes can be located away from the remainder of the chromosome from which they are transcribed[16,18–20], that the promoter regions of active genes in B compartments can form local associations with the A compartment[21], and that transcribed genomic loci (measured by interactions between pre-mRNAs) do not form A/B compartments[16].

Yet, there are other observations to suggest that transcription may occur within both A and B compartments: many A/B compartment boundaries remain the same between distinct cell states despite major changes in gene expression programs (Dixon et al., 2015), and direct recruitment of various gene loci to the nuclear lamina (a compartment associated with transcriptional silencing and located within B compartments) does not always lead to transcriptional repression for all genes[22–25]. Accordingly, whether localization of genes within B compartments or other nuclear structures that have been associated with inactive Pol II transcription and repressive heterochromatin (such as the nucleolus and nuclear lamina)[22,26,27] precludes Pol II transcription or is simply correlated with inactive transcription remains unclear.

Recently, we developed RNA & DNA SPRITE (RD-SPRITE), which enables simultaneous multi-way measurements of DNA and RNA organization in the nucleus[28]. In our previous study, we focused on the spatial localization of ncRNAs and their roles in seeding nuclear organization. However, RD-SPRITE also measures localization of mRNAs, including individual nascent pre-mRNAs at their transcriptional loci. Because RNA represents the functional output of transcription, this approach allows us to directly measure both 3D genome organization and transcription at the same location within the

nucleus. Here, we show that RD-SPRITE can be used to assess the relationship between structural organization and transcriptional activity within different structural compartments.

3.3. RESULTS

RD-SPRITE measures nascent and mature mRNAs at precise locations in the cell

RD-SPRITE uses split-and-pool barcoding to measure the spatial organization of individual RNA and DNA molecules within the cell. The fundamental measurement unit of RD-SPRITE is the SPRITE cluster, which contains multiple RNA and DNA molecules that are in close proximity within a single cell[28,29]. Using these clusters, we can measure multiway RNA and DNA contacts, including RNA-RNA, RNA-DNA, and DNA-DNA contacts, within higher-order structures in the cell (**Figure 1A**). We previously showed that RD-SPRITE can accurately measure the 3D spatial organization of DNA and RNA in the nucleus, including DNA structures such as chromosome territories, A/B compartments, TADs, and loops as well as DNA and RNA within nuclear bodies such as the nucleolus, nuclear speckles, and histone locus body[28,30].

Here, we sought to explore whether RD-SPRITE can measure the 3D organization of distinct populations of mRNAs — including nascent and mature mRNAs — and their quantitative levels at various locations in the cell. To do this, we examined the RNA-RNA and RNA-DNA contacts in our RD-SPRITE dataset collected from mouse embryonic stem cells. Specifically, we focused on intronic reads as a surrogate for nascent pre-mRNAs and exonic reads as a surrogate for mature mRNAs. We reasoned that newly transcribed (nascent) pre-mRNAs should be preferentially located on chromatin in proximity to their genomic DNA locus, while fully spliced (mature), mRNAs should be associated with ribosomal RNAs (rRNAs) in the cytoplasm. Consistent with this, we find that intronic reads in RD-SPRITE represent nascent pre-mRNAs in that they are (i) enriched on chromatin, (ii) enriched for contacts with various small nuclear RNAs (snRNAs) such as U1 and U2,

which are involved in pre-mRNA splicing in the nucleus, and (iii) depleted for contacts with cytoplasmic RNAs, such as rRNAs (**Figure 1B**). In contrast, exonic reads show properties consistent with mature mRNAs in that they are: (i) depleted on chromatin, (ii) depleted for contacts with snRNAs, and (iii) enriched for contacts with rRNAs (**Figure 1B**). Together, these data demonstrate that RD-SPRITE can detect both classes of mRNAs located in different parts of the cell and distinguish between their localization patterns.

We next tested whether RD-SPRITE can quantitatively measure the relative abundance of these distinct mRNA populations. First, we measured whether the overall RNA levels measured by RD-SPRITE correlate with total RNA-Seq measurements and observed a strong correlation between the levels of RNAs measured by each approach (spearman p=0.79) (**Figure 1C**)[31]. Next, we focused specifically on nascent pre-mRNA levels by comparing transcription levels estimated from intronic reads in RD-SPRITE and found them to be highly correlated with those estimated from global run-on and sequencing (GRO-Seq) assays[32], which measure transcription levels of mRNAs (spearman p=0.86). Finally, we observed a strong correlation between exonic reads measured by RD-SPRITE and mature mRNA levels measured by polyA-selected RNA-Seq (spearman p=0.88).

To confirm the localization of nascent RNAs at their genomic loci, we measured the DNA contacts of pre-mRNAs (RNA-DNA contacts) and found them to be enriched for contacts with their genomic loci (**Figure 1D-E**). Next, we explored whether RD-SPRITE can detect the 3D structure at these actively transcribing DNA loci. To do this, we mapped the DNA-DNA contacts of SPRITE clusters containing a specific nascent pre-mRNA and found that the DNA contacts are highly enriched surrounding the locus from which the pre-mRNA is transcribed (**Figure 1F**).

Taken together, these results demonstrate that RD-SPRITE accurately distinguishes distinct populations of mRNAs within the cell, enables quantitative measurement of their transcription levels, and detects the genomic contacts and 3D structure around individual pre-mRNAs.

Genomic DNA located within B compartments can be actively transcribed

Because RD-SPRITE accurately measures both nascent RNA transcripts and higher-order DNA organization genome-wide, we used it to explore the global structure of genomic DNA regions undergoing Pol II transcription. Specifically, we generated a genome-wide DNA-DNA contact matrix using SPRITE clusters containing nascent pre-mRNAs. We reasoned that if most genes within B compartments loop out and reposition into A compartments when actively transcribed (the "looping out" model), then we would see a single active compartment in the DNA-DNA contact matrix of actively transcribed regions. Conversely, if genes are transcribed within B compartments, then we would observe both A and B compartments within this DNA-DNA contact matrix (**Figure 2A**). In fact, the genomic DNA structures generated from only actively transcribed clusters show clear chromosome territories and intra-compartment structures comparable to those observed when measuring DNA-DNA contacts across all SPRITE clusters (**Figure 2B**). The A/B compartment structure seen in transcribed clusters closely corresponds to A/B compartments defined using principal eigenvector analysis on the DNA contacts measured from all SPRITE clusters (**Figure S1,** see **Methods)**. This suggests that genes in the B compartment do not "loop out" as they are transcribed but instead remain in the B compartment.

While it is commonly described as a single compartment, the B compartment is in fact heterogenous. Compartment structures can also be defined using 5-subcompartments, three of which (B1, B2, B3) are considered B-like but differ in repressive chromatin modifications, gene density, and nuclear location[33,34] (**Figure 2C**); B2 and B3 are highly enriched for chromatin features associated with transcriptional repression while B1 has chromatin features more closely resembling the A2 sub-compartment. Because of this, we considered the possibility that our observations of transcription within the B compartment might be restricted to B1. To explore this, we focused on a set of highly-expressed nascent pre-mRNAs in RD-SPRITE and found these genes to be located within all three B sub-

compartments (**Figure 2C,** see **Methods**). Focusing specifically on the sub-compartments associated with repressive features (B2 or B3), we measured the DNA-organization (DNA-DNA contacts) when pre-mRNAs are actively transcribed. We selected individual SPRITE clusters that contain reads for nascent pre-mRNAs located within B2 or B3 and generated a DNA-DNA heatmap (**Figure 2D**). We found that actively transcribed genomic regions within these sub-compartments maintain DNA-DNA contacts with other B2 and B3 regions and do not contact neighboring A1 sub-compartment genomic regions. Conversely, when we used clusters containing pre-mRNAs from genes within the A1 sub-compartment to generate a DNA-DNA heatmap, we observed preferential contacts with other A1 regions but not contacts with neighboring B compartment DNA regions. Together, these results demonstrate that active transcription can occur within all B sub-compartments.

To further validate that B compartment structures are observed when B compartment genes are actively transcribed, we explored the DNA contacts of nascent pre-mRNAs. RD-SPRITE detects long-distance RNA-DNA interactions between nascent pre-mRNAs and genomic DNA sites beyond their transcriptional loci (**Figure 1E**). To investigate the underlying genomic DNA structure during active transcription of B compartment genes, we looked for A/B compartment structures in these long-range RNA-DNA contacts. Indeed, beyond RNA-DNA contacts between pre-mRNAs and their own loci, B compartment pre-mRNAs are enriched for contacts with DNA regions located in neighboring B compartments and depleted for contacts with DNA regions located within A compartments (**Figure 2E-G**). Because these nascent RNA transcripts are located near their gene locus, this confirms that genes contained within B compartments do not “loop out” when transcribed.

Together, these results indicate that localization of genes within B compartments does not preclude transcription.

Nascent pre-mRNAs organize within genome-wide structures resembling A/B compartments

We next wondered whether multiple, simultaneously transcribed genes organize together within the B compartment. To explore this, we generated an RNA-RNA contact matrix to measure the genome-wide spatial organization of nascent pre-mRNAs (**Figures 3A-B**). Because the number of observed RNA contacts is dependent on expression level, we focused on the 2000 most highly expressed genes to ensure high-confidence measurements of individual pre-mRNA contacts (**Table S1**). These highly expressed genes include those located within both A and B compartments (1216 A genes and 784 B genes) and display comparable expression levels. (**Figure S1C, S2A**). We sorted these pre-mRNAs by the genomic position of their gene locus and observed clear structural patterns, including: (i) preferential contacts between pre-mRNAs that are transcribed from the same chromosome reminiscent of chromosome territories (**Figure 3A**); and (ii) alternating blocks of highly interacting pre-mRNAs within individual chromosomes reminiscent of A/B compartments (**Figure 3B**). In contrast, contact matrices generated between mature mRNAs (exons) do not display preferential contact frequencies based on their genomic positions, consistent with their localization in the cytoplasm (**Figure S3A-B**).

To determine whether these intrachromosomal structural patterns correspond to A/B compartments, we compared them to the 3D structure of their corresponding genomic DNA loci. We generated a DNA-DNA contact matrix for these highly expressed genes (gene-level heatmap) and observed highly similar intrachromosomal patterns in the DNA-DNA and the pre-mRNA RNA-RNA contact maps (Pearson r = 0.83), but not between gene-level DNA and mature mRNA contact maps (Pearson r = 0.04) (**Figure 3C, S3C**). Next, we defined A/B compartments using nascent RNA-RNA contacts and asked whether their quantitative (eigenvector) values matched those defined using DNA-DNA contacts. First, we ensured that A/B compartment scores based on the gene-level DNA-DNA contacts were similar to those measured across the genome (Pearson r = 0.87, see **Methods**) to confirm that this gene-level analysis is comparable to genome-wide analysis (**Figure 3D**). Second, we compared the gene-level DNA-DNA eigenvectors to those calculated from the

nascent RNA-RNA contact matrix and found a strong correlation (Pearson r = 0.75) (**Figure 3E**). Finally, we grouped RNA-RNA contacts based on A/B compartment definitions from genomic DNA and found that pre-mRNAs transcribed from B compartments display a high contact frequency with other pre-mRNAs transcribed from B compartments, but not with pre-mRNAs transcribed from loci contained within A compartments, and vice versa (**Figure 3F**). In contrast, mature mRNAs do not display any preferential interactions between A/B regions (**Figure S3D-F**).

To ensure that the observed compartmentalization of nascent RNAs is not a unique feature of highly expressed genes, we explored compartmentalization properties across mRNAs that span a broad range of expression levels (i.e., the top 10,000 most abundant pre-mRNAs) (**Figure S4A)** and observed preferential A-A and B-B contacts and depletion of neighboring A-B contacts (**Figure S4B)**. Indeed, zooming-in on chromosome 2, we detected clear B-A-B compartment structures, comparable to those measured for the most abundant 2000 pre-mRNAs, within the RNA-RNA contacts of lower expression genes (**Figure S4C**). This indicates that the organization of pre-mRNAs within A/B compartments and transcription within the B compartment is observed across a range of expression levels and all classes of transcribed Pol II genes.

These results are consistent with our observations that actively transcribed genes are spatially organized into A/B compartments and that multiple genes are simultaneously transcribed within B compartments (**Figure 3G**). If transcription only occurred in a single active compartment, we would expect nascent RNAs to globally interact with each other (**Figure S3G)**. Instead, our RNA-RNA heatmaps clearly demonstrate that compartmentalization occurs among nascent transcripts; we observe distinct groups of pre-mRNAs interacting with each other while excluding other nearby transcripts. Importantly, this pre-mRNA compartmentalization, which closely matches the corresponding A/B compartment definitions of DNA, is observed in the RNA-RNA contacts *a priori*, independent of any compartment calls from DNA-DNA contacts.

To confirm these observations using an orthogonal assay, we performed RNA-FISH and measured whether B compartment pre-mRNAs interact more closely with each other than with pre-mRNAs in neighboring A compartments. Specifically, we generated probes against introns of 6 pre-mRNAs within chromosome 2; each set of three probes corresponded to two mRNAs from distinct B compartments and one from an intervening A compartment (**Figure 3H**). Consistent with our RD-SPRITE measurements, we find that the pre-mRNAs transcribed from the two B compartments are closer in 3D space than they are to the pre-mRNA transcribed from the A compartment. This occurs even though the two genes in the B-compartments (B-B pairs) are farther apart in linear space than the A-B pairs (**Figure 3I-J**).

Together, these results indicate that nascent pre-mRNA transcripts from genes located in both the A and B compartments organize into structures such as chromosomal territories and A/B compartment structures. This highlights the power of RD-SPRITE and its ability to measure long-distance interactions of nascent pre-mRNAs genome-wide to uncover the spatial organization of RNA in the nucleus.

Transcription of RNA Pol II genes can occur in proximity to the nucleolus

We next explored whether Pol II transcription occurs near the nucleolus, a nuclear body that is organized around active transcription and processing of RNA Polymerase I (Pol I) transcribed pre-ribosomal RNAs[35,36]. Previous studies have shown that genomic DNA regions positioned near the nucleolus are associated with inactive Pol II transcription and heterochromatin marks[26,27,37,38]. However, whether proximity to the nucleolus is simply correlated with inactive transcription or whether organization around the nucleolus precludes Pol II transcription remains unknown.

To explore this, we utilized the ability of RD-SPRITE to measure long-range RNA and DNA organization around the nucleolus (**Figure 4A**)[28,30]. We reasoned that if proximity to the nucleolus precludes transcription, DNA regions would loop away when they are

transcribed (**Figure 4B**) and this would result in SPRITE clusters containing nascent pre-mRNAs depleted for nucleolar contacts. In contrast, if transcription can occur near the nucleolus, we would detect preferential contacts between nascent pre-mRNAs of nucleolar-proximal genes and nucleolar RNAs.

First, we defined the genomic DNA regions proximal to the nucleolus (the "nucleolar hub") based on DNA contact frequency with nucleolar RNAs, such as 45S pre-ribosomal RNAs (rRNAs) and small nucleolar RNAs (snoRNAs), and inter-chromosomal DNA-DNA contacts (**Table S2**). We previously showed that these genomic DNA regions are proximal to the nucleolus[30] (see **Methods**, **Figure 4C, S5A**). Next, we analyzed whether nascent pre-mRNAs transcribed from these nucleolar-proximal DNA loci co-occur in SPRITE clusters with snoRNAs, suggesting that they are transcribed when they are physically close to the nucleolus and do not "loop out" during transcription. Indeed, pre-mRNAs from nucleolar-proximal genes display strong enrichment for snoRNA contacts, whereas nascent pre-mRNAs transcribed from nucleolar-distal genes exhibit few snoRNA contacts (**Figure 4D-E**). In fact, the frequency of snoRNA to pre-mRNA contacts was positively correlated with the nucleolar proximity of the pre-mRNA's genomic locus (Pearson r = 0.75; **Figure 4F, S5B**), while the transcriptional levels of the pre-mRNAs were not (Pearson r = -0.02; **Figure S2B**). This suggests that RNA Pol II transcription can occur close to the nucleolus.

To explore the underlying genomic DNA structure of nucleolar-proximal genes when they are actively transcribed, we measured the RNA-DNA contacts for these pre-mRNAs (**Figure 4G**). We reasoned that if these DNA loci loop away from the nucleolus when they are transcribed, nucleolar proximal pre-mRNAs would exhibit reduced interactions with nucleolar hub DNA regions and increased interactions with neighboring non-nucleolar hub regions. Instead, we observed that nascent pre-mRNAs of these genes frequently contact other nucleolar-proximal DNA regions and are depleted at neighboring non-nucleolar hub regions.

Because our results suggest that genes are transcribed when they are near the nucleolus, we wondered whether multiple actively transcribed genes organize together around the

nucleolus. Specifically, we explored if nascent pre-mRNAs of nucleolar-associated genes display preferential inter-chromosomal contacts. To do this, we took the genome-wide inter-chromosomal contact matrix between all nascent pre-mRNAs (**Figure 3A**) and aggregated the mRNAs into three groups: speckle hub genes, nucleolar hub genes or neither (**Figure 4H, Table S2, S3)**. We observed enrichment of inter-chromosomal contacts between these nucleolar hub nascent pre-mRNAs (p-value < 0.01, see **Methods and Figure S5C**), suggesting that the nucleolar hub DNA regions from multiple chromosomes remain organized together in space during transcription and that genes from multiple chromosomes are simultaneously transcribed at the nucleolus.

To confirm this observation, we performed intron RNA FISH for two genes located on chromosome 19, one within the nucleolar hub (Carnmt1) and one far from it (Btrc, **Figure 4I, Video S1, S2**). Carnmt1 is actively transcribed while positioned adjacent to the nucleolus (**Figure 4I**). We measured the distance to the nucleolus for 41 alleles of each gene across 22 cells and observed that ~50% (21/41) of Carnmt1 alleles are transcribed within 0.1 μm of the nucleolar surface. Indeed, even when the allele is directly contacting the nucleolus (distance ≤ 0 μm), we observe transcription in ~1/3 of measured Carnmt1 alleles (**Figure 4J**). In contrast, we rarely observe Btrc transcribed near the nucleolus, even though it is located on the same chromosome as Carnmt1 (**Figure 4K**).

Together, these results demonstrate that proximity to the nucleolus does not preclude transcription of Pol II genes.

3.4. DISCUSSION

Here, we showed that RD-SPRITE enables simultaneous measurement of 3D DNA structure and nascent RNA transcription to map the DNA contacts of pre-mRNAs, the 3D structure of actively transcribed genomic loci, and the global organization of nascent pre-RNAs. Previously, the question of whether certain nuclear structures are impermissive to transcription was unresolved because we lacked the ability to map DNA and RNA contacts

simultaneously at high resolution across the genome. Existing methods are unable to do this because they either focus exclusively on DNA structure (e.g. Hi-C) or map RNA and DNA via proximity ligation (e.g. GRID-seq), which is limited to pairwise interactions and therefore cannot simultaneously measure 3D DNA structure and nascent RNA localization or RNA-RNA interactions between pre-mRNAs. Using RD-SPRITE, we can simultaneously measure RNA-RNA, RNA-DNA, and DNA-DNA contacts genome-wide and therefore generate global profiles of the 3D-structures associated with nascent transcripts.

We leveraged these features of RD-SPRITE to show that transcription of genomic DNA occurs within both A and B compartments as well as at genomic DNA regions that are proximal to the nucleolus. Our results demonstrate that DNA does not need to reposition into an "active" compartment in the nucleus to be transcribed and that gene localization within B compartments or near the nucleolus does not preclude Pol II transcription. Furthermore, we found that nascent pre-mRNAs — including those within A compartments, B compartments, and near the nucleolus — localize with other transcripts from their respective nuclear structures; this is reminiscent of DNA organization such as chromosomal territories, A/B compartments and inter-chromosomal interactions around the nucleolus. While we focused on exploration of inactive compartments, we note that this approach can also be used to explore other structural features and transcription, including enhancer-promoter contacts.

Our findings argue against the "looping out" model whereby active genes need to move out of inactive compartments to contact active compartments when transcribed. Previous evidence for this model came primarily from imaging studies which observed that individual DNA loci can loop away from their chromosome territories when transcriptionally active[18,39]. Additional studies using nascent RNA-FISH did not detect chromosome territories or compartment-like structures, suggesting that pre-mRNAs organized within a single active compartment[16]. It was therefore postulated that DNA structure detected by Hi-C and similar approaches may capture ensemble DNA organization across a population of cells rather than the organization of DNA loci that are

actively transcribed. However, these imaging approaches were not able to simultaneously measure both RNA and DNA with high sensitivity on a genome-wide scale; therefore, their inability to measure these structures likely reflects limited resolution rather than support for this model. Using RD-SPRITE, we can detect the long-range RNA-RNA interactions of lower abundance RNAs, such as nascent pre-mRNAs, which enables us to generate high-depth genome-wide RNA-RNA contact maps for thousands of RNAs.

Altogether, our results demonstrate that transcription can occur throughout the nucleus, including within regions that have typically been viewed as inactive. Thus, the simple idea of "active" and "inactive" compartments — distinct structural domains within the nucleus that are globally permissive or impermissive for transcription — is likely inaccurate. Consistent with this, previous studies have shown that transcription can occur at genes proximal to the nuclear lamina[22,40–42] and that Pol II can freely access the inactive X chromosome heterochromatin domain during X-chromosome inactivation[43].

Because spatial organization does not appear to dictate transcriptional state, arrangement of DNA into A/B compartments likely reflects other features of these genomic regions. Indeed, our study and others[9] suggest that transcription is unlikely to be the sole factor driving compartmentalization of the genome. Genomic DNA regions within transcriptionally inert sperm cells[9,44,45] as well as cells treated with various transcriptional inhibitors[46,47] are still partitioned into A/B compartments. Additionally, A/B compartments can be invariant across cell states — even when undergoing large-scale changes in gene expression[12]. One possibility is that these compartments reflect differential gene density: DNA regions contained within A compartments are generally gene dense, whereas those in B compartments are generally gene poor. This would explain why A/B compartments are correlated with transcriptional activity but may not regulate transcription state, because gene-dense regions are more likely to be transcriptionally active. Other possible contributors to compartmentalization include A/T sequence content of the genome, the prevalence of SINE and LINE elements[48], and the replication timing of DNA[49,50]. In fact, multiple studies have found that early and late replicating domains correspond to A and B compartments, respectively[49,51]. Yet another possibility is that these compartments reflect

patterns of histone modifications. While the precise features that drive compartment organization are unknown, our results suggest that these compartments do not define transcriptional state and additional work is needed to understand what role, if any, spatial compartments play in gene regulation.

Limitations of the study

Because RD-SPRITE does not quantify absolute physical 3D distances, our results do not directly measure how close actively transcribed genes are to each other or from structures such as the nucleolus. Instead, we assess the relative proximity between molecules by calculating contact frequency. While our results suggest that transcription can occur near the nucleolus, the measurements cannot determine whether transcription is occurring directly within — or at a defined distance from — the nucleolus. However, we have validated our observations for representative loci using FISH and find strong concordance between distances measured by microscopy and SPRITE data (here and in Quinodoz et al., 2018).

While our data suggests that large scale, global repositioning of B compartment genomic regions into an A compartment is not required for transcription, we cannot exclude the possibility that small scale, local structural reorganization may occur (e.g., promoter regions loop out and contact each other[21]) or that individual genes may relocate upon transcription. For instance, certain transcribed nucleolar genes may be located further from the nucleolus than their inactive counterparts but remain within the B compartment and in proximity to the nucleolus. Alternatively, multiple B compartment genes may undergo local structural changes to organize together when transcribed, while remaining distinct from other active A compartment genes. While these questions remain to be addressed, our data clearly indicate that genes remain compartmentalized when transcribed and do not reorganize into a single active compartment.

Finally, our study focused on mouse ES cells and therefore we cannot exclude the possibility that other cell types might display distinct properties. Furthermore, it remains possible that other spatial compartments in the nucleus that have not yet been studied might preclude Pol II transcription. Future work will be needed to comprehensively map transcriptional states and 3D genome structure in other cell types and extend these observations to additional cell-types and nuclear compartments.

3.5. MAIN FIGURES

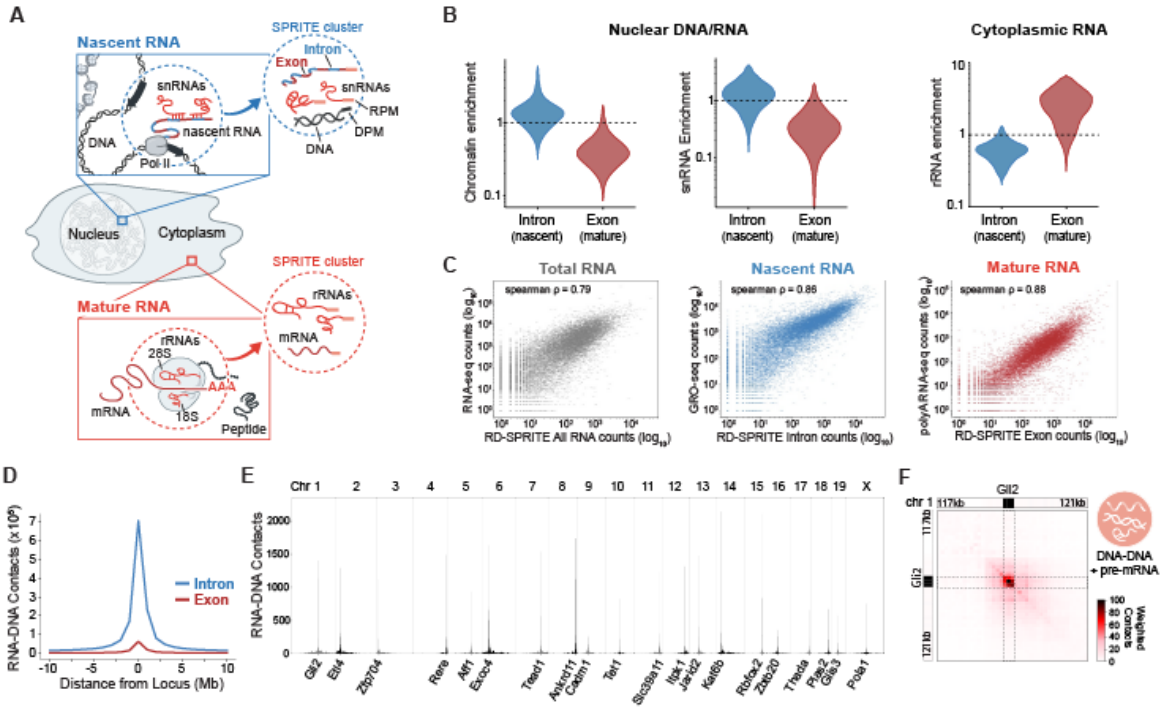


Figure 1: RD-SPRITE measures nascent and mature mRNAs at precise locations in the cell

(A) Schematic of nascent pre-mRNAs (blue), mature mRNAs (red), and their respective molecular interactions mapped using RNA-DNA SPRITE. Zoom-ins show nascent pre-mRNA contacts in the nucleus (top) and mature mRNA contacts in the cytoplasm (bottom). The specific RNA (RPM) or DNA (DPM) molecules measured within RD-SPRITE clusters are shown in the dotted circles. **(B)** Contact frequency enrichment scores of introns (blue) or exons (red) with chromatin (left), snRNAs (middle) or rRNAs (right) measured using RNA-DNA or RNA-RNA interactions. **(C)** Correlations between RD-SPRITE RNA abundance and total RNA-seq[31] (left), RD-SPRITE introns and GRO-seq[32] (middle), and RD-SPRITE exons and polyA-selected RNA-seq (right). **(D)** Aggregated total RNA-DNA contacts of introns or exons with DNA regions surrounding their genomic loci. Shown is the total weighted contact frequency of all RNAs within these populations contacting 1 megabase (Mb) genomic DNA windows from 10 Mbs up- and down-stream from the

transcriptional start site. **(E)** Examples of weighted RNA-DNA interactions for selected pre-mRNAs (1Mb resolution). The genomic locus for each pre-mRNA is annotated on the x-axis. **(F)** Weighted DNA-DNA interactions for transcriptionally active loci at the Gli2 gene locus on chromosome 1 (100 kb resolution). Interactions of transcriptionally active loci are defined as the DNA contacts occurring within multi-way SPRITE clusters containing both nascent Gli2 pre-mRNA transcripts and multiple DNA reads.

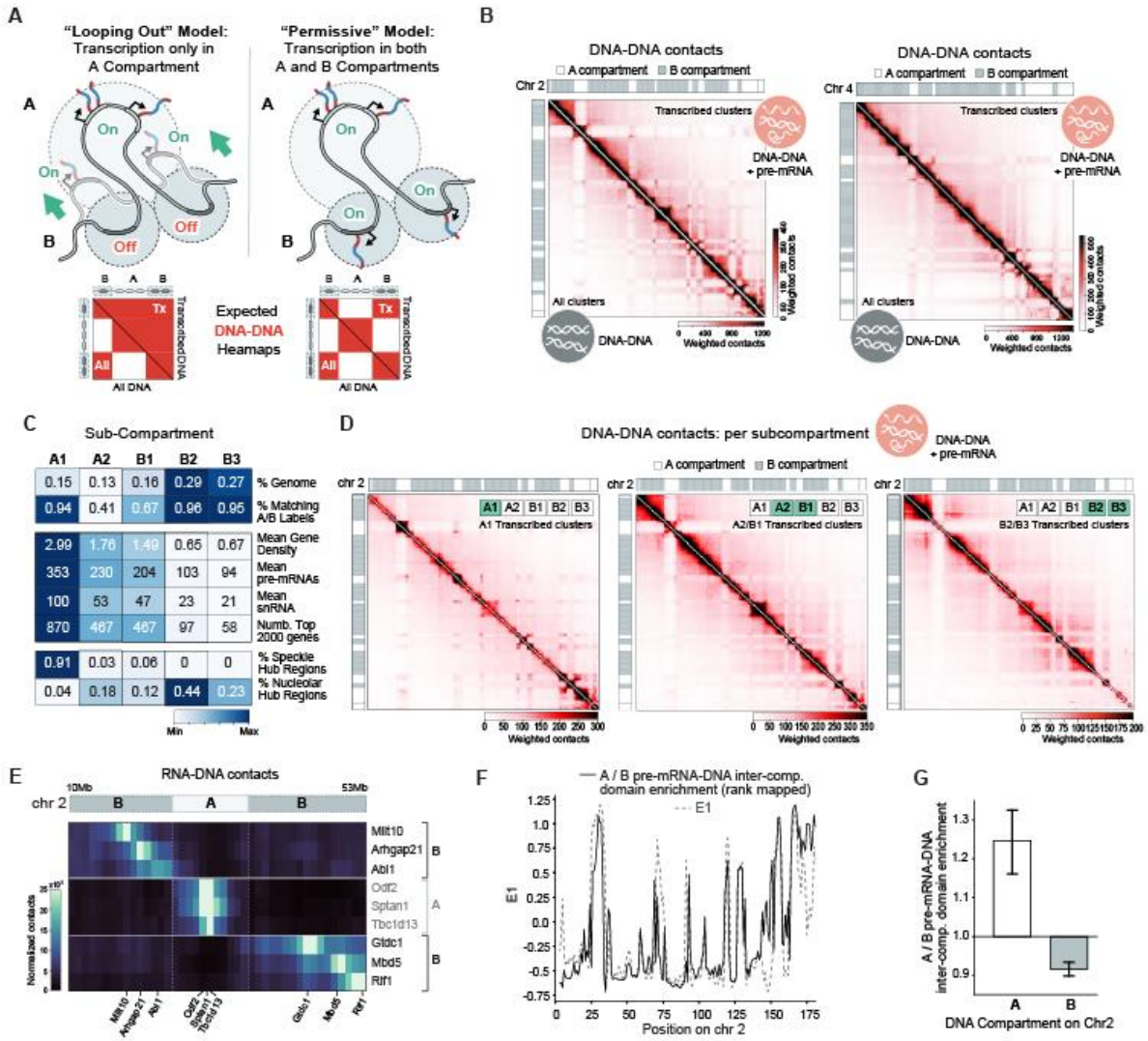


Figure 2: Genomic DNA located within B compartments can be actively transcribed.

(A) Two models of RNA Pol II gene transcription within A or B compartments and the expected DNA-DNA interaction matrices for actively transcribed loci. The "looping out" model requires B compartment genes to loop into the A compartment to be transcribed and the corresponding DNA-DNA matrix generated from transcribed DNA regions (left heatmap, upper diagonal) would not have compartment structure. In the "permissive" model, transcription of B compartment genes occurs without a change in genomic structure and the corresponding DNA-DNA matrix from transcribed DNA regions (right heatmap, upper diagonal) would have A/B compartment structure. **(B)** Weighted DNA-DNA

interaction heatmaps for SPRITE clusters containing nascent pre-mRNAs (upper diagonal) versus all SPRITE clusters (lower diagonal). Chromosomes 2 and 4 are shown as examples. **(C)** Feature profiles of A/B sub-compartments at 100 kilobase (kb) resolution. Characteristics include percentage (%) of genome assigned to each sub-compartment (top), % of 100 kb regions for each sub-compartment matching the corresponding "super-compartment" labels (i.e., A1-A, B1-B) calculated by principal eigenvector analysis of RD-SPRITE (second), mean number of protein-coding genes per 100 kb DNA region (third), mean weighted RNA-DNA contacts of pre-mRNAs per 100 kb DNA region (fourth), mean weighted RNA-DNA contacts of small nuclear RNAs (snRNAs) per 100 kb DNA region (fifth), number of top 2000 genes within each sub-compartment (sixth), and % of speckle or nucleolar hub regions within each sub-compartment (seventh and eighth). **(D)** Weighted DNA-DNA interaction heatmaps for actively transcribing SPRITE clusters containing nascent pre-mRNAs of genes in various sub-compartments. **(E)** Unweighted RNA-DNA interactions of nascent pre-mRNAs at the B-A-B compartment boundaries near the front end of chromosome 2 (10Mb-53Mb). **(F)** Inter-compartment pre-mRNA-DNA contact enrichment score for A versus B compartment genes (black solid line) and the first eigenvector (E1) (grey dotted line) along chromosome 2. Enrichment scores were rank-remapped to E1 for direct comparison (see **Methods**). **(G)** Mean inter-compartment pre-mRNA-DNA enrichment scores for A versus B compartment genes on A (left) or B (right) compartment genomic regions of chromosome 2. Error bars show 95% bootstrapped confidence intervals.

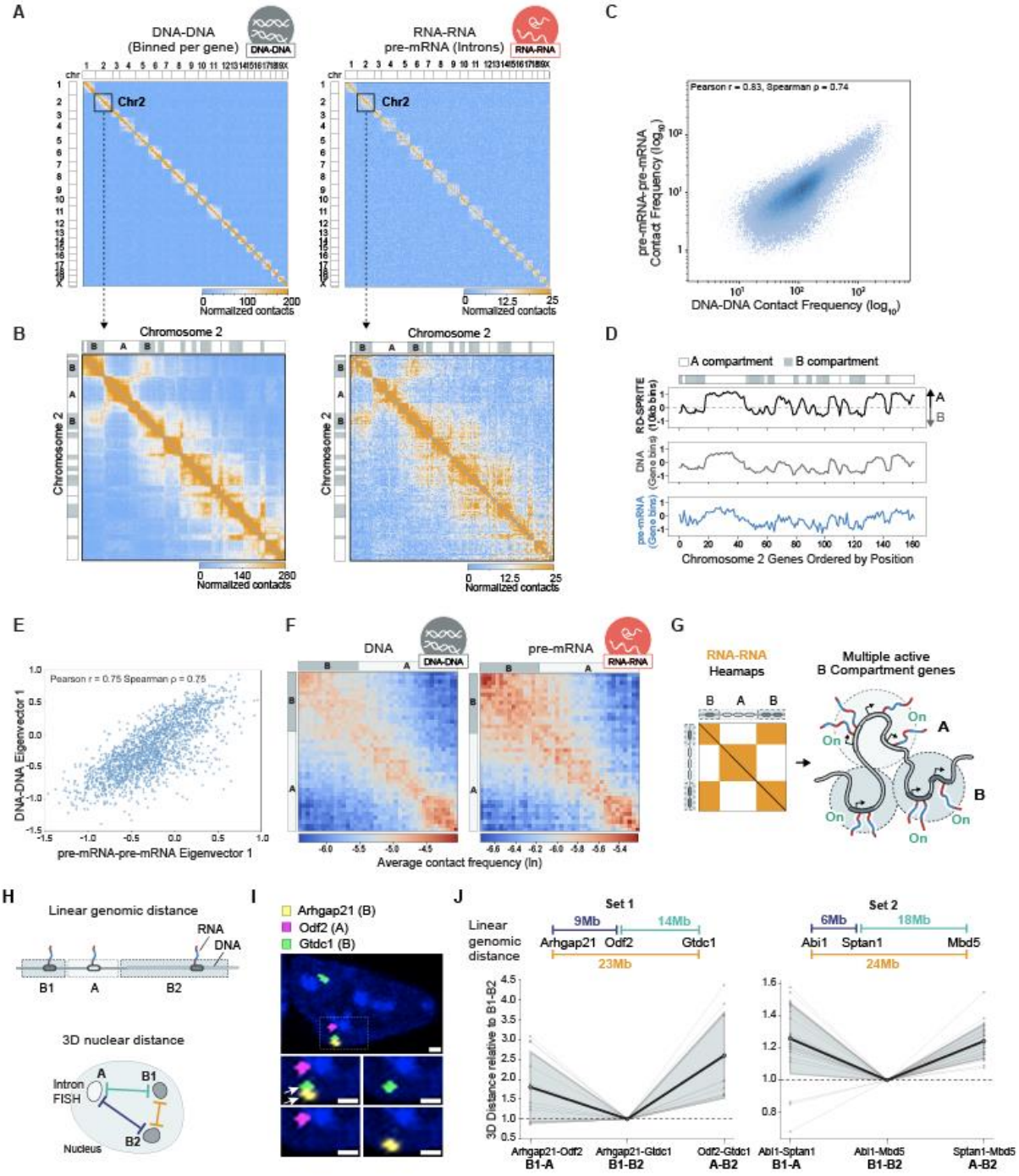


Figure 3: Nascent pre-mRNAs organize within genome-wide structures resembling A/B compartments.

(A) Gene-level DNA-DNA and nascent pre-mRNA RNA-RNA contact matrixes. Unweighted RNA-RNA contacts between the top 2000 expressed pre-mRNAs are shown (see **Methods**). DNA-DNA matrixes are binned by the genomic loci of genes used in the RNA-RNA matrix. Genes (and pre-mRNAs) are sorted based on their genomic position. **(B)** Zoom-in of gene-level DNA-DNA and nascent pre-mRNA RNA-RNA contact matrices for chromosome 2. **(C)** Correlation of genome-wide, intra-chromosomal contact frequencies for gene-level DNA-DNA (x-axis) and nascent pre-mRNA RNA-RNA (y-axis) contact matrices. **(D)** Comparison of the first eigenvector (E1) calculated from a genome-wide 10 kb-binned DNA-DNA contact matrix (top), gene-level binned DNA-DNA contact matrix (middle), and nascent pre-mRNA RNA-RNA contact matrix (bottom) along chromosome 2. A/B indicator bar along the top shows compartment assignments based on the value of the 10 kb-binned E1. **(E)** Correlation of E1 calculated from gene-level DNA-DNA (y-axis) and nascent pre-mRNA RNA-RNA (x-axis) contact matrices. **(F)** Saddle plots generated from the gene-level DNA-DNA and nascent pre-mRNA RNA-RNA contact matrices. Plots show the average interactions between groups of genes ordered by their compartment signals calculated from a 10 kb binned DNA-DNA matrix. A/B indicator bars along the axes indicate the compartments of the genes. **(G)** Model of RNA Pol II transcription of multiple genes within B compartments and the expected RNA-RNA interaction matrix. **(H)** Schematic of intron RNA-FISH design. Nascent transcripts from two B compartment genes located on opposite sides of an A compartment gene in linear genomic space (left) were probed and the 3D distance between pairs (right) was measured. **(I)** Representative microscopy image of intron RNA FISH for Arhgap21 (B compartment gene), Odf2 (A compartment gene), and Gtdc1 (B compartment gene). Both alleles of Gtdc1 are expressed. Transcripts from a single chromosome are boxed. Arrows highlight the B compartment genes coming together in 3D space. Scale bar is 1 μm. **(J)** Parallel coordinates plot of pairwise 3D distances measured by intron RNA-FISH (B1-A and A-B2). Distances were normalized to the B1-B2 pair distance for each cell to account differences in for cell size. Each gray line indicates a measurement from a single cell, bolded black lines indicate the mean, and shaded gray regions indicate the standard deviation. Probed genes and the linear genomic distances between each DNA loci are listed

on top of each plot. Measurements are from n = 12 cells containing Arhgap21, Gtdc1, and Odf2 triplets and n = 27 cells containing Abi1, Mbd5, and Sptan1 triplets.

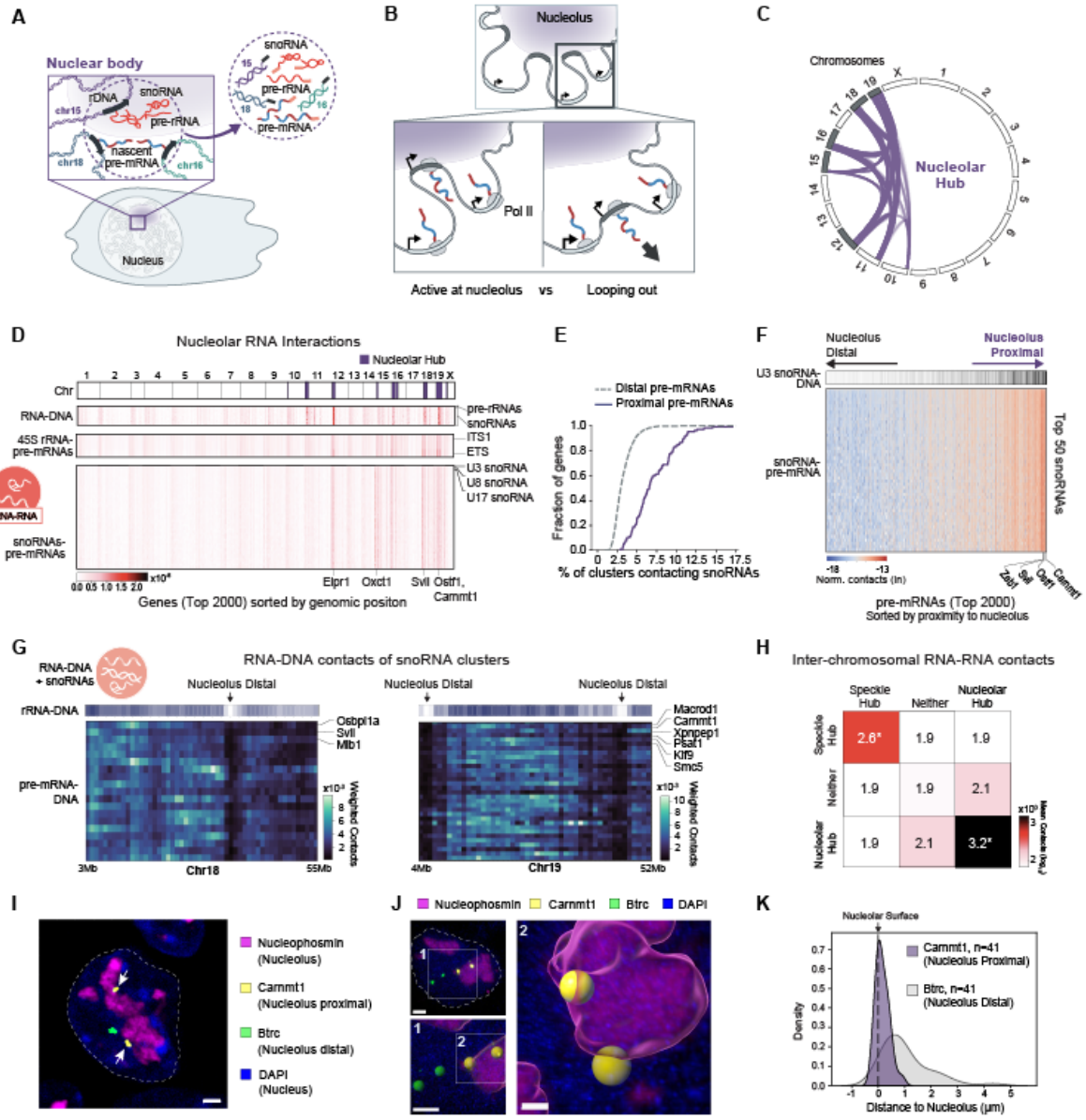


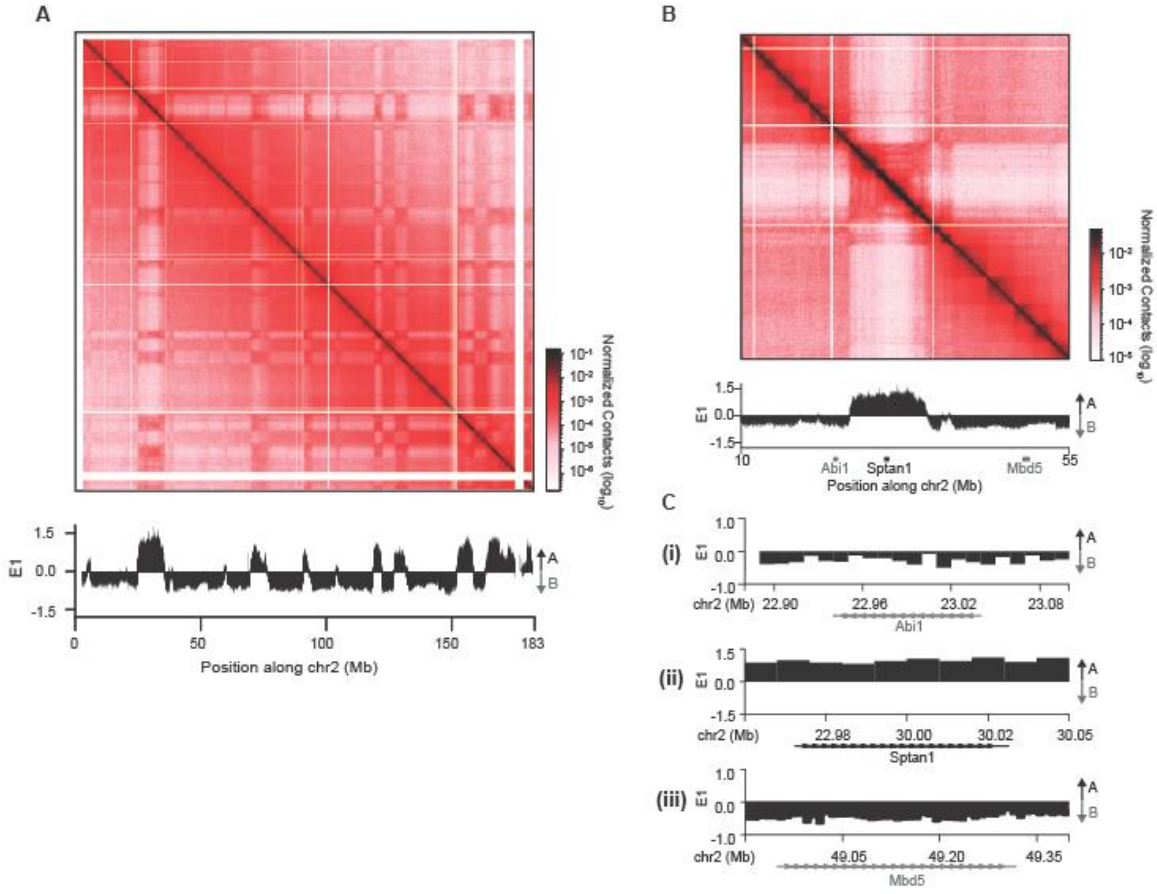
Figure 4: Transcription of RNA Polymerase II genes can occur in proximity to the nucleolus.

(A) Schematic of the molecular interactions occurring near the nucleolus and their corresponding RNA and DNA interactions measured within an RD-SPRITE cluster (circle). Because RD-SPRITE clusters can capture long-distance interactions, a single cluster can measure multiple interacting RNA (pre-mRNAs) and DNA molecules (genomic loci) around RNA bodies (containing 45S rRNA and snoRNAs). **(B)** Two models of RNA

Pol II gene transcription for nucleolar-proximal genes. **(C)** Diagram of inter-chromosomal DNA contacts between nucleolar hub regions. Chromosomes in gray contain ribosomal DNA genes. **(D)** RNA-DNA interactions (top box) corresponding to RNAs known to reside in the nucleolus (y-axis; snoRNAs and 45S pre-rRNAs) with genomic loci (x-axis; binned per gene) are shown along the top. RNA-RNA interactions (bottom 2 boxes) between the top 2000 expressed pre-mRNAs (x-axis; see **Methods**) and nucleolar RNAs (y-axis; 45S pre-rRNA and snoRNAs) are shown. Genes are ordered along the x-axis based on genomic position. Three components of 45S pre-rRNA spacers (ITS1, ITS2, 3'ETS) and the top 50 snoRNAs in descending order by contact frequency are along the y-axis. DNA loci within the nucleolar hub are annotated in purple. **(E)** Cumulative density of pre-mRNA contacts with snoRNAs for nucleolar proximal (purple) and nucleolar distal (gray) genes. **(F)** Nascent pre-mRNA – snoRNA contact matrix for the top 2000 genes. Genes are ordered based on their distance to the nucleolus, defined by contact frequency of the genomic locus to nucleolar hub regions in RD-SPRITE. Heatmap of U3 RNA-DNA density at the nucleolar distance corresponding to each gene is shown. **(G)** Unweighted pre-mRNA-DNA contacts occurring in SPRITE clusters containing snoRNAs for nucleolar genes of chromosome 18 (left) and 19 (right). 45S pre-rRNA density (RNA-DNA contact frequency) is shown as a heatmap, indicating nucleolar close (white) and far (purple) regions. **(H)** Average unweighted inter-chromosomal RNA-RNA contacts of the top 2000 nascent pre-mRNAs grouped by hub (speckle hub, nucleolar hub, neither). P-values were calculated relative to an expected distribution generated by randomizing RNAs reads across SPRITE clusters and calculating the resulting inter-chromosomal RNA-RNA contact frequency (see **Methods**). **(I)** Immunofluorescence (IF) combined with intron RNA-FISH for two genes on chromosome 19: Carnmt1 (yellow), a nucleolar-proximal gene, and Btrc (green), a nucleolar-distal gene. Both alleles of Carnmt1 are transcribed while located adjacent to the nucleolus (Nucleolin; purple). Nucleus is demarcated with DAPI. Arrows highlight the nucleolar-proximal pre-mRNAs located adjacent to nucleoli. Scale bar is 2 μm. **(J)** 3D surface representation of intron RNA-FISH for Carnmt1 (yellow) and Btrc (green) and IF for Nucleolin (purple). Zoom-out (upper left, scale bar = 2 μm) shows original FISH and IF signals in the entire cell. Zoom-ins show spheres

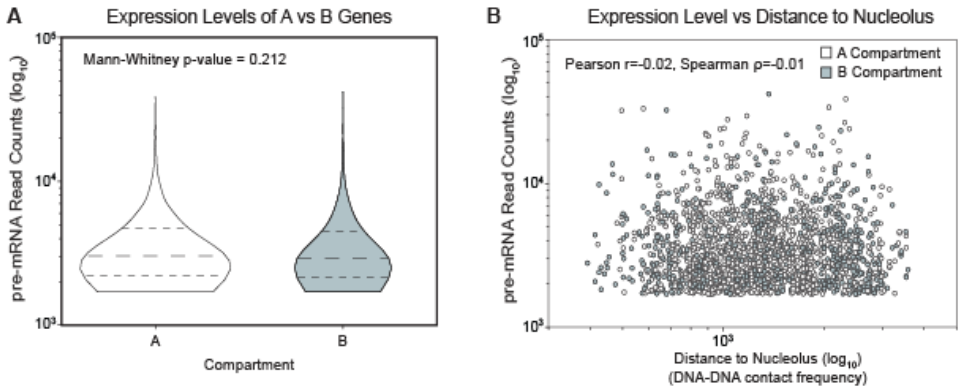
corresponding to FISH signal and nucleolar surface (lower left, scale bar = 2 μm) and the two transcribed Carnmt1 alleles intersecting the nucleolar surface (right, scale bar = 0.7 μm). **(K)** Distribution of 3D distances between the nucleolar surface and Carnmt1 (purple) or Btrc (grey) nascent transcripts quantified from intron RNA-FISH and IF images (n = 22 cells).

3.6. SUPPLEMENTAL FIGURES

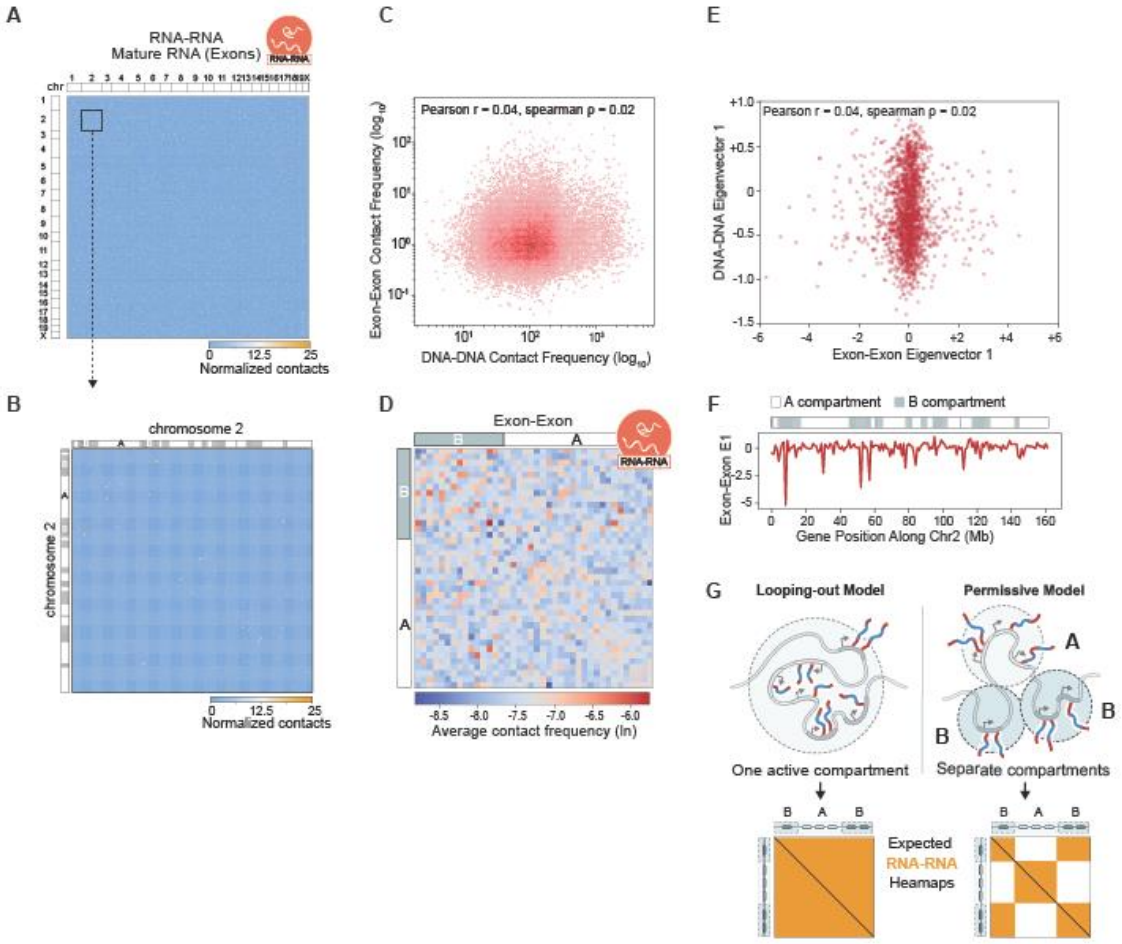


Supplemental Figure 1: Eigenvector 1 (E1) and A/B compartment assignments calculated from RD-SPRITE clusters, corresponding to Figure 2 and Figure 3. (A) Weighted DNA-DNA contact matrix at 100kb resolution (top) and eigenvector 1 (E1) at 10kb resolution (bottom) along chromosome 2. Both the contact matrix and E1 were calculated from RD-SPRITE clusters containing 2-1000 DNA reads (see **Methods**). **(B)** Zoom-in of the 100kb DNA-DNA contact matrix and E1 from (A) near the front of chromosome 2. The positions of one of the A (black) and B (gray) compartment gene triplicates measured by RNA-FISH in Figure 3 (Abi1, Sptan1, and Mbd5) are annotated. **(C)** Zoom-in of E1 at 10kb resolution near the gene annotations of the three RNA FISH-measured genes in (B). (i) Abi1 — a B compartment gene, (ii) Sptan1 — an A compartment

gene, and (iii) Mbd5 – a B compartment gene. The compartments for each gene were assigned based on the sign of E1 (A compartment > 0; B compartment < 0).



Supplemental Figure 2: Expression level profiles of top 2000 expressed introns, corresponding to Figure 3 and Figure 4. (A) Violin plot of read counts of selected, top 2000 expressed introns grouped by A/B compartment assignment of the individual genes. Median and quartiles are shown with dotted lines. The bottom of each violin was set to the lowest read count. Mann-Whitney p-value (top) shows no significant difference in the distributions of read counts between A and B compartment genes. **(B)** Scatterplot of read counts versus genomic loci distance to nucleolus for top 2000 expressed introns. Distance to nucleolus was calculated based on DNA-DNA contact frequency of the DNA region containing the gene locus and nucleolar hub regions (see **Methods**). Neither Pearson nor Spearman statistics show a correlation between counts and distance.



Supplemental Figure 3: RNA-RNA interaction matrix of mature mRNAs (exons), corresponding to Figure 3. (A) Mature mRNA (exon) gene-level RNA-RNA contact matrix for the exons of genes corresponding to the top 2000 expressed introns, analogous to **Figure 3A**. Genes are sorted based on their genomic position. Chromosomes are annotated along the top and left axes. **(B)** Zoom-in of mature mRNA (exon) RNA-RNA contact matrix for chromosome 2. **(C)** Correlation of genome-wide, intra-chromosome contact frequencies for gene-level DNA-DNA (x-axis) and mature mRNA (exon) RNA-RNA (y-axis) contact matrices. **(D)** Saddle plots generated from mature mRNA (exon) RNA-RNA contact matrix, analogous to **Figure 3F**. Plot shows the average interactions between groups of genes ordered by their compartment signals calculated from a 10 kb-binned DNA-DNA matrix. A/B indicator bar along the axes indicate the compartment

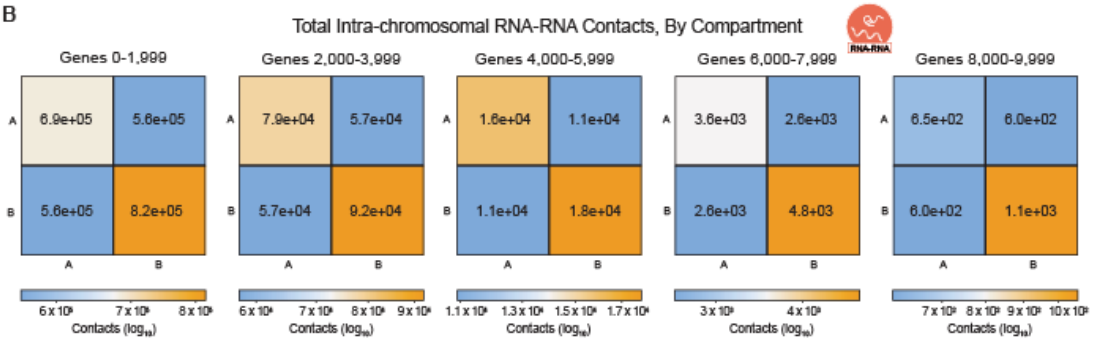
assignments of the genes. **(E)** Correlation of E1 calculated from gene-level DNA-DNA (y-axis) and mature mRNA (exon) RNA-RNA (x-axis) contact matrices. **(F)** E1 calculated from a mature mRNA (exon) RNA-RNA contact matrix along chromosome 2. A/B indicator bar along the top shows compartment assignments based on the sign of E1 generated from a 10 kb-binned DNA-DNA heatmap. **(G)** Schematic of the "looping out" model and the corresponding predicted RNA-RNA matrix. If transcription only occurs in the A compartment, nascent transcripts of both A and B compartment genes would interact within a single "active compartment" and there would be no observable compartmentalized structure in an RNA-RNA contact matrix.

A

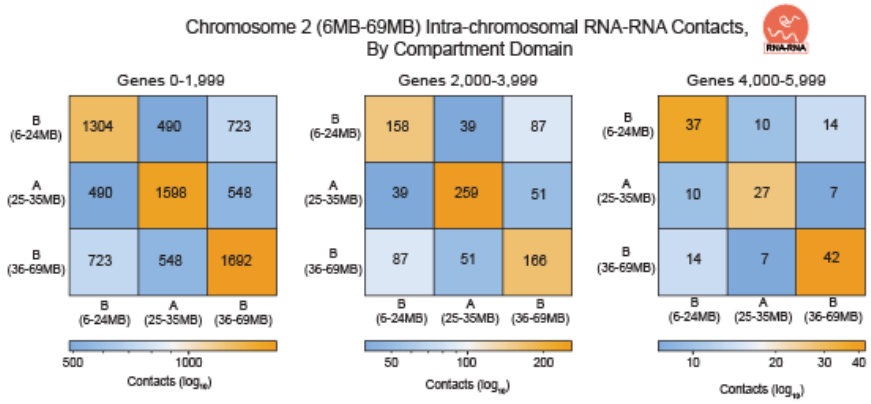
Gene Expression Groups Summary Statistics

Expression Group	Total Expression (Num. Reads)	Maximum Expression (Num. Reads)	Minimum Expression (Num. Reads)	Number of A Compartment Genes	Number of B Compartment Genes
0-1,999	8,421,550	41,983	1731	1214	786
2,000-3,999	2,271,638	1,730	726	1164	836
4,000-5,999	1,001,154	725	338	1220	780
6,000-7,999	468,427	338	151	1274	726
8,000-9,999	200,253	151	61	1264	736

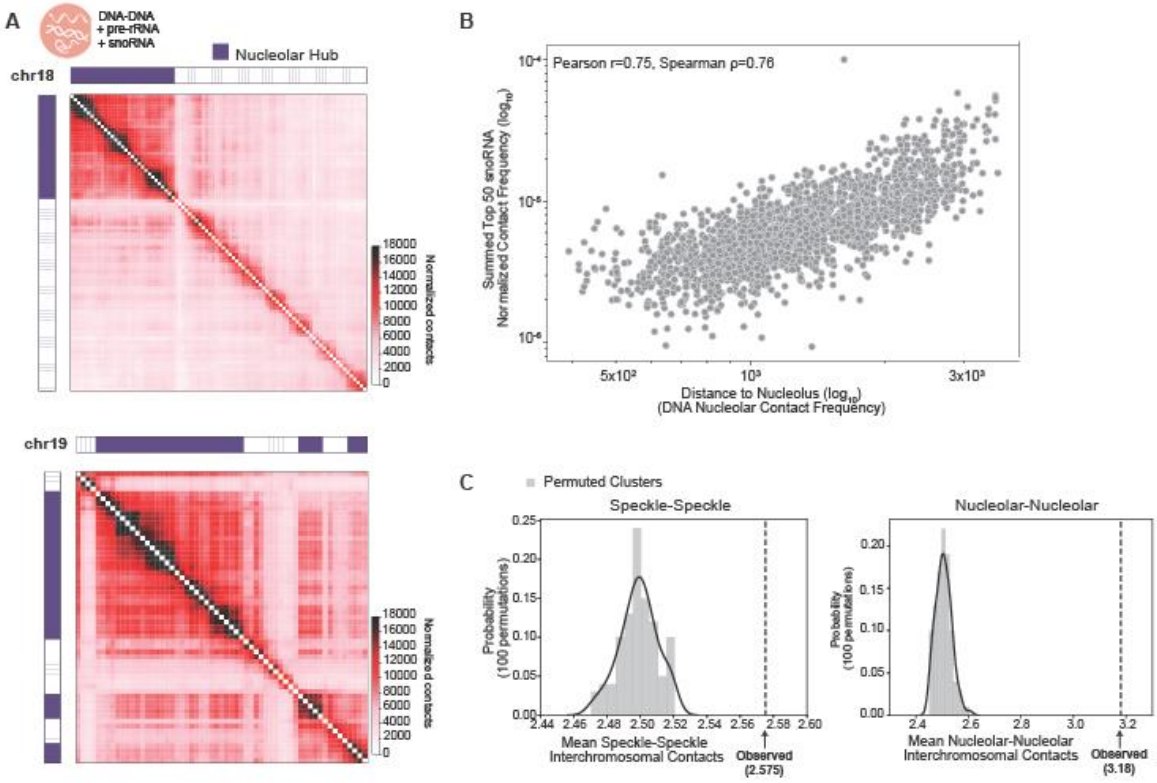
B



C



Supplemental Figure 4: RNA-RNA contacts for genes of varied expression levels, corresponding to Figure 3. (A) Summary statistics for five gene expression groups spanning the top 10,000 expressed pre-mRNA in RD-SPRITE. Each group contains 2,000 genes. **(B)** Genome-wide, intrachromosomal RNA-RNA contact matrices for pre-mRNAs within the five gene expression groups, collapsed by A/B compartment assignments of the individual genes. **(C)** RNA-RNA contact matrices of pre-mRNAs at the start of chromosome 2 for top three gene expression groups, collapsed by compartment domain assignments of the individual genes.



Supplemental Figure 5: Nucleolar Hub definition and interactions, corresponding to Figure 4. (A) DNA-DNA contact matrix of snoRNA or pre-rRNA containing clusters for chromosome 18 (top) and chromosome 19 (bottom). Nucleolar hub regions (purple) were defined by clustering of the interchromosomal contacts of the genome-wide DNA-DNA contact matrix generated from clusters containing snoRNA or 45S pre-rRNAs (see **Methods**). Hub regions are annotated with purple bars along the top and left axes. **(B)** Scatterplot of nucleolar hub distance versus snoRNA contacts for top 2000 expressed introns. SnoRNA contacts correspond to the total, normalized contact frequency between each pre-mRNA and the top 50 snoRNAs, shown in **Figure 4D**. Pearson and spearman correlation coefficients (top) show positive correlation between distance and contact frequency. **(C)** Expected distribution of mean interchromosomal RNA-RNA contacts between speckle hub genes (left) and nucleolar hub genes (right), used for significance testing of **Figure 4H**. RNA reads were randomly permuted among SPRITE clusters 100 times and interchromosomal RNA-RNA contacts were recalculated for each randomization (see **Methods**). Observed interchromosomal contact frequencies from **Figure 4H** are

shown as dotted lines and are located at considerably larger values than the expected distributions (permuted).

3.7. SUPPLEMENTAL TABLE AND VIDEO LEGENDS

Supplemental Table 1: Names and genomic locations of selected top 2000 expressed introns, related to Figures 3 and 4.

The 2,000 top expressed introns were identified in RD-SPRITE clusters of sizes 2-1000 (see **Methods**). The genes of these introns were used for profiling nascent pre-mRNA contacts with each other (**Figure 3**) and nascent transcription around the nucleolus (**Figure 4**).

Supplemental Table 2: Nucleolar Hub Regions, related to Figure 4.

Genomic DNA regions (1Mb resolution) that associate together around the nucleolus were annotated using RD-SPRITE clusters enriched for snoRNAs and pre-rRNAs (see **Methods**; mm10 genome).

Supplemental Table 3: Speckle Hub Regions, related to Figure 4.

Genomic DNA regions (1Mb resolution) that associate together around nuclear speckles were annotated using RD-SPRITE clusters enriched for snRNAs (see **Methods**; mm10 genome).

Supplemental Video 1: Intron FISH for Genes near the Nucleolus, corresponding to Figure 4I.

Video of intron RNA-FISH combined with immunofluorescence for Carnmt1 (yellow), a nucleolar proximal gene located on chromosome 19, and Btrc (green), a nucleolar distal gene also located on chromosome 19. Both alleles of Carnmt1 are transcribed while located adjacent to the nucleolus (Nucleolin; purple).

Supplemental Video 2: Intron FISH for Genes near the Nucleolus, corresponding to Figure 4J.

Video of intron RNA-FISH combined with immunofluorescence for Carnmt1 (yellow), a nucleolar proximal gene located on chromosome 19, and Btrc (green), a nucleolar distal gene also located on chromosome 19. Both alleles of Carnmt1 are transcribed while located on the surface of the nucleolus (Nucleolin; purple).

3.8. METHODS

Cell lines used in this study

We used the following cell line in this study: Female ES cells (pSM44 ES cell line) derived from a 129 × castaneous F1 mouse cross. These cells express Xist from the endogenous locus under control of a tetracycline-inducible promoter. The dox-inducible Xist gene is present on the 129 allele, enabling allele-specific analysis of Xist induction and X chromosome silencing.

Cell culture conditions

All mouse ES cells were grown at 37°C under 7% CO_2 on plates coated with 0.2% gelatin (Sigma, G1393-100ML) and 1.75 µg/mL laminin (Life Technologies Corporation, #23017015) in serum-free 2i/LIF media composed as follows: 1:1 mix of DMEM/F-12 (GIBCO) and Neurobasal (GIBCO) supplemented with 1x N2 (GIBCO), 0.5x B-27 (GIBCO 17504-044), 2 mg/mL bovine insulin (Sigma), 1.37 µg/mL progesterone (Sigma), 5 mg/mL BSA Fraction V (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), 5 ng/mL murine LIF (GlobalStem), 0.125 µM PD0325901 (SelleckChem) and 0.375 µM CHIR99021 (SelleckChem). 2i inhibitors were added fresh with each medium change. Medium was replaced every 24-48 hours depending on culture density, and cells were passaged every 72 hours using 0.025% Trypsin (Life Technologies) supplemented with 1mM EDTA and chicken serum (1/100 diluted; Sigma), rinsing dissociated cells from the plates with DMEM/F12 containing 0.038% BSA Fraction V.

RD-SPRITE Dataset and Computational Pipeline

The RNA-DNA SPRITE dataset was previously generated in female PSM44 mouse embryonic stem cells treated for 24hr with doxycycline for induction of Xist expression and can be found under the GEO accession number GSE151515[28]. RD-SPRITE data processing pipeline details were described in https://github.com/GuttmanLab/sprite2.0-pipeline. The final pipeline output is a cluster file containing reads grouped into SPRITE

clusters based on shared SPRITE barcode sequences. Each SPRITE cluster may contain repeat-masked DNA reads aligned to mm10, RNA reads annotated with Gencode vM25 gene annotations and repeatmasker annotations and/or repeat RNA reads aligned to a custom genome of repeat RNAs.

Gene identification was improved to avoid mis-annotations of genes. Intronic RNA reads were defined as RPM- containing reads that aligned to the genome and were uniquely annotated as the intron of a protein-coding gene. Similarly, exonic RNA reads were defined as RPM-containing reads that aligned to the genome and were uniquely annotated as the exon of a protein-coding gene. For example, mRNA reads overlapping with snoRNAs or other repetitive sequences were excluded. Unless stated otherwise, all analyses were based on SPRITE clusters of size 2-1000 reads. These cluster sizes were chosen to be consistent with the analysis in our previous papers, where we showed that many known structures such as TADs, compartments, RNA-DNA and RNA-RNA interactions, etc., occur within SPRITE clusters containing 2-1000 reads[28,30]. We also previously showed that the A/B compartments identified in these cluster sizes are highly correlated with those observed by Hi-C.

RD-SPRITE Data Analysis

Exon and Intron Enrichment Scores

A “contact enrichment score” was devised to compare the contact profiles of intronic and exonic RNA reads. Specifically, the frequency of an intronic or exonic read co-occurring in the same SPRITE cluster as another molecular species, such as chromatin (DNA reads, defined by DPM tags), small nuclear RNAs (U1, U2), or ribosomal RNAs (18S, 28S), was calculated. Contact scores were generated by sorting all clusters with a gene of interest based on whether they also contained the other molecule of interest. The contact and no-contact frequency scores were computed by taking the sum of 1/(cluster size) for all clusters with or without the other molecule, respectively. Summing 1/(cluster size) accounts for contact distance; larger clusters indicate further contact distances and are

proportionally down-weighted. A final contact score for a gene is the ratio of contact to no-contact frequency scores. This procedure was performed for intronic and exonic reads. Enrichment was computed by normalizing contact scores to the median contact score of all intronic and exonic contact scores. Outlier genes with very few annotated exonic or intronic reads had extreme scores that were not necessarily representative of intronic or exonic reads as a class. These were removed by setting a minimum contact and no contact threshold of two clusters in each category for a given gene. The median 90% of intronic and exonic contact enrichment scores were plotted as violin plots using the python plotting package seaborn.

Locus Enrichment Scores

To map the genome-wide localization profiles of intronic or exonic RNA reads relative to their genomic locus, the contact frequency between RNA transcripts (intronic or exonic reads) and each region of the genome (binned at 1Mb resolution) was calculated. 1Mb resolution was selected for this analysis such that most gene loci were located within only a single genomic bin (genomic locus bin). The raw contact frequency was defined as the number of SPRITE clusters in which a specific RNA transcript read (intronic or exonic) and a given genomic bin co-occur. The normalized contact frequency was calculated by weighting each SPRITE cluster by a scaling factor proportional to its size (2/n, where n is the total number of reads in the cluster). The normalized contact frequency profiles for intronic and exonic reads of each gene were summed over all genes, with the contact profiles centered on the genomic locus bin +/- 10Mb. The contact profiles for antisense (- strand genes) were reversed before summing to account for gene orientation.

Expression Correlations

Expression levels for each gene were calculated by counting the number of annotated intronic RNA reads and/or annotated exonic RNA reads in SPRITE clusters. The intronic expression levels were compared to GRO-Seq expression levels, from NCBI GEO (GSE48895 accession)[32]. The exonic expression levels were compared to a polyA-selected

RNA-seq library generated from PSM44 after 24hr dox induction (the identical cell line/conditions used for RD-SPRITE). The total expression levels were compared to Ribo-depleted RNA-seq expression levels from GEO Accession: GSM903663[31].

Selecting the Top 2000 introns

The top 2000 pre-mRNAs with highest coverage in the RD-SPRITE dataset were determined by counting the number of RNA reads in SPRITE clusters of size 2-1000 that were annotated as an intron of each protein-coding gene. To remove potential mis-annotations or non-representative genes, genes with transcript lengths greater than 1 Mb were filtered out. Additionally, to remove redundant or overlapping annotations, genes with intersecting annotations or annotations within 1000 bp of each other were filtered, keeping only the most abundant. This second filter allowed for any DNA locus to be uniquely assigned to only a single gene. From the final filtered list, the top 2000 pre-mRNA genes were selected.

DNA-DNA Contact Matrices

DNA-DNA contact matrices were generated from all SPRITE clusters of total size 2-1000 reads. Raw contact frequency was calculated at 1Mb resolution by counting the number of SPRITE clusters in which pairs of genomic bins co-occur. 1Mb resolution was selected for DNA-DNA heatmaps to enable visualization of genome-wide patterns (i.e. chromosomal territories, A/B compartments). Normalized contact frequency was calculated by dividing each genomic contact by a scaling factor proportional to SPRITE cluster size (specifically, n/2 where n is the total number of reads in the SPRITE cluster)[29]. Normalized contact frequency maps were corrected using ICE normalization, a matrix balancing algorithm commonly used for correcting Hi-C contact maps using CoolTools[52].

To analyze the 3D structure associated with a single active gene locus (Figure 1F), DNA-DNA contact maps were generated from a subset of SPRITE clusters that contained an intronic RNA transcript of the gene of interest. Cluster size normalized contact maps were generated at 100kb resolution by mapping the interactions between all pairs of DNA within

these clusters, as described above. 100kb resolution was selected to allow for visualization of contact enrichment at the gene-locus level. These contact frequency maps were corrected using a modified ICE normalization strategy. Specifically, because cluster subsets may be enriched or depleted for certain genomic regions or contacts, the assumptions for typical ICE normalization of a matrix do not apply. To correct these matrices for any genome coverage bias present in the entire SPRITE dataset, ICE bias factors from DNA-DNA contact matrices generated with all clusters were applied to the matrices generated from cluster subsets.

To analyze the 3D-structure associated with nascent transcription for all active regions or sets of genes (e.g. A1, A2, B1, B2, etc.), DNA-DNA contact maps were generated from a subset of SPRITE clusters that contained a specific set of RNA transcripts. In the case of mapping all active regions, SPRITE clusters that had DNA reads and at least one RNA read annotated as the intron of a protein coding gene were selected. Cluster size normalized contact maps were generated at 1Mb resolution from this subset of clusters, as described above. 1Mb resolution was selected to enable visualization of genome-wide structural patterns (i.e. chromosomal territories and A/B compartments). These contact maps were ICE normalized using the modified ICE normalization strategy for cluster subsets, as described above.

Genome-wide Eigenvector Calculations for A/B Compartment Identification

Genome-wide eigenvectors were calculated from SPRITE DNA-DNA contact maps to define reference A/B compartments. First, SPRITE clusters of sizes 2-1000 DNA reads were converted to a cooler format, a standard format for HiC interaction data, using the ‘cloud-pairs’ function of cooler[53]. The pairs of contacts within SPRITE clusters were individually written out and weighted by the n/2 scaling factor, where n is the number of DNA reads in the cluster. Next, cooler files were generated at various resolutions (10kb, 100kb, 1Mb) using the cooler function “coarsen” and matrix balancing weights were calculated using the cooler function “balance”. Finally, eigenvectors were calculated at these resolutions using the HiC analysis software cooltools. 1Mb resolution eigenvectors

were used to define A/B compartment domains for genome-wide or chromosome-wide analysis; 100kb resolution eigenvectors were used to match sub compartment resolution (see below); 10kb resolution eigenvectors were used to define compartments on a gene-resolution level and assign individual genes to either compartment.

RNA-DNA Contact Maps

Genome-wide localization profiles were generated for individual pre-mRNAs by calculating the contact frequency of intronic RNA reads for that gene and genomic DNA binned at various resolutions (10Mb, 1Mb, 100kb). A range of DNA binning resolutions was used to measure contacts occurring at different size scales — i.e., gene-locus specific contacts (100kb), intra-compartment contacts (1Mb), and long-range, chromosome-wide contacts (10Mb). Raw contact frequency was calculated by counting the number of SPRITE clusters in which an intronic RNA read and a DNA read, mapped to its corresponding genomic bin, co-occurred. Weighted contact frequency was calculated by scaling raw contacts with a scaling factor proportional to cluster size, as described for DNA-DNA contact matrices. To account for differences in gene expression when comparing RNA-DNA localization profiles across genes, the genome-wide RNA-DNA contacts for each gene were normalized to one.

Aggregate inter-compartment domain RNA-DNA contact frequencies were computed using the unweighted RNA-DNA contact profiles of the top 2000 genes at 1Mb genomic bin resolution. 1Mb resolution was selected to define A/B compartment domains and resolve compartment boundaries. First, for each gene, the A/B compartment domain containing the gene locus was masked. Then, the inter-compartment domain contacts for all A compartment genes and all B compartment genes were summed separately. To account for the difference in number of A and B compartment genes, the aggregated contact frequencies were normalized to their respective medians on a per-chromosome basis. Finally, the ratio of A-to-B frequency was used to generate an enrichment score. When the ratio is >1, the inter-compartment domain contact frequency with A genes is higher; when the ratio is <1, the inter-compartment domain contact frequency with B genes is higher.

This enrichment score was further aggregated across all A compartment regions and all B compartment regions on chromosome 2. To compare the magnitude of the enrichment score to the magnitude of eigenvector 1 (E1), the enrichment score along chromosome 2 was plotted after rank re-mapping to E1; all enrichment score values and all eigenvector values along chromosome 2 were ordered each from greatest to least. The top enrichment score was assigned the value of the top eigenvector, the second highest enrichment score was assigned the value of the second highest eigenvector and so forth.

RNA-RNA Contact Matrices

RNA-RNA contact matrices were generated by computing the contact frequency between RNA-RNA pairs. Contact frequency was defined as the number of SPRITE clusters containing both transcripts. RNA-RNA contacts were not weighted by cluster size.

For all of the top 2000 genes, their intronic or exonic RNA transcripts were used to generate RNA-RNA contact matrices between nascent pre-mRNAs or mature mRNAs, respectively. Matrices were ordered by the genomic position of these genes and normalized using ICE normalization (also used for matrix balancing HiC data) to account for differences in RNA expression.

DNA-DNA Contact Matrices By Gene

DNA-DNA contact matrices were generated on a gene-level to directly compare to the corresponding RNA-RNA matrices of the same genes. Specifically, instead of calculating the contact frequency between genomic bins of DNA, raw frequencies were calculated by annotating each DNA fragment with its respective gene locus (similar to RNA annotations but ignoring strand) and counting the number of SPRITE clusters containing pairs of interacting gene loci. These matrices were normalized using ICE normalization.

Eigenvector Calculations for Gene-Binned Contact Maps

Eigenvectors were calculated from RNA-RNA contact matrices or gene-resolution DNA-DNA contact matrices using these same HiC analysis software packages to define A/B

compartments (see above). Pre-computed, raw contact matrices were converted into a cooler format by assigning the contacts of each gene to the 10kb genomic bin located in the center of the gene annotation. Coolers were balanced using the cooler function "balance". Finally, eigenvectors were calculated using the HiC analysis software cooltools. Signs of eigenvectors for individual chromosomes were matched to the eigenvectors calculated from the entire genome.

Saddle Plots

Saddle plots were generated from the RNA-RNA and gene-resolution DNA-DNA contact matrices using cooltools. To enable direct comparisons, the ordering of genes was the same for all saddle plots. The genes were sorted and grouped into 40 bins based on eigenvector 1 (E1) of their genomic positions from the genome-wide eigenvector calculation using RD-SPRITE. The RNA-RNA and gene-resolution DNA-DNA contact matrices were then aggregated based on these groups. The total interaction sum and count were used to calculate an average contact frequency per group.

RNA-RNA Contacts for Gene Expression Levels

Genes were grouped into five expression levels of 2000 genes each based on pre-mRNA abundance in RD-SPRITE clusters. Specifically, genes were ordered by number of associated intron-containing RNA reads and grouped into 0-1999, 2000-3999, 4000-5999, 6000-7999, and 8000-9999, with the 0-1999 group containing the most abundant pre-mRNA genes. For each group, RNA-RNA contacts were mapped between pre-mRNAs as previously described. Individual gene-based contacts were collapsed into A and B compartment contacts for each chromosome. A and B compartments for each gene were assigned based on E1 calculated at 10kb resolution from RD-SPRITE data, as described above. The genome-wide, by-chromosome A/B compartment contact matrix was normalized using ICE normalization. Finally, the total intrachromosomal A-A, A-B and B-B, contacts were calculated and displayed as a 2-by-2 matrix. For the top three expression levels (0-1999, 2000-3999, 4000-5999), the gene-based RNA-RNA contacts were

additionally collapsed into contiguous domains of A and B along each chromosome (instead of one A and one B group per chromosome). The resulting contact matrix was normalized using ICE normalization. A similar domain level analysis was not performed for the lowest expression levels because of sparsity in pre-mRNA reads.

Sub-compartment Analysis

Annotations for the A/B sub-compartments in mouse embryonic stem cells were kindly provided by the Jian Ma laboratory at Carnegie Mellon University. These subcompartment annotations were only used for the analyses shown in Figure 2C and Figure 2D. Sub-compartment annotations were based on a 5 state Gaussian Hidden Markov Model and were at 100 kb resolution. The features of the sub-compartments were profiled using the RD-SPRITE dataset. Specifically, A/B compartment labels, assigned based on the principal eigenvector calculated at 100 kb resolution using RD-SPRITE, were compared to the A1/A2/B1/B2/B3 and the number of mismatched compartments (e.g. A compartment with B1/B2/B3 sub-compartment) was calculated. Next, the weighted RNA-DNA contacts of nascent pre-mRNAs or of small nuclear RNAs (U1, U2) across all regions of a single sub-compartment annotation were averaged. To determine whether genes in all sub-compartments were expressed, the top 2000 pre-mRNA genes were assigned to their respective sub-compartments. If a gene annotation intersected multiple sub-compartments, it was assigned to the sub-compartment with maximum representation (i.e. most covered base pairs); each gene could only be counted for one sub-compartment. Chromosome X genes were excluded from this analysis because of the lack of sub-compartment assignments.

Weighted DNA-DNA contact matrices for transcribing regions containing pre-mRNAs from each sub-compartment were generated as described above. Specifically, a subset of SPRITE clusters containing intronic RNA transcripts of genes located in a given sub-compartment were selected. Then, the DNA-DNA contact frequency from this cluster subset was calculated. Any of the top 2000 genes were assigned to a given sub-

compartment if a portion of the gene annotation intersected the given sub-compartment. Thus, genes could be included in more than one sub-compartment for this analysis.

Nucleolar and Speckle Hub Definition

Nucleolar and Speckle Hubs were previously defined using inter-chromosomal contacts of DNA SPRITE at 1Mb resolution in mES cells[30]. Briefly, it was found that two mutually-exclusive sets of DNA regions showed enriched inter-chromosomal contacts within a set but not with DNA regions of the other set of interacting loci.

To improve these annotations in this manuscript, the hubs were recalculated using the RD-SPRITE dataset and including RNA enrichment information. Instead of all SPRITE clusters being included to generate a DNA-DNA heatmap and measure inter-chromosomal contact enrichment, we only used SPRITE clusters containing known RNAs functionally associated with the respective nuclear body (nucleolus/speckles) being mapped. For the nucleolar hub, clusters were selected using small nucleolar RNAs (snoRNAs) or pre-rRNAs (45s rRNA); for the speckle hub, clusters were selected using small nuclear RNAs (e.g. U1, U2, or other snRNA 'biotype' genes). DNA-DNA contact frequency was calculated at 1Mb resolution (the same resolution as used for the original hub definition) from these cluster subsets and was not weighted by cluster size, in order to maximize the information from larger clusters which we have found are enriched for interactions around nuclear bodies[30]. The resulting raw heatmaps were balanced using the ICE bias factors of the DNA-DNA heatmap calculated using all clusters, as described above. Inter-chromosomal contacts were hierarchically clustered using the python package g.cluster.hierarchy[54]. Hierarchical clustering was converted into flat clusters using the fcluster function. Upon clustering, a single cluster of genomic bins nearly matching the previously annotated "inactive" hub (for the snoRNA/pre-rRNA workup) or "active" hub (for the snRNA workup) was apparent. These genomic bins within these clusters were redefined as the "Nucleolar" Hub and "Speckle" Hub.

Distance to Nucleolus using SPRITE Contacts

Genes located within the nucleolar hub annotations are defined as “nucleolus proximal” while genes located outside are defined as “nucleolus distal”.

Additionally, a continuous metric for distance to nucleolus was generated using DNA-DNA contact frequencies. For each 1Mb bin of the genome, the total inter-chromosomal contact frequency with nucleolar hub DNA region was calculated and genes were assigned the distance of the 1Mb genomic bin in which they are located.

snoRNA-RNA and pre-rRNA to pre-mRNA Contacts

For each of the top 2000 genes, the contact frequency between nascent pre-mRNAs and 2 sets of nucleolar hub RNAs, defined as individual snoRNAs or the three components of 45S pre-rRNA (ITS1, ITS2, 3’ETS), was calculated by counting the number of SPRITE clusters containing both an intronic read and a snoRNA/pre-rRNA read. A heatmap of snoRNA-RNA contacts was generated using only the top 50 snoRNAs with the highest contact frequency to the set of top 2000 genes. To account for differences in gene expression, the contacts for each gene were normalized to the total number of intronic reads for that gene (independent of contact with snoRNAs). To account for differences in snoRNA abundance, the total contacts of each snoRNA with the set of top 2000 genes was normalized to 1. The contact matrix between pre-rRNA and pre-mRNAs was similarly normalized.

Contact matrices were ordered in two ways: by genomic position and by distance to nucleolus. In both cases, snoRNAs/pre-rRNAs are ordered along the y-axis with pre-rRNAs on top, followed by snoRNAs in the order from most frequently to least frequently contacting the set of genes.

snoRNA-DNA and pre-rRNA to DNA Contacts at Gene Level

For comparison to the snoRNA-RNA and pre-rRNA-RNA contact matrices, the snoRNA-DNA and pre-rRNA-DNA contacts per gene were generated. Raw contact frequencies were calculated by counting the number of clusters in which a specific snoRNA or pre-

rRNA and a DNA read overlapping a gene annotation co-occur. To account for biases in DNA coverage, the raw frequencies per gene were divided by the total DNA coverage of that gene locus. To account for differences in snoRNA or pre-RNA abundance, the total contact of each individual RNA was normalized to 1.

RNA-DNA Contacts for snoRNA containing clusters

DNA localization profiles for nascent pre-mRNAs of nucleolar genes were calculated using a subset of clusters containing snoRNAs. Specifically, we selected clusters that contained the top 100 snoRNAs and mapped the RNA-DNA contacts within these clusters. We calculated the contact frequency between intronic RNA reads of nucleolar proximal genes and genomic DNA binned at 1Mb. 1Mb resolution was selected because we previously defined nucleolar hub regions at this resolution[30]. Raw contact frequencies and normalized contact profiles were generated as described in the RNA-DNA contact maps section above.

RNA-seq Experiments and Data Processing

PolyA-selected RNA-seq of mES cells

RNA-seq libraries of dox-induced PSM44 mouse embryonic stem cells (mES cells) were prepared using a double poly-A selection step prior to RNA library preparation (described in the Guttman Lab CLAP protocol; https://guttmanlab.caltech.edu/files/2021/08/CLAPprotocol_combined_word.pdf). Specifically, NEBNext Magnetic Oligo d(T)25 Beads (NEB, S1419S) were prepared by washing twice with RNA binding buffer (50 mM HEPES pH 7.5, 1000 mM LiCl, 2.5 mM EDTA, 0.1% Triton-X100). Total RNA was diluted in HEPES buffer, heated to 65°C for 5 minutes and then cooled to 4°C to denature RNA. Prepared Oligo dT(25) beads were mixed with denatured RNA and incubated at room temperature for 10 minutes to allow for RNA binding. These beads were then washed twice with RNA Wash Buffer (50 mM HEPES pH 7.5, 300 mM LiCl, 2.5 mM EDTA, 0.1% Triton-X100). Polyadenylated RNA

was eluted from beads by heating to 80°C for 2 minutes in HEPES Elution Buffer (5mM HEPES pH 7.5, 1.0 mM EDTA), followed by a hold at 25°C. This capture step (bind, wash and elute) was repeated, for two total capture steps, using the same Oligo dT beads. Specifically, to re-capture the polyA-selected RNA, RNA binding buffer was added to the mixture of beads and eluted RNA and the mixture was incubated at RT for 10 minutes. These beads were then washed twice with RNA Wash Buffer. The final selected polyA transcripts were eluted in HEPES Elution buffer by heating to 80°C for 2 minutes and holding at 25°C. The eluted beads were immediately placed on a magnet until the solution cleared and the cleared solution was transferred to a new tube. cDNA generation and library prep were performed as described in the Guttman Lab CLAP protocol after this.

Data Processing and Read Annotation

Libraries were sequenced on a HiSeq 2500 (90 cycle x 125 cycle). Adapters were trimmed from raw paired-end fastq files using Trimmomatic v0.38. Trimmed reads were then aligned to GRCm38.p6 with the Ensembl GRCm38 v95 gene model annotation using Hisat2 v2.1.01 with a high penalty for soft-clipping (--sp 1000,1000) and excluding mixed or discordant alignments (--no-mixed --no-discordant). Unmapped reads and reads with a low MapQ score (samtools view -bq 20) were filtered out. Mapped reads were annotated with the featureCounts tool from the subread package v1.6.4[55,56] using the Gencode release M25 annotations for GRCm38.p6 and a subset of the Repeat and Transposable element annotation from the Hammel lab, identical to the annotation strategy for genome-aligned RNA reads of RD-SPRITE. Reads that received a single annotation for a protein coding gene were counted and correlated with intronic read counts from RD-SPRITE.

Microscopy Experiments

Intron RNA fluorescence *in situ* (RNA-FISH)

RNA-FISH experiments were performed with ViewRNA ISH Cell Assay (ThermoFisher, QVC0001) protocol following manufacturer instructions with minor modifications[28,57]. First, pSM44 mES cells were fixed on coverslips with 4% formaldehyde in PBS for 15 minutes at room temperature followed by permeabilization 0.5% Triton X-100 in 1x PBS (RNAse-free) for 10 minutes at room temperature. Then, coverslips with cells were washed twice with 1x PBS (RNAse-free) and either dehydrated with 70% ethanol and stored for up to one week at -20C or used directly for the next step. Next, coverslips were washed one more time with 1x PBS and incubated with the desired combination of RNA FISH probes (custom probe design from Affymetrix) in Probe Set Diluent at 40°C for at least three hours. Coverslips were then rinsed once with 1x PBS, twice with Wash Buffer for 10 minutes, and rinsed once more with PBS before incubating in PreAmplifier Mix Solution at 40°C for 45 minutes. This step was repeated for the Amplifier Mix Solution and Label Probe Solution. After all three steps of amplification were performed followed by washes, coverslips were incubated with 1x DAPI in PBS at room temperature for 15 minutes and subsequently mounted onto glass slides using ProLong Gold with DAPI (Invitrogen, P36935).

RNA-FISH & Immunofluorescence

For IF combined with *in situ* RNA visualization, the ViewRNA Cell Plus (Thermo Fisher Scientific, 88-19000-99) kit was used following the RNA-FISH part of protocol from above with minor modifications. First, pSM44 mES cells were fixed on coverslips with 4% formaldehyde in PBS for 15 minutes at room temperature followed by permeabilization with 0.5% Triton X-100 in 1x PBS for 10 minutes at room temperature. Next, immunostaining was performed starting with two washes of coverslips with 1x PBS (RNAse-free) and blocking with blocking buffer (kit) with addition of RNAse inhibitor (kit) for 30 minutes. Then, coverslips were incubated with primary antibody for 3 hours at room temperature in a blocking buffer with RNAse inhibitor (anti-Nucleolin Abcam Cat# ab22758, RRID:AB 776878, 1:500). After incubation, cells were washed 3 times in 1x PBS (RNAse-free) and incubated for 1 hour at room temperature with secondary antibody labeled with Alexa fluorophores (Invitrogen, Alexa 555) diluted in 1x PBS (1:500). Next,

coverslips were washed three times in 1x PBS (RNAse-free) and RNA-FISH protocol was performed starting with probe incubation step (described above). After the final wash, coverslips were rinsed in ddH2O, mounted with ProLong Gold with DAPI (Invitrogen, P36935), and stored at 4°C until acquisition.

Image quantification and analysis

RNA-FISH only images were acquired with Zeiss LSM 800 with the 63x oil objective and collected every 0.3 µm for 16 Z-stacks, IF/RNA-FISH images were acquired with Zeiss LSM 980 with the 63x oil objective and collected every 0.3 µm for 16 Z-stacks.

Image analysis was performed using an Icy (v2.3) software followed by custom written python script for x, y, z Euclidean distance measurements. Briefly, a region of interest corresponding to each nucleus was determined using DAPI staining. Next, in each nucleus, intron spots were identified based on a local intensity threshold. Only nuclei with at least one spot for each target probed and a maximum of two spots per individual target (corresponding to the individual alleles) were kept for further analysis; nuclei that did not meet criteria were discarded. Then, the x, y, z position of each intron spot was determined and used to calculate Euclidean distance between all possible pairs of gene alleles. Using this matrix of interactions, nuclei were selected for further analysis only if they contained one or two full sets of triplet alleles (B-A-B) and each pair of alleles within the triplet was in a proximity of less than 20 units. This allows us to focus on triplets of genes that come from the same allele.

Imaris software v8 from Bitplane (Oxford Instruments Company) was used to visualize Nucleolin and intron-RNA-FISH localization. Distances were measured from the middle of the 3D spot constructed from allele intensity to the 3D surface constructed from the nucleolin signal.

Quantification and statistical analysis

Details of statistical analyses performed in this paper including analyses packages can be found in the figure legends, main text, and STAR Methods. Spearman correlation coefficients and Pearson correlation coefficients were calculated using the stats module of the scipy python package[54]. Mann-Whitney U test was performed using the stats module of the scipy python package[54]. Precision measures such as mean, median, quartiles, standard deviation, and bootstrapped confidence intervals are described in the corresponding figure legends.

Significance of inter-chromosomal RNA-RNA contacts

To compute the significance of inter-chromosomal RNA-RNA contacts between speckle hub or nucleolar hub genes, the RNA-RNA contacts between the top 2000 pre-mRNAs were randomly permuted to generate an expected distribution for contact frequency. Specifically, the pre-mRNA reads associated with these genes were randomized across the clusters containing them. The RNA-RNA contacts for the permuted clusters were calculated, the gene-based RNA-RNA contact map were normalized using ICE, and the inter-chromosomal contacts were collapsed into speckle hub genes, nucleolar hub genes, or neither. This procedure was repeated 100 times to generate an expected distribution of mean inter-chromosomal contacts. The observed value was compared to the expected distribution to generate a p-value.

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-Nucleolin	Abcam	Cat# ab22758; RRID:AB_776878
Chemicals, peptides, and recombinant proteins		
Doxycycline	Sigma	D9891
Deposited data		
SPRITE data	Quinodoz et al., 2021	GEO:GSE151515
Ribo-Depleted RNA RNA-seq data in mESC	Sigova et al., 2013	GEO:GSM903663
PolyA RNA-seq data in mESC	This Study	GEO: GSE211287
GRO-seq data in mES cells	Jonkers et al., 2014	GEO:GSE48895
Experimental models: Cell lines		
Mouse: pSM44 ES cell line	This Study	pSM44 (dox-inducible Xist)
Software and algorithms		
SPRITE pipeline 2.0 (v0.2)	This Study	https://github.com/GuttmanLab/sprite2.0-pipeline https://doi.org/10.5281/zenodo. 7030136
Bowtie2 (v2.3.5)	Langmead and Salzberg, 2012	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Hisat (v2.1.0)	Kim, Langmead, and Salzberg, 2015	http://www.ccb.jhu.edu/software/hisat/index.shtml
Samtools (v1.4)	Li et al., 2009	http://samtools.sourceforge.net/
Bedtools (v2.30.0)	Quinlan and Hall, 2010	https://bedtools.readthedocs.io/en/latest/

Trim Galore! (v0.6.2)	Felix Krueger (The Babraham Institute)	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Subread (v2.0.3)	Liao et al., 2013, 2014	http://subread.sourceforge.net/
Cooler (v0.8.5)	Abdennur and Mirny, 2019	https://github.com/open2c/cooler
Cooltools (v0.4.1)	10.5281/zenodo.5214125	https://github.com/open2c/cooltools
Scipy (v1.7.1)	Virtanen et al., 2020	https://scipy.org/

3.9. REFERENCES

1. Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat Rev Mol Cell Biol *16*, 245–257. 10.1038/nrm3965.

2. 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., et al. (2017). The 4D nucleome project. Nature. 10.1038/nature23884.

3. Misteli, T. (2020). The Self-Organizing Genome: Principles of Genome Architecture and Function. Cell *183*, 28–45. 10.1016/j.cell.2020.09.014.

4. Dekker, J. (2002). Capturing Chromosome Conformation. Science (1979) *295*, 1306–1311. 10.1126/science.1067799.

5. Dekker, J. (2016). Mapping the 3D genome: Aiming for consilience. Nat Rev Mol Cell Biol *17*, 741–742. 10.1038/nrm.2016.151.

6. Gibcus, J.H., and Dekker, J. (2013). The Hierarchy of the 3D Genome. Mol Cell *49*, 773–782. 10.1016/j.molcel.2013.02.011.

7. Lieberman-aiden, E., Berkum, N.L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Revelas Folding Principles of the Human Genome. Science *326*, 289–294. 10.1126/science.1181369.

8. Kempfer, R., and Pombo, A. (2019). Methods for mapping 3D chromosome architecture. Nat Rev Genet. 10.1038/s41576-019-0195-2.

9. van Steensel, B., and Furlong, E.E.M. (2019). The role of transcription in shaping the spatial organization of the genome. Nat Rev Mol Cell Biol *20*. 10.1038/s41580-019-0114-6.

10. Oudelaar, A.M., and Higgs, D.R. (2021). The relationship between genome structure and function. Nat Rev Genet *22*. 10.1038/s41576-020-00303-x.

11. Finn, E.H., Pegoraro, G., Brandão, H.B., Valton, A.-L., Oomen, M.E., Dekker, J., Mirny, L., and Misteli, T. (2017). Heterogeneity and Intrinsic Variation in Spatial Genome Organziation. BioRxiv. 10.1101/171801.

12. Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336. 10.1038/nature14222.

13. Rowley, M.J., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nat Rev Genet *19*. 10.1038/s41576-018-0060-8.

14. Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B.R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G., Solovei, I., et al. (2019). Heterochromatin drives compartmentalization of inverted and conventional nuclei. Nature. 10.1038/s41586-019-1275-3.

15. Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. Cell *164*. 10.1016/j.cell.2016.02.007.

16. Shah, S., Takei, Y., Zhou, W., Lubeck, E., Yun, J., Eng, C.H.L., Koulena, N., Cronin, C., Karp, C., Liaw, E.J., et al. (2018). Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. Cell. 10.1016/j.cell.2018.05.035.

17. Ferrai, C., de Castro, I.J., Lavitas, L., Chotalia, M., and Pombo, A. (2010). Gene positioning. Cold Spring Harb Perspect Biol *2*. 10.1101/cshperspect.a000588.

18. Mahy, N.L., Perry, P.E., Gilchrist, S., Baldock, R.A., and Bickmore, W.A. (2002). Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. Journal of Cell Biology *157*. 10.1083/jcb.200111071.

19. Mahy, N.L., Perry, P.E., and Bickmore, W.A. (2002). Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. Journal of Cell Biology *159*, 753–763. 10.1083/jcb.200207115.

20. Williamson, I., Berlivet, S., Eskeland, R., Boyle, S., Illingworth, R.S., Paquette, D., Dostie, J., and Bickmore, W.A. (2014). Spatial genome organization: Contrasting views from chromosome conformation capture and fluorescence in situ hybridization. Genes Dev *28*, 2778–2791. 10.1101/gad.251694.114.

21. Gu, H., Harris, H., Olshansky, M., Eliaz, Y., Krishna, A., Kalluchi, A., Jacobs, M., Cauer, G., Pham, M., Rao, S.S.P., et al. (2021). Fine-mapping of nuclear compartments using ultra-deep Hi-C shows that active promoter and enhancer elements localize in the active A compartment even when adjacent sequences do not.

22. van Steensel, B., and Belmont, A.S. (2017). Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. Cell *169*. 10.1016/j.cell.2017.04.022.

23. Reddy, K.L., Zullo, J.M., Bertolino, E., and Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature *452*, 243–247. 10.1038/nature06727.

24. Finlan, L.E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J.R., and Bickmore, W.A. (2008). Recruitment to the Nuclear Periphery Can Alter Expression of Genes in Human Cells. PLoS Genet *4*, e1000039. 10.1371/journal.pgen.1000039.

25. Kumaran, R.I., and Spector, D.L. (2008). A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. Journal of Cell Biology *180*, 51–65. 10.1083/jcb.200706060.

26. Wang, Y., Zhang, Y., Zhang, R., van Schaik, T., Zhang, L., Sasaki, T., Peric-Hupkes, D., Chen, Y., Gilbert, D.M., van Steensel, B., et al. (2021). SPIN reveals genome-wide landscape of nuclear compartmentalization. Genome Biol *22*. 10.1186/s13059-020-02253-3.

27. Padeken, J., and Heun, P. (2014). Nucleolus and nuclear periphery: Velcro for heterochromatin. Curr Opin Cell Biol *28*, 54–60. 10.1016/j.ceb.2014.03.001.

28. 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., et al. (2021). RNA promotes the formation of spatial compartments in the nucleus. Cell *184*, 5775-5790.e30. 10.1016/j.cell.2021.10.014.

29. Quinodoz, S.A., Bhat, P., Chovanec, P., Jachowicz, J.W., Ollikainen, N., Detmar, E., Soehalim, E., and Guttman, M. (2022). SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. Nat Protoc *17*. 10.1038/s41596-021-00633-y.

30. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell. 10.1016/j.cell.2018.05.024.

31. Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., et al. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc Natl Acad Sci U S A *110*. 10.1073/pnas.1221904110.

32. Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *2014*. 10.7554/eLife.02407.

33. 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., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. 10.1016/j.cell.2014.11.021.

34. Xiong, K., and Ma, J. (2019). Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. Nat Commun *10*. 10.1038/s41467-019-12954-4.

35. Pederson, T. (2011). The nucleolus. Cold Spring Harb Perspect Biol *3*, 1–15. 10.1101/cshperspect.a000638.

36. Mélèse, T., and Xue, Z. (1995). The nucleolus: an organelle formed by the act of building a ribosome. Curr Opin Cell Biol. 10.1016/0955-0674(95)80085-9.

37. Németh, A., and Längst, G. (2011). Genome organization in and around the nucleolus. Trends in Genetics *27*, 149–156. 10.1016/j.tig.2011.01.002.

38. Kresoja-Rakic, J., and Santoro, R. (2019). Nucleolus and rRNA Gene Chromatin in Early Embryo Development. Trends in Genetics. 10.1016/j.tig.2019.06.005.

39. Boyle, S., Rodesch, M.J., Halvensleben, H.A., Jeddeloh, J.A., and Bickmore, W.A. (2011). Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Research *19*. 10.1007/s10577-011-9245-0.

40. Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*. 10.1038/nature06947.

41. Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. Mol Cell *38*, 603–613. 10.1016/j.molcel.2010.03.016.

42. Kumaran, R.I., and Spector, D.L. (2008). A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. Journal of Cell Biology *180*. 10.1083/jcb.200706060.

43. Collombet, S., Rall, I., Dugast-Darzacq, C., Heckert, A., Halavatyi, A., le Saux, A., Dailey, G., Darzacq, X., and Heard, E. (2021). RNA polymerase II depletion from the inactive X chromosome territory is not mediated by physical compartmentalization. bioRxiv.

44. Jung, Y.H., Sauria, M.E.G., Lyu, X., Cheema, M.S., Ausio, J., Taylor, J., and Corces, V.G. (2017). Chromatin States in Mouse Sperm Correlate with Embryonic and Adult Regulatory Landscapes. Cell Rep *18*. 10.1016/j.celrep.2017.01.034.

45. Battulin, N., Fishman, V.S., Mazur, A.M., Pomaznoy, M., Khabarova, A.A., Afonnikov, D.A., Prokhortchouk, E.B., and Serov, O.L. (2015). Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. Genome Biol *16*. 10.1186/s13059-015-0642-0.

46. Jiang, Y., Huang, J., Lun, K., Li, B., Zheng, H., Li, Y., Zhou, R., Duan, W., Wang, C., Feng, Y., et al. (2020). Genome-wide analyses of chromatin interactions after the loss of Pol I, Pol II, and Pol III. Genome Biol *21*. 10.1186/s13059-020-02067-3.

47. Barutcu, A.R., Blencowe, B.J., and Rinn, J.L. (2019). Differential contribution of steady-state RNA and active transcription in chromatin organization . EMBO Rep. 10.15252/embr.201948068.

48. Lu, J.Y., Chang, L., Li, T., Wang, T., Yin, Y., Zhan, G., Han, X., Zhang, K., Tao, Y., Percharde, M., et al. (2021). Homotypic clustering of L1 and B1/Alu repeats compartmentalizes the 3D genome. Cell Res *31*. 10.1038/s41422-020-00466-6.

49. Sima, J., Chakraborty, A., Dileep, V., Michalski, M., Klein, K.N., Holcomb, N.P., Turner, J.L., Paulsen, M.T., Rivera-Mulia, J.C., Trevilla-Garcia, C., et al. (2019). Identifying cis Elements for Spatiotemporal Control of Mammalian DNA Replication. Cell *176*. 10.1016/j.cell.2018.11.036.

50. Lu, J., and Gilbert, D.M. (2007). Proliferation-dependent and cell cycle-regulated transcription of mouse pericentric heterochromatin. Journal of Cell Biology. 10.1083/jcb.200706176.

51. Marchal, C., Sima, J., and Gilbert, D.M. (2019). Control of DNA replication timing in the 3D genome. Nat Rev Mol Cell Biol *20*. 10.1038/s41580-019-0162-y.

52. Venev, S., Abdennur, N., Goloborodko, A., Flyamer, I., Fudenberg, G., Nuebler, J., Galitsyna, A., Akgol, B., Abraham, S., Kerpedjiev, P., et al. (2021). open2c/cooltools: v0.4.1.

53. Abdennur, N., and Mirny, L.A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics *36*, 311–316. 10.1093/bioinformatics/btz540.

54. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods *17*, 261–272. 10.1038/s41592-019-0686-2.

55. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. 10.1093/bioinformatics/btt656.

56. Liao, Y., Smyth, G.K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res *41*, e108. 10.1093/nar/gkt214.

57. Jachowicz, J.W., Strehle, M., Banerjee, A.K., Blanco, M.R., Thai, J., and Guttman, M. (2022). Xist spatially amplifies SHARP/SPEN recruitment to balance chromosome-wide silencing and specificity to the X chromosome. Nat Struct Mol Biol *29*, 239–249. 10.1038/s41594-022-00739-1.

Chapter 4

RNA PROMOTES THE FORMATION OF SPATIAL COMPARTMENTS IN THE NUCLEUS

Sofia A. Quinodoz, Joanna W. Jachowicz, Prashant Bhat, Noah Ollikainen, Abhik K. Banerjee, Isabel N. Goronzy, Mario R. Blanco, Peter Chovanec, Amy Chow, Yolanda Markaki, Jasmine Thai, Kathrin Plath, and Mitchell Guttman

A modified version of this chapter was published as “RNA promotes the formation of spatial compartments in the nucleus” in *Cell*. 184(23):5775-5790. 2021. doi: 10.1016/j.cell.2021.10.014.

4.1. SUMMARY

The nucleus is a highly organized arrangement of RNA, DNA, and protein molecules that are spatially organized within three-dimensional (3D) structures. Although RNA has long been proposed to play a global role in organizing nuclear structure, exploring this has remained a challenge because no existing methods can simultaneously measure RNA and DNA contacts within 3D structures. To address this, we developed a genome-wide approach to comprehensively map the spatial organization of all RNAs relative to DNA called RNA & DNA SPRITE (RD-SPRITE). Using this approach, we detect higher-order RNA-chromatin structures associated with three major classes of nuclear function: RNA processing (including ribosome biogenesis, mRNA splicing, snRNA biogenesis, and histone mRNA processing), heterochromatin assembly, and gene regulation. We identify hundreds of ncRNAs that form high-concentration territories in spatial proximity to their transcriptional loci. Focusing on several examples, we show that RNA is required to recruit ncRNA and protein regulators into dozens of precise 3D structures in the nucleus. We show that specific ncRNAs can shape long-range DNA contacts, heterochromatin assembly, and gene expression within these spatial territories. Together, our results demonstrate a unique mechanism by which RNAs can act to shape nuclear structure by forming high concentration spatial territories immediately upon transcription, binding to diffusible regulators, and guiding them into spatial compartments to regulate a range of essential nuclear functions.

4.2. INTRODUCTION

The nucleus is spatially organized in three-dimensional (3D) structures that are important for various functions including transcription and RNA processing[1–6]. To date, genome-wide studies of nuclear organization have focused primarily on the role of DNA[7–9], yet nuclear structures are known to contain multiple DNA, RNA, and protein molecules that are involved in shared functional and regulatory processes[1–6]. These include classical compartments like the nucleolus[10] (which contains transcribed ribosomal RNAs and their processing molecules) and nuclear speckles[11] (which contain nascent pre-mRNAs and mRNA splicing components), as well as more recently described transcriptional condensates[12,13] (which contain Mediator and RNA Polymerase II). Because the complete molecular architecture of the nucleus has not been globally explored, the full extent to which such compartments exist and contribute to nuclear function remains unknown. Even for the specific nuclear compartments that have been molecularly characterized, the mechanism by which intrinsically diffusible RNA and protein molecules become spatially organized within these structures remains largely unknown.

Nuclear RNA has long been proposed to play a central role in shaping nuclear structure[14–19]. Initial experiments performed more than 30 years ago found that global disruption of RNA (using RNase) leads to large scale morphological deficits in the nucleus[14]. Over the past decade it has become clear that mammalian genomes encode thousands of nuclear-enriched ncRNAs[20–22], several of which play critical roles in the regulation of essential nuclear functions[23,24]. These include ncRNAs involved in splicing of pre-mRNAs (snRNAs)[25,26], cleavage and modification of pre-ribosomal RNAs (snoRNAs, Rnase MRP)[27–29], 3’-end cleavage and processing of the non-polyadenylated histone pre-mRNAs (U7 snRNA)[30–33], and transcriptional regulation (e.g. Xist[34–36] and 7SK[37–39]). Interestingly, many of these functionally important ncRNAs localize within specific spatial compartments in the nucleus[6,40,41]. For example, snoRNAs and the 45S pre-ribosomal RNA localize within the nucleolus[10,42–44], the Xist lncRNA localizes on the inactive X chromosome (Barr body)[45–48], and snRNAs and Malat1 lncRNA localize within nuclear speckles[11,49].

In each of these examples, multiple RNA, DNA, and protein components simultaneously interact within precise three-dimensional structures to coordinate specific nuclear functions. While the roles of these specific ncRNAs have been well studied, comprehensively mapping the localization patterns of most nuclear ncRNAs relative to other RNAs and DNAs in 3D space remains a challenge because no existing method can simultaneously measure higher-order RNA-RNA, RNA-DNA, and DNA-DNA contacts within 3D structures. As a result, it is unclear (i) which specific RNAs might be involved in nuclear organization[15,17,19], (ii) which specific nuclear compartments are dependent on RNA, and (iii) what mechanisms RNA might utilize to organize nuclear structures.

Microscopy is currently the only way to relate RNA and DNA molecules in 3D space. However, this approach is limited to examining a small number of simultaneous interactions and therefore requires *a priori* knowledge of which RNAs and nuclear structures to explore. An alternative approach is genomic mapping of RNA-DNA contacts using proximity-ligation methods[50–54]. While these approaches can provide genome-wide pairwise maps of RNA-DNA interactions, they do not provide information about the 3D organization of these molecules in the nucleus. Moreover, we recently showed that proximity-ligation methods can fail to identify pairwise contacts between molecules that are organized within certain nuclear compartments because these methods only identify interactions where components are close enough in space to be directly ligated[55]. Consistent with this observation, existing RNA-DNA proximity-ligation methods fail to identify known RNA-DNA contacts that are contained within various well-established nuclear bodies, such as nucleoli, histone locus bodies (HLBs), and Cajal bodies[52–54].

We recently developed SPRITE, a proximity-ligation independent method that utilizes split-and-pool barcoding to generate accurate, comprehensive, and multi-way 3D spatial maps of the nucleus across a wide range of distances[55]. Importantly, we showed that this approach can accurately map the spatial organization of DNA arranged around two nuclear bodies: nucleoli and nuclear speckles[55]. However, our original version of the technique could not detect the vast majority of RNAs — including low abundance ncRNAs known to organize within several well-defined nuclear structures — thereby precluding a

comprehensive map of RNA localization within the nucleus. Here, we introduce a dramatically improved method, RNA & DNA SPRITE (RD-SPRITE), which enables simultaneous and high-resolution mapping of thousands of RNAs — including low abundance RNAs such as individual nascent pre-mRNAs and ncRNAs — relative to all other RNA and DNA molecules in 3D space. Using this approach, we identify several higher-order RNA-chromatin hubs as well as hundreds of ncRNAs that form high concentration territories throughout the nucleus. Focusing on specific examples, we show that many of these RNAs act to recruit diffusible ncRNA and protein regulators and can shape long-range DNA contacts, heterochromatin assembly, and gene expression within these spatial territories. Together, our results highlight a unique role for RNA in the formation of nuclear compartments that are involved in a wide range of essential nuclear functions including RNA processing, heterochromatin assembly, and gene regulation.

4.3. RESULTS

RD-SPRITE generates accurate maps of higher-order RNA and DNA contacts throughout the cell

To explore the role of RNA in shaping nuclear structure, we developed RNA & DNA SPRITE (RD-SPRITE), which enables simultaneous mapping of multi-way DNA-DNA, RNA-DNA, and RNA-RNA contacts. Specifically, we improved the efficiency of the RNA-tagging steps of our SPRITE method[55] to enable detection of all classes of RNA, from highly abundant ribosomal RNAs and snRNAs to less abundant lncRNAs and individual nascent pre-mRNAs (**Supplemental Note 1**). Briefly, our approach works as follows: (i) RNA, DNA, and protein contacts are crosslinked to preserve their spatial relationships *in situ*, (ii) cells are lysed and the contents are fragmented into smaller crosslinked complexes, (iii) DNA and RNA within each complex are tagged with a sequence-specific adaptor, (iv) barcoded using an iterative split-and-pool strategy to uniquely assign a shared barcode to all DNA and RNA components contained within a

crosslinked complex, (v) DNA and RNA are sequenced, and (vi) all reads sharing identical barcodes are merged into a group that we refer to as a SPRITE cluster (**Figure 1A, Supplemental Figure 1A**, see **Methods**). Because RD-SPRITE does not rely on proximity ligation, it can detect multiple RNA and DNA molecules that simultaneously associate within the nucleus.

We performed RD-SPRITE in an F1 hybrid female mouse ES cell line that was engineered to induce Xist from a single allele (see **Methods**). We sequenced these libraries on a NovaSeq S4 run to generate ~8 billion reads corresponding to ~720 million SPRITE clusters (**Supplemental Figure 1C, Supplemental Table 2**). We confirmed that we accurately identify RNA- and DNA-specific reads (**Supplemental Figure 1A-B**) and that the data measure *bona fide* RNA interactions — including well-described RNA-DNA and RNA-RNA contacts not only in the nucleus, but throughout the cell.

First, we explored RNA-DNA contacts captured in our data and compared their interactions to those of several ncRNAs that were previously mapped to chromatin and reflect a range of known *cis* and *trans* localization patterns. Specifically, we observed strong enrichment of (i) Xist over the inactive X (Xi), but not the active X chromosome (Xa)[56,57] (**Figure 1B, Supplemental Figure 1D**); (ii) Malat1 and U1 over actively transcribed RNA Polymerase II genes[58,59] (**Figure 1B**); and (iii) telomerase RNA component (Terc) over telomere-proximal regions of all chromosomes (**Supplemental Figure 1E**)[60,61].

Second, we explored known RNA-RNA contacts that occur in different locations in the cell. For example, we observed a large number of contacts between translation-associated RNAs in the cytoplasm, including all RNA components of the ribosome (5S, 5.8S, 18S, 28S rRNA) and ~8000 individual mRNAs (exons), but not with pre-mRNAs (introns). Conversely, we observed many contacts between the small nuclear RNA (snRNA) components of the spliceosome (e.g. U1, U2, U4, U5, U6) in the nucleus and individual pre-mRNAs (**Figure 1C**).

Together, these results demonstrate that RD-SPRITE accurately measures known RNA-DNA and RNA-RNA localization patterns in the nucleus and cytoplasm. While we focus primarily on RNA localization within the nucleus, we note that RD-SPRITE can also be utilized to study RNA compartments beyond the nucleus[62–64].

Multiple non-coding RNAs co-localize within spatial compartments in the nucleus

Because RD-SPRITE generates comprehensive maps of RNA and DNA localization in the nucleus, we explored which specific RNAs localize within spatial compartments. To do this, we first mapped pairwise RNA-RNA and RNA-DNA contacts and identified several groups of RNAs in which member RNAs display high pairwise contact frequencies with each other, but low contact frequencies with RNAs in other groups (**Figure 1D**). Interestingly, the multiple pairwise interacting RNAs within the same group also localize to similar genomic DNA regions. For example, we observe pairwise contacts between snRNAs (e.g. U1, U2) and other RNAs known to localize in nuclear speckles (e.g. Malat1, 7SK), and these RNAs display similar localization over actively transcribed DNA regions (**Supplemental Figure 1G-H**). Using a combination of RNA FISH to visualize RNAs and immunofluorescence to visualize different cellular compartments, we confirmed that RNAs within groups spatially co-localize (**Supplemental Figure 1I**), while RNAs in distinct groups localize to different regions of the cell (**Supplemental Figure 1J**).

We next explored whether groups of pairwise interacting RNAs form higher-order structures within the nucleus. To do this, we computed the frequency of observing simultaneous contacts between 3 or more distinct RNAs and compared this to the expected frequency if these RNAs were randomly distributed (see **Methods**). We observed many significant multi-way contacts within each group (see **Supplemental Table 1**). For example, we observe multi-way contacts between all RNA components of the spliceosome (5-way contacts between U1, U2, U4, U5, and U6, $p<0.01$, z-score=9.9). Overall, RNAs contained within a group exhibit a significantly higher number of multi-way contacts than

RNAs from distinct groups (~50-fold for 3-way contacts, **Supplemental Figure 1F**). We refer to these higher-order, multi-way structures as "hubs" and explore them below.

Non-coding RNAs form processing hubs around genomic DNA encoding their nascent targets

We first explored the RNA-DNA hubs associated with RNA processing. Specifically, we examined the RNA components in these hubs (RNA-RNA interactions), the location of these RNAs relative to genomic DNA (RNA-DNA interactions), and whether the multiple DNA loci come together in 3D space (DNA-DNA interactions). We observed that:

(i) ***ncRNAs involved in ribosomal RNA processing organize within a 3D compartment containing transcribed ribosomal RNA genes.*** We identified a hub that includes the 45S pre-ribosomal RNA (pre-rRNA), RNase MRP, and dozens of snoRNAs that are involved in rRNA biogenesis (**Figure 1D, Supplemental Figure 2A**). rRNA is transcribed as a single 45S precursor RNA and is cleaved by RNAse MRP and modified by various snoRNAs to generate the mature 18S, 5.8S, and 28S rRNAs[65–67]. We found that these ncRNAs form multi-way contacts with each other ($p<0.01$, z-score=31, **Supplemental Table 1**) and localize at genomic locations that are proximal to ribosomal DNA repeats that encode the 45S pre-rRNA and other genomic regions that organize around the nucleolus[55] (**Figure 2A, Supplemental Figure 2B,** see **Methods**). We explored the DNA-DNA interactions that occur within SPRITE clusters containing multiple nucleolar hub RNAs (45S pre-rRNA and snoRNAs, ≥4-way contacts) and observed that these RNAs and genomic DNA regions are organized together in 3D space (**Figure 2B**, **Supplemental Figure 2C,** see **Methods**). Our results demonstrate that the nascent 45S pre-rRNA, along with the diffusible snoRNAs and RNase MRP, is spatially enriched near the DNA loci from which it is transcribed.

(ii) ncRNAs involved in mRNA splicing are spatially concentrated around a high-density of transcribed Pol II genes. We identified a hub that contains nascent pre-mRNAs along

with all of the major (e.g. U1, U2, U4, U5, U6) and minor (U11, U12) spliceosomal ncRNAs and other ncRNAs associated with transcriptional regulation and mRNA splicing (e.g. 7SK and Malat1) (**Figure 1D, Supplemental Table 1**). Nascent pre-mRNAs are known to be directly bound and cleaved by spliceosomal RNAs to generate mature mRNA transcripts[25,68], yet it has been unclear how spliceosomal RNAs are organized in the nucleus relative to target pre-mRNAs and genomic DNA[69–74]. We first explored the possibility that the localization of splicing RNAs to genomic DNA regions occurs primarily through their association with nascent pre-mRNAs. In this case, we would expect that the DNA occupancy of splicing RNAs would be proportional to mRNA transcription levels, regardless of the 3D spatial position of an individual gene in the nucleus (**Figure 2C**). However, we find that these splicing RNAs do not show a uniform occupancy over all genes when normalized for transcription levels. Instead, they are more highly enriched over DNA regions containing a high-density of actively transcribed Pol II genes (Pearson r = 0.84-0.90, **Figure 2A, Supplemental Figure 2D**). When we explored the higher-order DNA contacts of these RNAs (≥2 distinct RNAs, ≥4-way RNA-DNA contacts), we found that these genomic DNA regions form preferential inter-chromosomal contacts and are comparable to regions we previously showed are organized around nuclear speckles[55] (**Figure 2D, Supplemental Figure 2E**). Interestingly, we observed that snRNA localization was significantly higher over DNA regions that are close to the nuclear speckle relative to those located farther away (**Figure 2C**), even when focusing on genes with comparable levels of transcription in both sets (**Figure 2E**). These results demonstrate that spliceosomal RNAs are spatially enriched near clusters of actively transcribed Pol II genes and their associated nascent pre-mRNAs.

(iii) ncRNAs involved in snRNA biogenesis are spatially organized around snRNA gene clusters. We identified a hub containing several annotated small Cajal body-associated RNAs (scaRNAs), two previously unannotated scaRNAs, and several small nuclear RNAs (snRNAs) (**Figure 1D, Supplemental Table 1, Supplemental Figure 3F,** see **Methods**). snRNAs are Pol II transcripts produced from multiple locations throughout the genome that undergo 2'-O-methylation and pseudouridylation before acting as functional

components of the spliceosome at thousands of nascent pre-mRNA targets[75–77]. scaRNAs directly hybridize to snRNAs to guide these modifications[78–80]. We found that scaRNAs are highly enriched at discrete genomic regions containing multiple snRNA genes in close linear space (**Figure 3A**). Because nascent snRNAs are hard to distinguish from mature snRNAs, we are unable to directly observe the spatial localization of nascent snRNAs on genomic DNA. However, because scaRNAs are known to bind to nascent snRNAs[81,82], we focused on SPRITE clusters containing snRNAs and scaRNAs and observed that these clusters are highly enriched at genomic DNA regions containing snRNA genes **(Figure 3A)**, indicating that nascent snRNAs are enriched near their transcriptional loci. Interestingly, despite being separated by large genomic distances, these DNA regions form long-range contacts (**Figure 3B, Supplemental Figure 3G**). In fact, we observe that these scaRNAs, snRNAs, and the distal DNA loci from which the snRNAs are transcribed simultaneously interact within higher-order SPRITE clusters (**Figure 3A, Supplemental Figure 3I**). Together, these results demonstrate that these components simultaneously interact within a spatial compartment in the nucleus. We note that this snRNA biogenesis hub may be similar to Cajal bodies, which have been noted to contain snRNA genes and scaRNAs[81–85] (see **Supplementary Note 2**).

(iv) The histone processing U7 snRNA is spatially enriched around histone gene loci. We identified a hub containing the U7 snRNA and various histone mRNAs (**Figure 1D**). Unlike most pre-mRNAs, histone pre-mRNAs are not polyadenylated; instead their 3'ends are bound and cleaved by the U7 snRNP complex to produce mature histone mRNAs[86,87]. This process is thought to occur within nuclear structures called Histone Locus Bodies (HLBs)[33,78], demarcated by NPAT protein (**Supplemental Figure 3B**). We observed that the U7 snRNA localizes at genomic DNA regions containing histone mRNA genes, specifically at two histone gene clusters on chromosome 13 (**Figure 3A**). To determine whether the U7 snRNA, histone gene loci, and nascent histone pre-mRNAs form a 3D spatial compartment, we generated DNA-DNA interaction maps from U7 snRNA-containing clusters (≥3-way RNA-DNA contacts) and observed long-range DNA contacts between the two histone gene clusters on chromosome 13 (**Figure 3C, Supplemental**

Figure 3H,J). We observed that scaRNAs also localize to these histone gene clusters (**Figure 3A)**, consistent with previous observations that HLBs and Cajal bodies are often found adjacent to each other in the nucleus[83,88] (see **Supplemental Note 2, Supplemental Figure 3C-E**).

Taken together, these results indicate that higher-order spatial organization of diffusible regulators around shared DNA sites and their corresponding nascent RNA targets is a common feature of many distinct forms of RNA processing, including ribosomal RNA, mRNA, snRNA, and histone mRNA biogenesis.

Spatial organization of processing compartments is dependent on transcription of nascent RNA

In each of these examples, we observed spatial compartments that consist of: (i) nascent RNAs localized near their DNA loci, (ii) these DNA loci forming long-range 3D contacts, and (iii) diffusible ncRNAs associating with these nascent RNAs and DNA loci within the compartment. Because many of these diffusible ncRNAs are known to directly bind to the nascent RNA (e.g. snoRNAs bind 45S pre-rRNA[27,89,90], U7 binds histone pre-mRNA[91–93], and scaRNAs bind pre-snRNAs[79,81]), we hypothesized that nascent transcription of RNA might act to form a high-concentration territory at these genomic DNA sites and recruit these diffusible ncRNAs into these spatial compartments.

To test whether transcription of nascent RNA is critical for the spatial organization of these compartments, we treated cells with actinomycin D (ActD), a drug that inhibits RNA Pol I and Pol II transcription[94], for 4 hours and performed RD-SPRITE (**Figure 4A, Supplemental Figure 4A**). We confirmed that ActD treatment led to robust inhibition of various nascent RNAs (>10-fold reduction, 45S pre-rRNA, histone mRNAs), but did not impact the steady-state RNA levels of their associated diffusible ncRNAs (e.g. snoRNAs, U7, scaRNAs) (**Figure 4B, Supplemental Figure 4B-C**).

We then explored the spatial organization of DNA and RNA after ActD treatment. Strikingly, while we did not observe structural changes of most DNA structural features (e.g., chromosome territories, A/B compartments, **Supplemental Figure 4I**), we observed large-scale disruption of DNA and RNA in the nuclear structures associated with the ribosome, snRNA, and histone biogenesis. Focusing on the nucleolar hub, we observed a strong depletion of RNA-RNA contacts between the various snoRNAs (**Figure 4C**) and global disruption of snoRNA localization at nucleolar DNA sites (**Figure 4D-E, Supplemental Figure 4D**). Instead, these snoRNAs, RMRP (another nucleolar-enriched RNA), and various proteins associated within the nucleolus (e.g. NPM1 and Fibrillarin), appeared to diffuse broadly throughout the nucleus (**Figure 4D, Supplemental Figure 4E,H**). Moreover, we observed a dramatic reduction in inter-chromosomal contacts between genomic DNA regions contained within the nucleolar hub (**Figure 4F, Supplemental Figure 4G**). These results indicate that transcription of 45S pre-rRNA (which is known to interact with snoRNAs and RNase MRP[23,65]) acts to concentrate these diffusible *trans*-acting regulatory ncRNAs and spatially organize DNA loci into the nucleolar compartment (**Figure 4G**).

We observed a similar loss of focal localization of scaRNAs at snRNA genes (**Figure 4E, Supplemental Figure 4D**), a change from focal to diffusive localization of scaRNAs in the nucleus (**Figure 4D**), and a striking reduction in the long-range DNA-DNA contacts between snRNA genes upon ActD treatment (**Figure 4F, Supplemental Figure 4G**). These results indicate that active transcription of nascent snRNAs (which are known to bind to scaRNAs[79,80]) acts to enrich diffusible scaRNAs and snRNA genomic loci into a defined spatial compartment (**Figure 4G**). We observed a loss of focal localization of U7 at the histone genes (**Figure 4E, Supplemental Figure 4D**) and specific loss of long-range DNA-DNA interactions occurring between the histone loci (**Figure 4F**). Furthermore, we observed an overall increase in the number of nuclear foci containing HLB-associated proteins (NPAT) within each cell (**Figure 4D, Supplemental Figure 4F**). These results indicate that nascent transcription of histone pre-mRNAs (which directly bind to U7) act to concentrate the *trans*-associating U7 ncRNA and other HLB proteins, and spatially

organize histone genomic DNA into the HLB compartment (**Figure 4G**). Consistent with this notion, previous studies have shown that histone pre-mRNAs are sufficient to seed the formation of the HLB and that the U7 binding site on the histone pre-mRNA is required for HLB formation[41,78,95].

Although we did not observe major changes in DNA-DNA or RNA-DNA contacts within the splicing hub, this may be because ActD only leads to a modest reduction (<2-fold) in nascent pre-mRNA (introns) levels (**Supplemental Figure 4A**). Consistent with this possibility, we previously observed significant changes in snRNA localization at active DNA sites following treatment with flavopiridol (FVP), a transcriptional inhibitor that runs off elongating Pol II and leads to robust reduction of nascent pre-mRNA levels[96].

4.4. DISCUSSION

We aimed to understand how intrinsically diffusible molecules become spatially enriched in 3D space within the nucleus. Our results demonstrate that ncRNAs can act as seeds to drive spatial localization of otherwise diffusive ncRNA and protein molecules in the nucleus. For example, we showed that experimental perturbations of several ncRNAs disrupt localization of diffusible proteins (HP1, SHARP) and ncRNAs (e.g. U7, snoRNAs, scaRNAs, etc.) in dozens of compartmentalized structures. In all of these cases, we observed a common theme where (i) specific RNAs localize at high concentrations in spatial proximity to their transcriptional loci and (ii) diffusible ncRNA and protein molecules that bind to these RNAs are enriched within these compartmentalized structures. More generally, we identified hundreds of additional ncRNAs that are spatially enriched near their transcription sites and, as such, may represent a widespread class of molecules that could act as localized seeds to guide spatial localization of regulatory factors throughout the nucleus. Together, these observations suggest a common mechanism by which RNA can mediate nuclear compartmentalization: nuclear RNAs can form high concentration spatial territories close to their transcriptional loci (“seed”), bind to diffusible regulatory ncRNAs and proteins through high affinity interactions (“bind”) and by doing so act to dynamically change the spatial distribution of these diffusible molecules in the

nucleus such that they are enriched within compartments composed of multiple DNA loci, regulatory and target RNAs, and proteins in 3D space ("recruit", **Figure 5**). By recruiting diffusible regulatory factors to multiple distinct DNA sites, these ncRNAs may also act to drive coalescence of distinct DNA regions into a shared territory in the nucleus. This may explain why various RNAs are critical for organizing long-range DNA interactions around specific nuclear bodies.

This mechanism may explain why many distinct types of RNA processing occur through compartmentalization of regulatory ncRNAs and proteins near their nascent RNA targets. Specifically, we show that each of these RNA processing hubs consists of a high concentration of nascent RNA near its transcriptional locus and enrichment of diffusible *trans*-associating ncRNAs — known to bind to the encoded nascent RNA — within the spatial compartment. In this way, these nuclear compartments contain high concentrations of regulatory RNAs and proteins in proximity to their nascent RNA targets, which are further organized within higher-order DNA structures that come together in 3D space to form distinct processing hubs. Because the efficiency of a biochemical reaction is increased when the substrate or enzyme concentration is increased, creating a high local concentration of regulators (e.g. spliceosomes) and targets (e.g. nascent pre-mRNAs) in 3D space may increase the kinetic efficiency of such reactions, and in turn increase the efficiency of co-transcriptional processing and regulation. This compartmentalization mechanism can also increase the rate at which regulators identify and engage targets, which may be particularly important in cases where the regulators (e.g. scaRNAs, U7) are expressed at low levels relative to their more abundant substrates (e.g. snRNAs, histone mRNAs). This spatial organization may be an important regulatory mechanism for ensuring the efficiency of co-transcriptional RNA processing and may explain how RNA processing and transcription are kinetically coupled.

Our results demonstrate that hundreds of nuclear ncRNAs are preferentially localized within precise territories in the nucleus, suggesting that this may be an important and common function exploited by additional nuclear RNAs to coordinate the spatial organization of diffusible molecules. This mechanism utilizes a unique role for RNA in the

nucleus (relative to DNA or proteins). Specifically, the process of transcription produces many copies of an RNA, which is present at high concentrations in proximity to its transcriptional locus[97,98]. In contrast, proteins are translated in the cytoplasm and therefore lack positional information in the nucleus, and DNA is only present at a single copy and therefore cannot achieve high local concentrations.

Central to this mechanism is the fact that ncRNAs can form high affinity interactions with both protein and RNA immediately following transcription. In this way, they can act to recruit proteins and RNAs within these high concentration spatial territories. In contrast, mRNAs are functional when translated into protein and therefore do not form stable interactions with regulatory molecules in the nucleus. Our results suggest that any RNA that functions independently of its translated product may similarly act as a ncRNA. For example, we note that nascent pre-mRNAs may have protein-coding functions and also form high-affinity interactions within the nucleus that are important for spatial organization. Indeed, we find that histone pre-mRNAs can seed organization of nuclear compartments even though their processed RNAs are also translated into protein products. This role for RNA as a seed for nuclear compartments might also explain formation of other recently described nuclear compartments such as transcriptional condensates[99,100], which inherently produce high levels of RNAs, including enhancer-associated RNAs and pre-mRNAs[101]. Nonetheless, not all ncRNAs — or even all nuclear ncRNAs — act to form compartments around their loci since nuclear ncRNAs can also localize within other regions in the nucleus (e.g. Malat1, scaRNAs, snoRNAs, and snRNAs). Future work will be needed to understand why some specific nuclear RNAs are constrained to local spatial compartments, while others diffuse throughout the nucleus.

This unique role for ncRNAs in the nucleus may explain why certain biological processes utilize ncRNA regulators rather than proteins or DNA. For example, coordinated regulation of multiple genomic DNA targets would be ideally controlled through the expression of a single ncRNA that could localize and recruit regulatory proteins to all of these targets simultaneously. Indeed, many multi-gene regulatory programs, such as X chromosome inactivation and imprinted gene silencing, utilize ncRNAs as regulators (e.g. Xist,

Kcqn1ot1, and Airn). In this way, ncRNAs can increase both the efficiency and specificity of gene regulation by enabling control of multiple target genes through the expression of a single regulatory RNA from its genomic locus. This strategy may also be advantageous even when modulating a single gene because establishment of an RNA territory can recruit effector proteins simultaneously to many genomic regions that are far away in linear distance but proximal in 3D space — including promoters and multiple enhancers — to enable higher concentration and more potent gene regulation. As an example, we observe high concentration of the Pvt1 lncRNA over the Myc gene and all of its known enhancer elements. This coordinated gene regulation model may extend to many of the hundreds of ncRNAs that we identified to be localized within discrete territories in the nucleus.

Taken together, these results provide a global picture of how spatial enrichment of ncRNAs in the nucleus can seed formation of compartments that coordinate the efficiency and specificity of a wide range of essential nuclear functions, including RNA processing, heterochromatin organization, and gene regulation (**Supplemental Figure 5**). While we focused our analysis on ncRNAs in this work, we note that RD-SPRITE can also be applied to measure how gene expression relates to genome organization because it can detect the arrangement of nascent pre-mRNAs relative other RNAs (e.g. enhancer RNAs, pre-mRNAs) and 3D DNA structure. Beyond the nucleus, we anticipate that RD-SPRITE will also provide a powerful method to study the molecular organization, function, and mechanisms of RNA compartments and granules throughout the cell.

LIMITATIONS OF STUDY

We note that there are several technical limitations of the RD-SPRITE method. For example, this approach requires crosslinking, which may lead to potential biases in the types of interactions that are detected. Moreover, because this approach takes a snapshot in time, it cannot measure dynamic events. In addition, while we showed several examples

of RNAs that are required for recruiting diffusible molecules into spatial compartments and identified hundreds more that localized in high concentration territories and therefore are potentially capable of acting in this way, this mechanism may not hold true for every RNA and future work will be needed to explore the functional and mechanistic roles of individual ncRNAs.

4.5. MAIN FIGURES

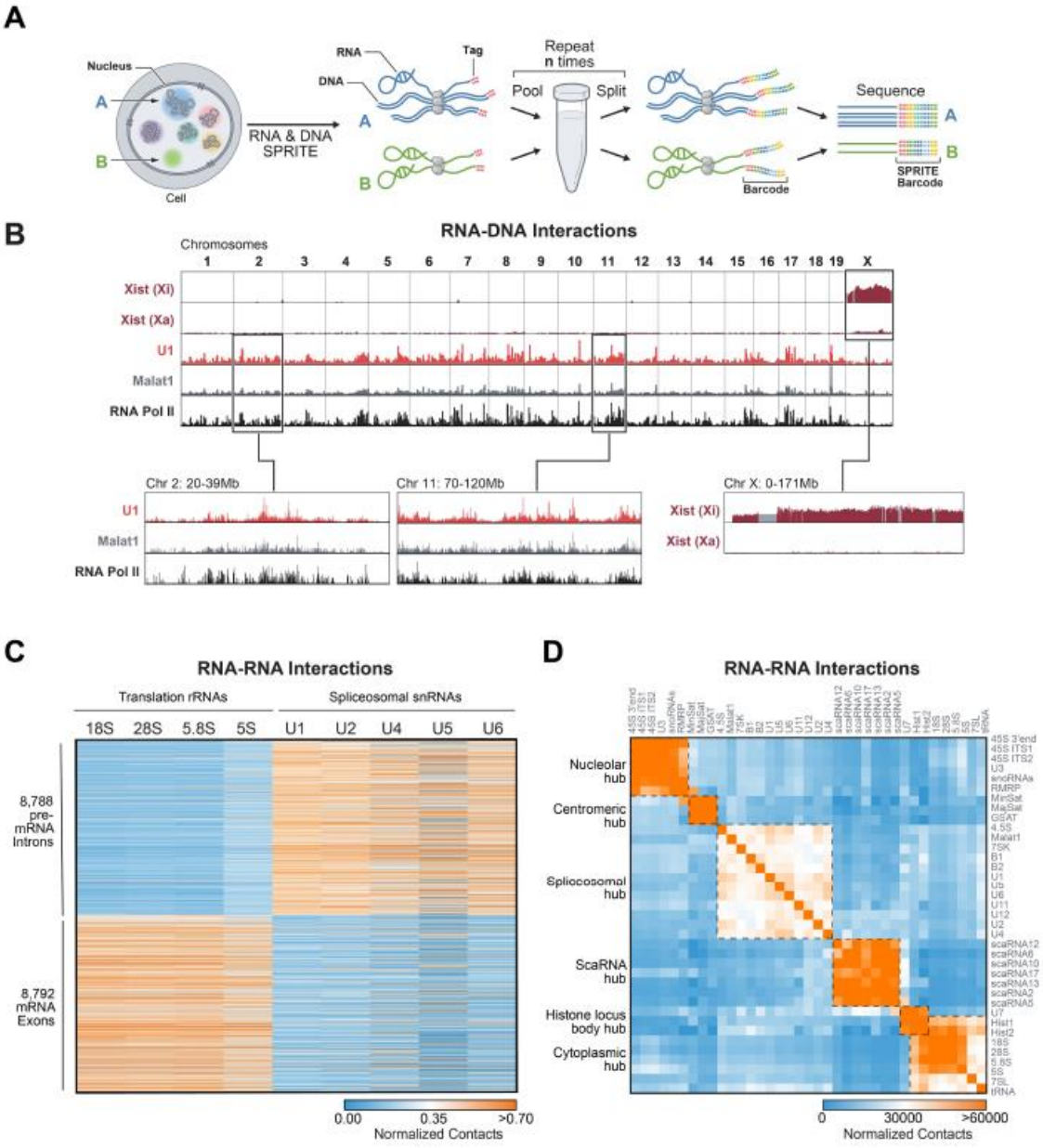


Figure 1: RD-SPRITE generates maps of higher-order RNA and DNA contacts throughout the cell.

(A) Schematic of the RD-SPRITE protocol. Crosslinked cells are fragmented into smaller crosslinked complexes (e.g. A, B). RNA and DNA are each tagged with a DNA-specific

or RNA-specific adaptor sequence (pink). The sample is processed through multiple rounds of split-and-pool barcoding (*n* times), where tag sequences are concatemerized during each round. A series of tags is referred to as a SPRITE barcode. RNA and DNA are sequenced, and barcodes are matched to generate SPRITE clusters to identify groups of interacting molecules. **(B)** RNA-DNA interactions of various non-coding RNAs in mouse embryonic stem (mES) cells. Xist (burgundy) unweighted contacts across the genome in female ES cells where Xist is induced exclusively on the 129 allele (inactive X chromosome; Xi), but not the Castaneous allele (active X chromosome; Xa). U1 spliceosomal RNA (red) and Malat1 lncRNA (grey) weighted contacts across the genome occur at highly transcribed RNA Pol II (ENCODE) genomic regions (black). Insets show zoom-ins of Xist (right) and U1/Malat1 along with genomic localization of RNA Pol II from ENCODE (middle and left). Masked regions on chromosome X plotted in gray. **(C)** A heatmap showing the number of unweighted RNA-RNA contacts between different classes of RNAs. Columns: translation-associated RNAs (18S, 28S, 5.8S, and 5S rRNA) and splicing-associated RNAs (U1, U2, U4, U5, U6 snRNA). Rows: Introns and exons of individual mRNAs. Orange represents high contact frequency and blue represents low contact frequency. **(D)** A heatmap showing unweighted RNA-RNA contact frequencies for several classes of RNAs. Orange represents high contact frequency and blue represents low contact frequency. Groups of pairwise interacting RNAs that have frequent higher-order (multi-way) contacts with each other, but not other groups of RNAs, are referred to as RNA hubs.

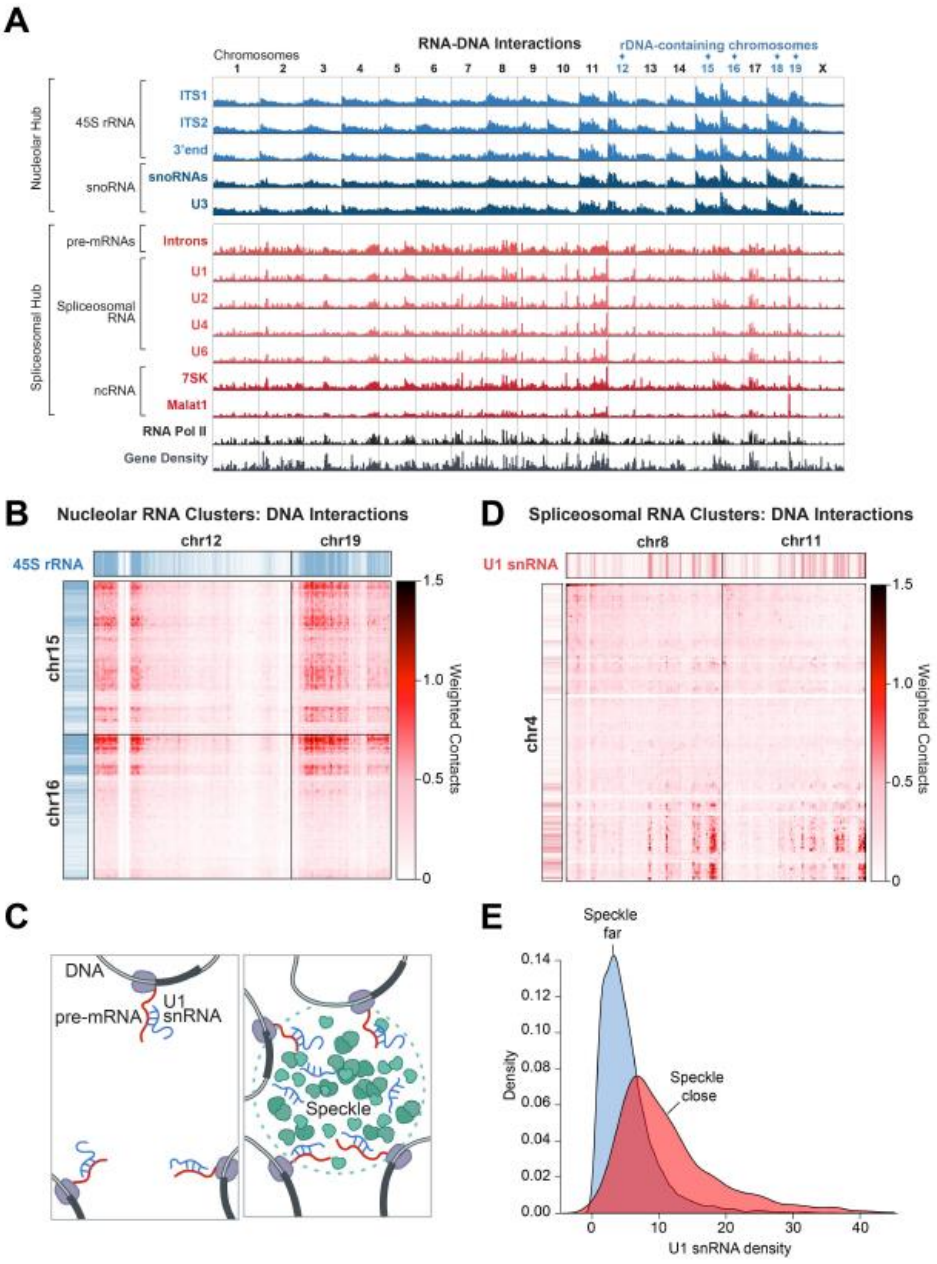


Figure 2: Nucleolar and spliceosomal RNAs form genome-wide interaction hubs.

(A) Genome-wide weighted RNA-DNA contacts (1Mb resolution) for several RNAs within the nucleolar (blue) and spliceosomal (red) hubs. RNA Pol II occupancy from ENCODE (black) is shown along with gene density (gray) across the genome. Chromosomes that contain ribosomal RNA genes at the centromere proximal regions of

the chromosome are demarcated in blue (chr. 12, 15, 16, 18, and 19). **(B)** Weighted DNA-DNA contacts occurring in SPRITE clusters containing nucleolar hub RNAs (e.g 45S pre-rRNAs, snoRNAs, RMRP). Long range, higher-order inter-chromosomal interactions are shown between chromosomes 12 and 19 and chromosomes 15 and 16 for nucleolar hub RNA-containing clusters. Red represents high DNA-DNA contact frequency and white represents low contact frequency. Weighted 45S rRNA-DNA contacts are shown along the top and right axis; blue and white represent high and low contact frequencies, respectively. **(C)** Schematic of two possible snRNA localization models. Model 1: snRNA localization at genomic regions occurs primarily through its association with nascent pre-mRNAs (left). Model 2: snRNA localization depends on 3D spatial position of an individual gene in the nucleus (right). **(D)** Weighted DNA-DNA contacts occurring in SPRITE clusters containing spliceosomal hub RNAs (e.g. U1, U2, Malat1, 7SK). Long range, higher-order inter-chromosomal interactions are shown between regions on chromosome 4 and chromosomes 8 and 11 (examples that have high Pol II occupancy) for all spliceosomal hub RNA-containing clusters. Red represents high DNA-DNA contact frequency and white represents low contact frequency. Weighted U1 snRNA-DNA contacts are shown along the top and right axis; red and white represent high and low contact frequencies, respectively. **(E)** Density of U1 snRNA over genomic DNA regions of actively transcribed Pol II genes separated into genes of comparable transcription levels whose genomic DNA regions are far from nuclear speckles (blue) or close to nuclear speckles (red). Distance from speckle is measured as previously described[55] (see **Methods**).

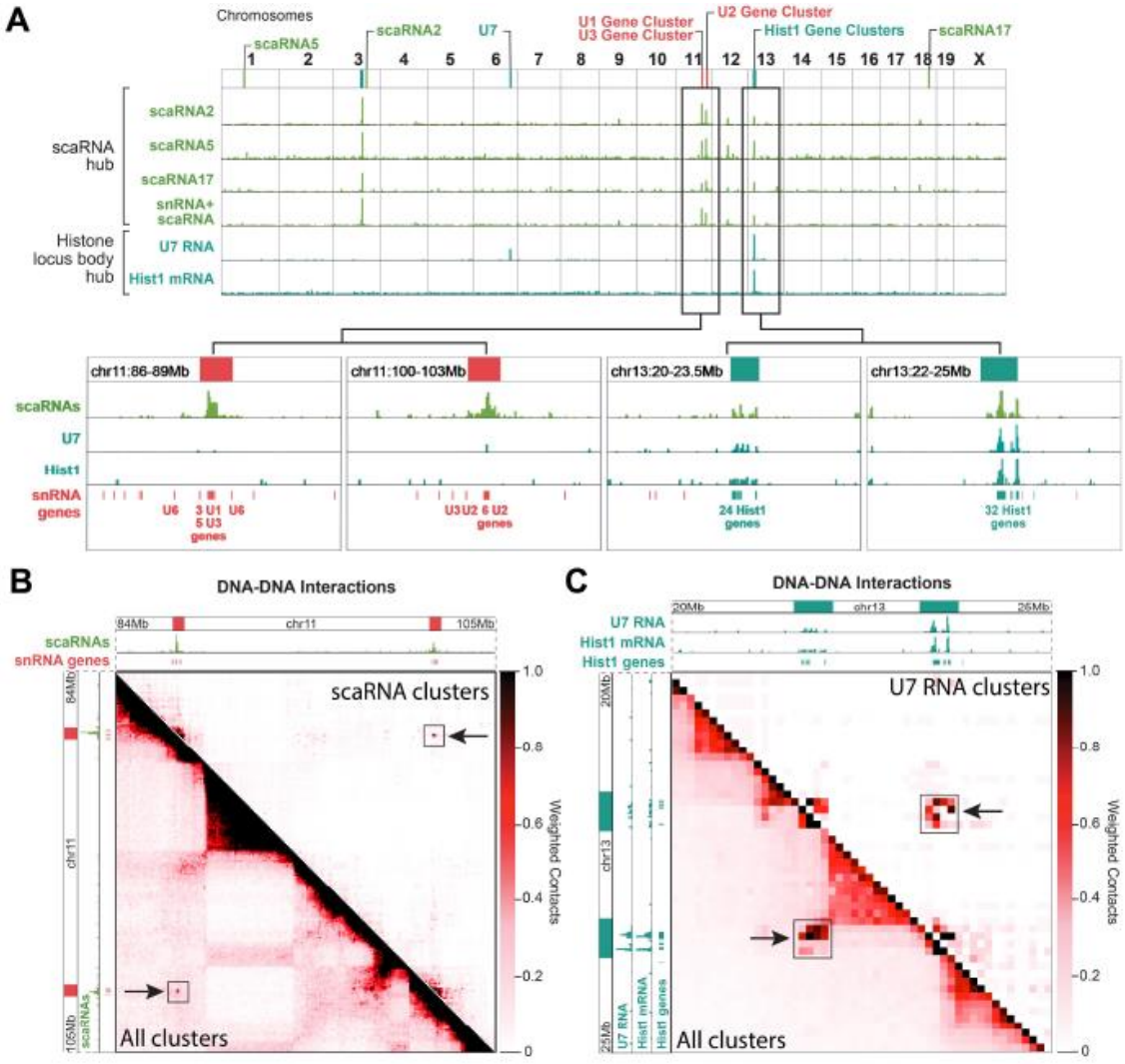


Figure 3: Non-coding RNAs involved in snRNA and histone mRNA biogenesis are spatially organized around snRNA and histone gene clusters.

(A) Weighted RNA-DNA contacts for scaRNA2, scaRNA5 (Gm25395), scaRNA17, or SPRITE clusters containing both scaRNAs and snRNAs (U1/U2) are plotted across the genome in green. Weighed RNA-DNA contacts for U7 snRNA and histone pre-mRNAs (Hist1 mRNA) are plotted in teal. Insets (bottom) show zoom-ins on specific regions of interest (snRNA gene clusters in red, histone gene clusters in teal). Lines (top) show the genomic locations of each RNA plotted and gene clusters of interest. **(B)** Weighted DNA-DNA contacts occurring within all SPRITE clusters (lower diagonal) or only SPRITE

clusters containing scaRNAs (upper diagonal) are shown across a region of chromosome 11 which contains clusters of snRNA genes. The weighted RNA-DNA contacts of scaRNAs are shown along the top and side axes. DNA-DNA contacts occurring between scaRNA-enriched loci are highlighted with black boxes and arrows. **(C)** Weighted DNA-DNA contacts within all SPRITE clusters (lower diagonal) or only SPRITE clusters containing the U7 RNA (upper diagonal) are shown across a region of chromosome 13 which contains histone genes. U7 and histone pre-mRNA occupancy is shown along the top and side axis. DNA-DNA contacts occurring between U7-enriched histone loci are highlighted with black boxes and arrows.

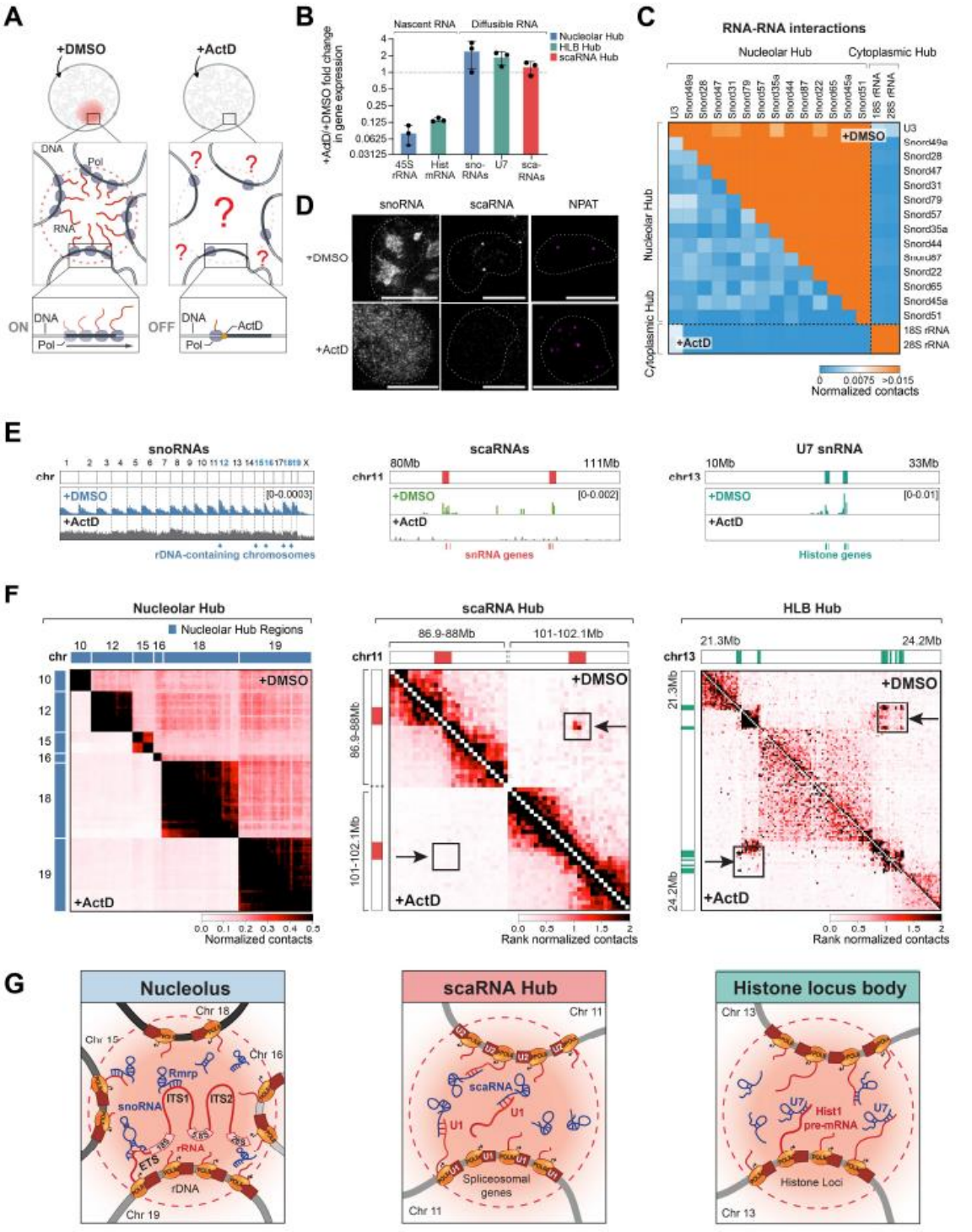


Figure 4: Inhibition of nascent RNA disrupts the spatial organization of RNA processing hubs.

(A) Schematic of transcriptional inhibition of RNA Pol I and Pol II in cells treated with Actinomycin D (+ActD) or control (+DMSO) and resulting effects on RNA-DNA hub organization (red circle). **(B)** Quantification of changes in gene expression following ActD treatment. Nascent transcripts (pre-rRNAs, histone RNAs) are drastically reduced, while diffusible ncRNAs (scaRNAs, snoRNAs, U7) are not. (Note: we cannot distinguish nascent and mature snRNAs using sequencing.) Raw RNA read counts were normalized to 28S rRNA read counts and then scaled to DMSO expression levels (see **Methods**). Error bars represent standard deviation of 3 replicate RD-SPRITE experiments. **(C)** RNA-RNA contact frequency of various snoRNAs contained within the nucleolar hub and cytoplasmic hub RNAs following ActD transcriptional inhibition (lower diagonal) or DMSO-control treatment (upper diagonal). **(D)** RNA FISH (columns 1,2) of nucleolar and scaRNA hub-associated RNAs (snoRNAs, scaRNAs) and IF of HLB-associated NPAT protein (column 3) following ActD treatment (bottom row) or DMSO-control treatment (top row). **(E)** RNA-DNA SPRITE contact profiles for diffusible RNAs within each hub following transcriptional inhibition with ActD. Contact profiles are normalized by RNA expression levels within each sample to enable comparison between treated and untreated samples. (Left) Genome-wide, weighted RNA-DNA contacts for snoRNAs following ActD transcriptional inhibition (+ActD, grey) or control treatment (+DMSO, blue). Contacts for top expressing snoRNAs in clusters size 1001-10000 were aggregated (see **Methods**) (Middle) Weighted RNA-DNA contacts for scaRNAs following ActD transcriptional inhibition (+ActD, grey) or control treatment (+DMSO, green). RNA localization is shown across a region of chromosome 11 which contains snRNA gene clusters (red boxes). (Right) Weighted RNA-DNA contacts for U7 snRNA following ActD transcriptional inhibition (+ActD, grey) or control treatment (+DMSO, teal). RNA localization is shown across a region of chromosome 13 which contains histone gene clusters (teal boxes). **(F)** Weighted DNA-DNA SPRITE contact matrixes at hub-associated genomic locations. (Left) Split DNA-DNA contact matrix for nucleolar-hub associated genomic regions (previously described[55]) on chromosomes 10, 12, 15, 16, 18, and 19 following ActD treatment (lower diagonal) or DMSO-control treatment (upper diagonal). Raw contact frequencies were rescaled to the mean intra-chromosomal contact frequency (see

Methods). (Middle) Split DNA-DNA contact matrix for two regions on chromosome 11 containing snRNA clusters following ActD treatment (lower diagonal) or DMSO-control treatment (upper diagonal). Weighted DNA contact frequencies were rescaled based on rank-ordering to enable comparison between samples (see **Methods**). Locations of snRNA gene clusters are demarcated by red boxes along the top and left side axes. (Right) Split DNA-DNA contact matrix for the region on chromosome 13 containing histone gene clusters following ActD treatment (lower diagonal) or DMSO-control treatment (upper diagonal). Weighted DNA contact frequencies were rescaled based on rank-ordering to enable comparison between samples (see **Methods**). Locations of histone genes are demarcated in teal along the top and left axis. **(G)** Model schematic of how nascent transcription of RNA acts to organize diffusible ncRNAs and genomic DNA to organize the nucleolar hub (left), scaRNA hub (middle), and histone hub (right).

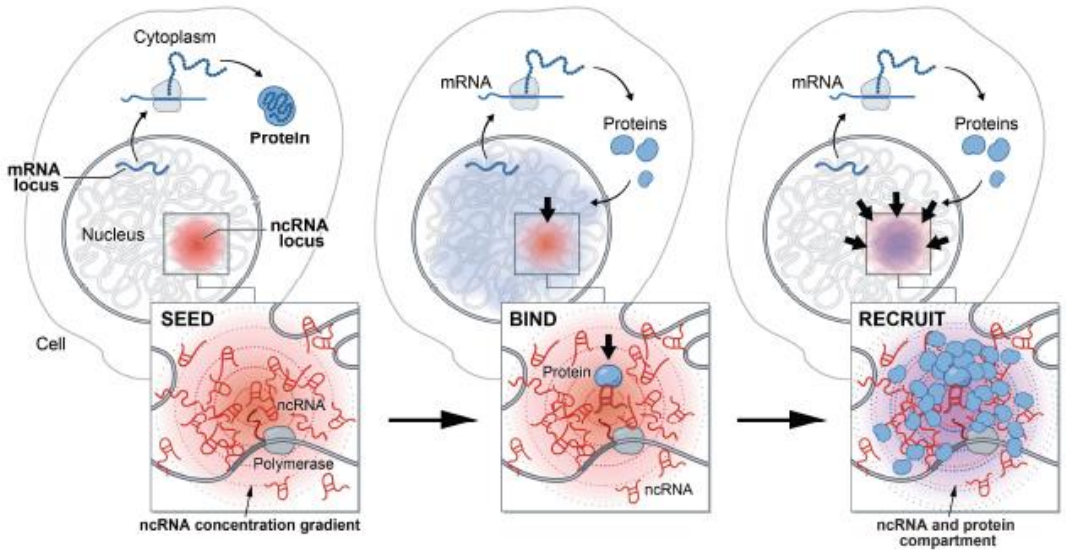
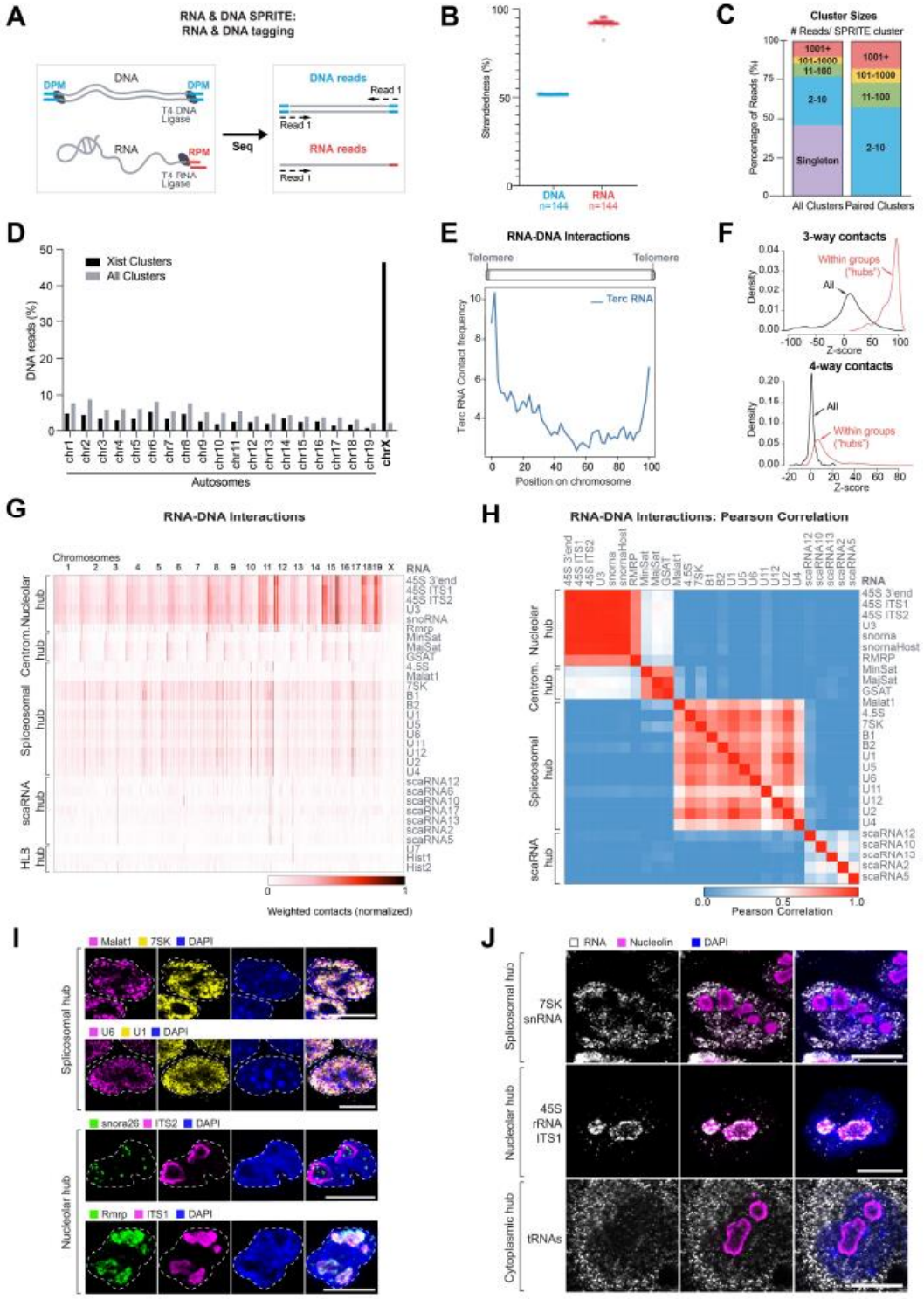


Figure 5: A model for the mechanism by which ncRNAs drive the formation of nuclear compartments.

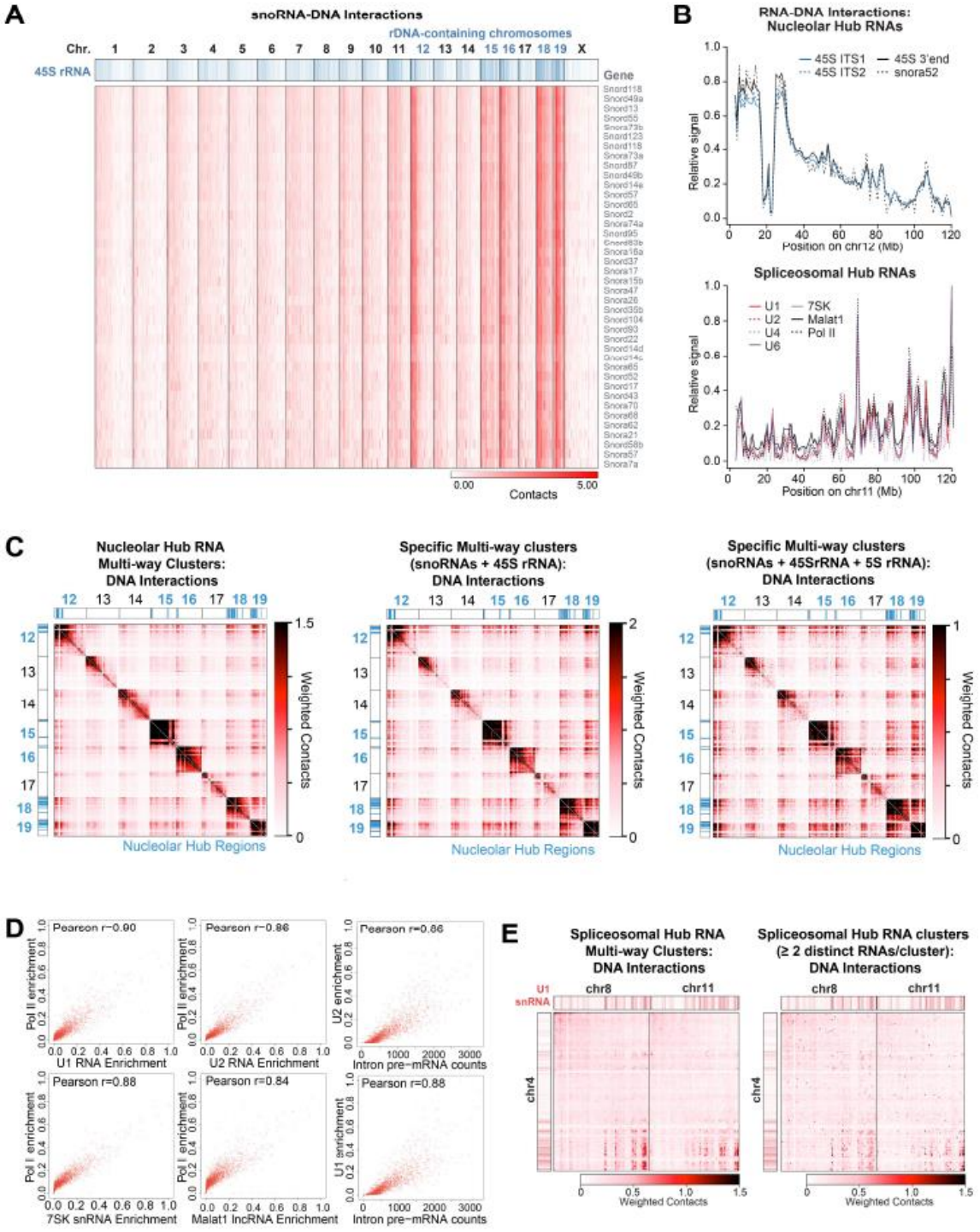
Upon transcription, mRNAs are exported to the cytoplasm (for translation to proteins) while ncRNAs are retained in the nucleus. The process of ncRNA transcription creates a concentration gradient of ncRNA transcript with the highest concentrations near its transcriptional locus (SEED, left panel). Because these RNAs are functional immediately upon transcription and can bind with high affinity to diffusible RNAs and proteins (BIND, middle panel), they can act to change the localization of these other RNAs and proteins to concentrate them in a spatial compartment (RECRUIT, right panel). In this way, ncRNAs may drive the organization of regulatory and functional nuclear compartments.

4.6. SUPPLEMENTAL FIGURES



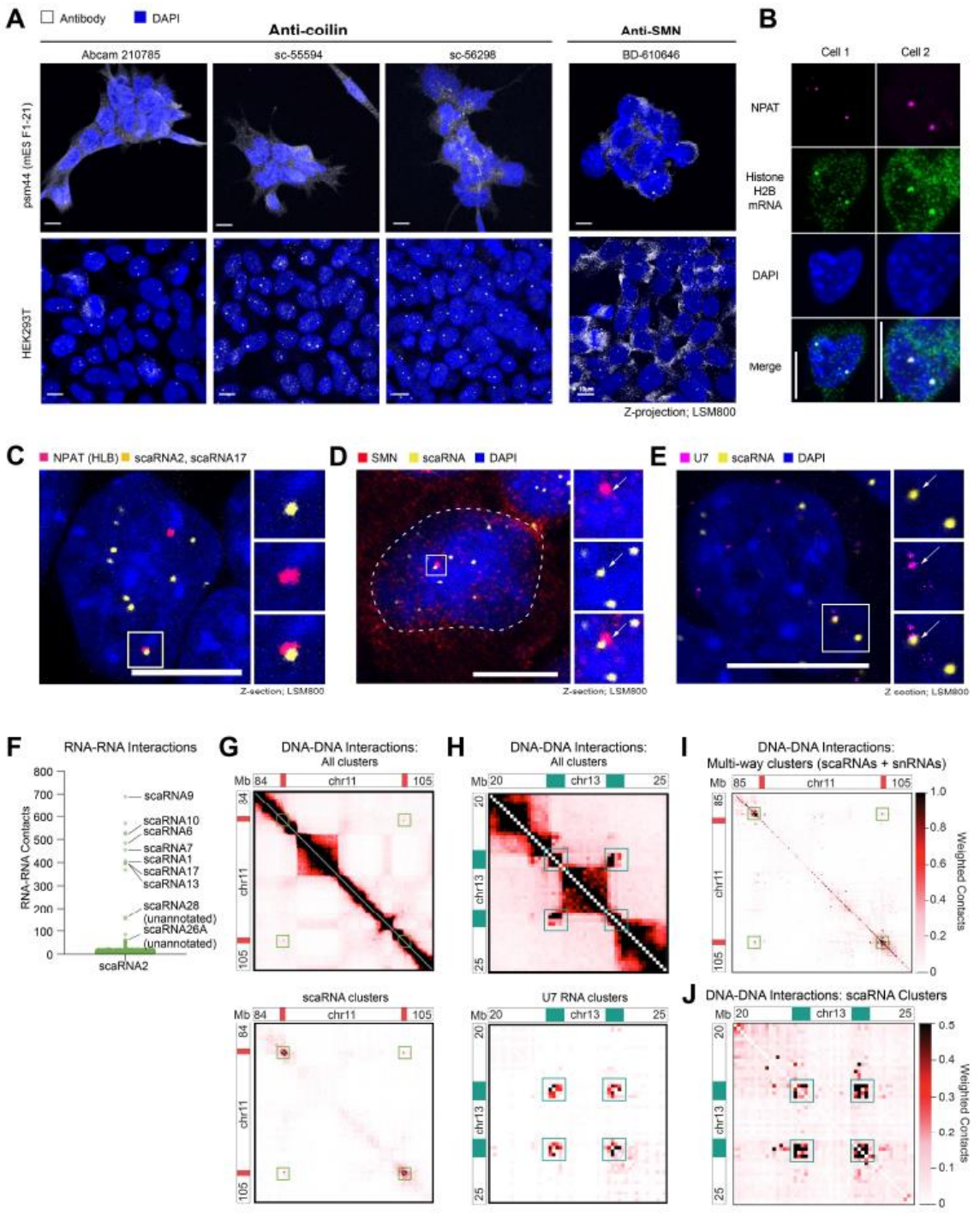
Supplemental Figure 1: RD-SPRITE accurately measures RNA and DNA contacts. (A) Schematic of tagging used to identify DNA- and RNA-specific reads through sequencing. DNA and RNA are each tagged with sequence-specific tags, namely "DNA Phosphate Modified" (DPM) tag and "RNA Phosphate Modified" (RPM) tags using T4 DNA and RNA Ligase, respectively. DNA is double stranded and therefore DPM will be read from both strands, while RNA is single stranded and therefore RPM will be read only from 1 strand. Additionally, the RPM and DPM tags have identical dsDNA sticky ends that enable subsequent split-pool barcoding with the same SPRITE tags. **(B)** The percentage of reads aligning to each DNA strand based on their DPM tag (DNA reads) or RPM tag (RNA reads) is shown across 144 independently amplified and sequenced SPRITE libraries from four SPRITE experiments (technical replicates). **(C)** Percentage of reads in SPRITE clusters of different sizes, stratified into categories of clusters containing 1, 2-10, 11-100, 101-1000, and 1001+ reads per cluster. Distributions shown for all clusters (left) and paired clusters (2+ reads per cluster) (right). **(D)** Percentage of DNA reads aligning to each chromosome from SPRITE clusters containing the Xist lncRNA (black) as compared to all SPRITE clusters (gray). **(E)** The aggregate unweighted RNA-DNA contact frequency of the Telomerase associated RNA Component (Terc) across all chromosomes. **(F)** Multiway contact analysis statistics for 3-way and 4-way RNA contacts co-occurring in SPRITE clusters. We calculated the expected frequency of multiway contacts if RNAs associated at random (n=100 iterations) versus the observed frequency within the RD-SPRITE dataset (see **Methods**). Z-scores are shown for 3-way (top) or 4-way (bottom) contacts among all RNAs (all, black) or RNAs within the same "group" (within group, red), defined by sets of pairwise interacting RNAs (Figure 1D). **(G)** Weighted genomic DNA localization heatmap of individual RNAs belonging to distinctive nuclear hubs. RNAs are organized by their RNA hub occupancy (shown in Figure 1D). Contacts are normalized from 0 to 1 to account for expression levels of each RNA. **(H)** Pearson correlation of RNA-DNA unweighted contact frequencies across the genome for all pairs of RNAs within the nuclear hubs (nucleolar, centromeric, spliceosomal, and scaRNA hubs). Red represents high correlation and blue represents low correlation. **(I)** RNA FISH of various non-coding RNAs within the spliceosomal hub (top rows) or nucleolar hub (bottom rows).

Spliceosomal hub (top, row 1): Malat1 lncRNA and 7SK RNA and (bottom, row 2): U6 and U1 spliceosomal RNAs. Nucleolar hub (top, row 3): snora26 snoRNA and 45S pre-rRNA ITS2 and (bottom, row 4): RNase MRP (Rmrp) and 45S pre-rRNA ITS1. Far-left and left-middle panels show individual RNAs; right-middle panel shows DAPI; far-right panels show overlays. Dashed lines demarcate the nuclear boundary identified with DAPI. Scalebar is 10μm. **(H)** RNA FISH (left) of specific, hub-associated ncRNA along with nucleolin immunofluorescence (middle) and DAPI (right). 7SK snRNA (top), ITS1 regions of 45S pre-rRNA (middle) and tRNAs (bottom). tRNAs are visualized using pooled RNA FISH probes (see **Methods**). Scalebar is 10μm.



Supplemental Figure 2: Nucleolar and spliceosomal hubs show higher-order interactions around ribosomal RNA genes and genomic regions with a high density of actively transcribed genes, respectively. (A) Genome-wide localization of each

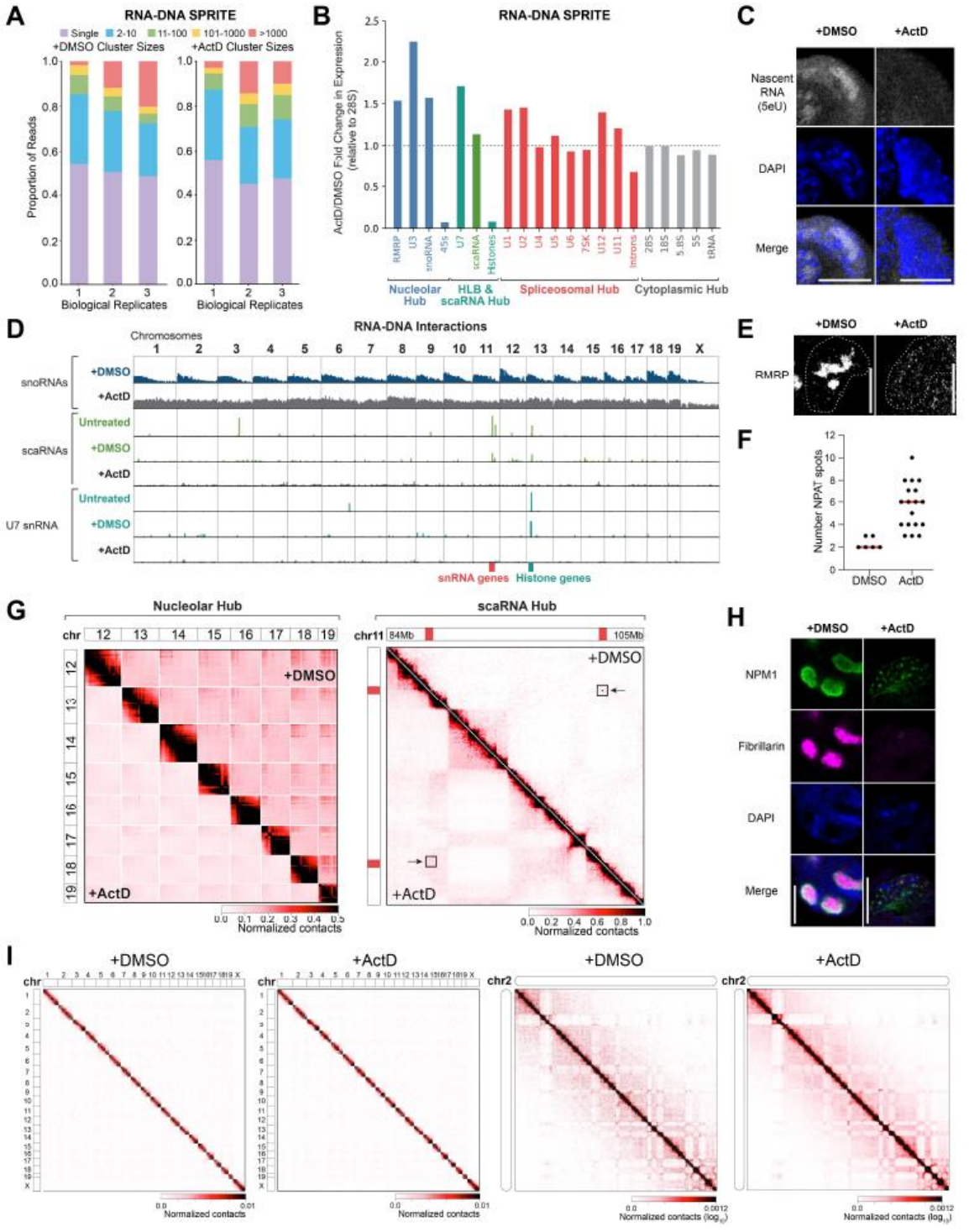
individual snoRNA, as determined by unweighted RNA-DNA contact frequency. Blue track shows 45S pre-rRNA localization on DNA. Chromosomes containing ribosomal DNA (rDNA) genes (chromosomes 12, 15, 16, 18, 19) are denoted in blue. **(B)** (Top) Overlay of RNA-DNA contact frequencies on chromosome 12 for various RNAs within the nucleolar hub. (Bottom) Overlay of RNA-DNA contact frequencies on chromosome 11 for various RNAs within the spliceosomal hub. **(C)** Weighted DNA-DNA contact heatmap is shown for SPRITE clusters containing any of the RNAs within the nucleolar hub (left), both snoRNAs and 45S pre-rRNA (middle), and snoRNAs, 45S, and 5S (right) simultaneously. (**D**) (Columns 1,2) Genome-wide 1Mb enrichment of several spliceosomal hub RNA-DNA interactions (U1 snRNA, U2 snRNAs, 7SK RNA, and Malat1 lncRNA) compared to enrichment of Pol II ChIP-seq signal (ENCODE). (Column 3) Genome-wide 1Mb enrichment of spliceosomal hub RNA-DNA interactions (U1 snRNA, U2 snRNA) compared to counts of pre-mRNA RNA-DNA interactions. Pearson correlation scores are provided for each set of comparisons. (**E**) Weighted DNA-DNA contacts that co-occur in a SPRITE cluster with at least one RNA in the splicing hub (left) or multiple (2 or more) RNAs in the splicing hub are shown (right). Weighted U1 snRNAs contacts on DNA are shown as a heatmap (red-white scale) along the top and side axes.



Supplemental Figure 3: Spatial relationship between snRNA biogenesis hub and histone locus bodies. (A) Immunofluorescence imaging of classical Cajal Body (Coilin) and nuclear gem (SMN) markers in mouse ES cells (top) and HEK293T cells (bottom).

Mouse ES cells do not contain visible Coilin foci for any of the three anti-Coilin antibodies tested. In contrast, HEK293T cells show visible Coilin foci. SMN foci, which are markers for nuclear Gemini of Cajal bodies ("gems"), are present in both mouse ES cells and HEK293T cells. Scalebar is 10μm. **(B)** IF of NPAT (magenta), RNA FISH of Histone H2B mRNA (green), nuclear stain with DAPI (blue) and overlaid images for two representative mES cells. NPAT and Histone H2B mRNA colocalize within the nucleus. Scalebar is 10μm. **(C)** Combined IF and RNA FISH image of a mouse ES cell co-stained for NPAT protein (magenta) and scaRNAs (pooled scaRNA2 and scaRNA17 probes, yellow) within the nucleus (DAPI). Inset shows an example of scaRNA localization near NPAT foci. Scalebar is 10μm. **(D)** Combined IF and RNA FISH image of a mouse ES cell co-stained for SMN protein (red) and scaRNAs (pooled scaRNA2 and scaRNA17 probes, yellow) within the nucleus (DAPI). Inset shows an example of scaRNA localization near SMN foci (arrow). Scalebar is 10μm. **(E)** RNA FISH image of mouse ES cell with probes targeting U7 (purple) and scaRNAs (pooled scaRNA2 and scaRNA17 probes, yellow) within the nucleus (DAPI). Inset shows an example of scaRNA localization near U7 (arrow). Scalebar is 10μm. **(F)** RNA-RNA contact frequency between scaRNA2 and all RNAs. Top hits include annotated scaRNAs and two previously unannotated scaRNAs, which we identified (see **Supplemental Methods**). **(G)** Weighted DNA-DNA contacts within all SPRITE clusters (top) and within SPRITE clusters containing scaRNAs (bottom) are shown across a region on chromosome 11 which contains snRNA gene clusters. scaRNA occupancy is demarcated with solid red boxes along the top and left axis. Contacts within or between snRNA gene clusters are outlined with green boxes. **(H)** Weighted DNA-DNA contacts with all SPRITE clusters (top) and within only SPRITE clusters containing the U7 snRNA (bottom) are shown across a region on chromosome 13 which contains the two Hist1 gene clusters. U7 and Hist1 RNA occupancy is demarcated with solid teal boxes along the top and left axis. Contacts within or between histone gene clusters are outlined in teal boxes. (**I**) Weighted DNA-DNA contacts within SPRITE clusters containing reads from both scaRNAs and snRNAs are shown across the same snRNA gene cluster containing region on chromosome 11 as panel (G). **(J)** Weighted DNA-DNA

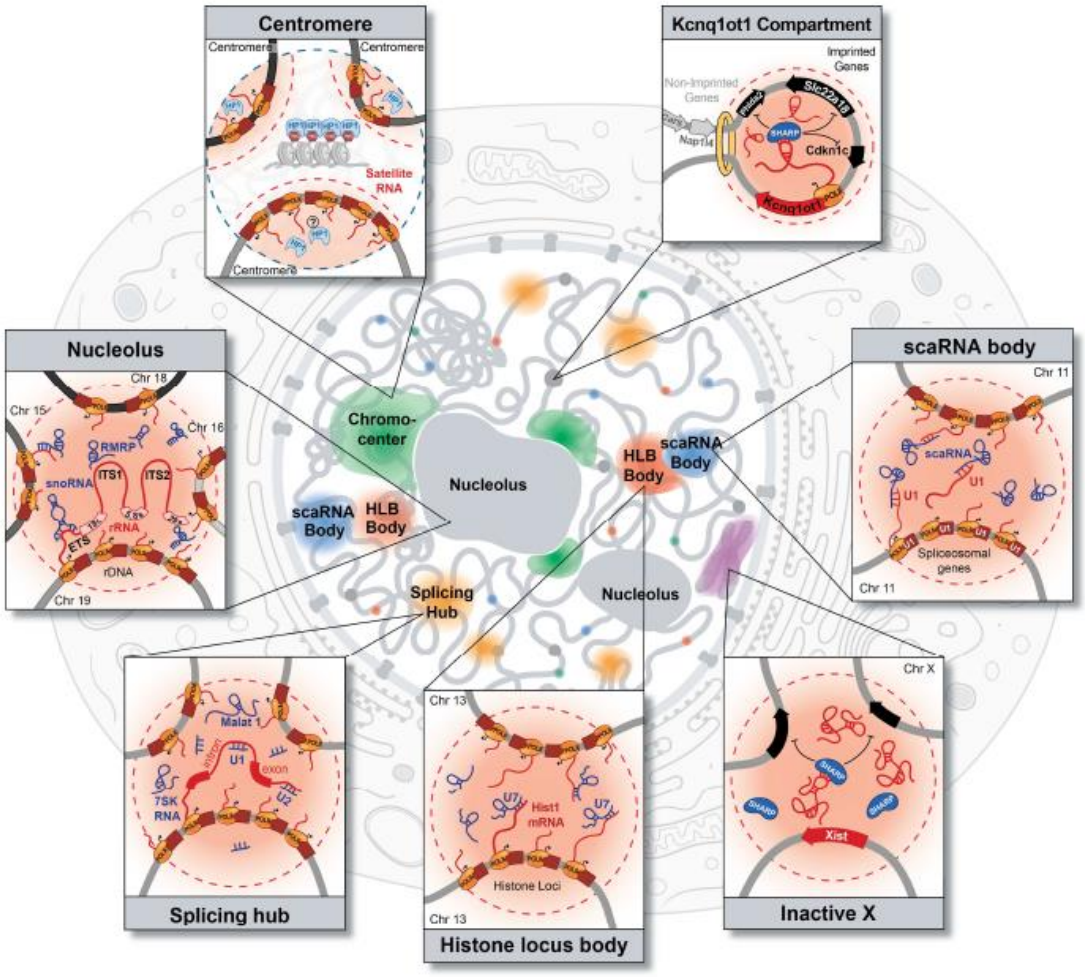
contacts within SPRITE clusters containing scaRNAs is shown across the same histone gene cluster containing region on chromosome 13 as (H).



Supplemental Figure 4: Transcriptional Inhibition with Actinomycin D leads to structural changes in the Nucleolar Hub, scaRNA Hub, and HLB Hubs. (A) Cluster size distribution in RD-SPRITE for DMSO-treated (left) and ActD-treated (right) samples.

Independent results from three biological replicates are shown. **(B)** Fold-changes in gene expression upon ActD treatment compared to control DMSO-treated samples for RNAs in the nucleolar, HLB, scaRNA, spliceosomal, and cytoplasmic hubs. Gene expression changes were computed in RD-SPRITE clusters containing 2-1000 reads/cluster. Raw RNA counts were normalized to 28S rRNA counts to account for differences in read depth prior to computing the ratio of ActD to DMSO counts. (see **Methods**) **(C)** Microscopy image of nascent RNA in DMSO-treated cells or ActD-treated cells. Nascent transcription was visualized by incubating cells with 5EU and labeling with click chemistry (see **Methods**). Scalebar is 10μm**. (D)** Genome-wide, weighted RNA-DNA contact frequencies for hub-associated RNAs in RD-SPRITE. (Top) DNA localization of snoRNAs following ActD transcriptional inhibition (+ActD, grey) or control treatment (+DMSO, blue). Contacts for top expressing snoRNAs in SPRITE clusters of size 1001-10000 reads were aggregated (see **Methods**) (Middle) DNA localization for scaRNAs following ActD transcriptional inhibition (+ActD, grey) or control treatment (+DMSO, green). (Bottom) DNA localization of U7 snRNA following ActD transcriptional inhibition (+ActD, grey) or control treatment (+DMSO, teal). Untreated tracks are from the original RD-SPRITE dataset used in this study. (**E**) RNA FISH of Rnase MRP (RMRP) following ActD treatment or DMSO-control treatment. Dashed lines demarcate the nuclear boundary identified with DAPI**.** (**F)** Quantification of the mean (red line) number of NPAT spots (HLBs) per cell in IF stained cells following ActD or DMSO-control treatment. DMSO: n=6 cells; ActD: n=18 cells. **(G)** DNA-DNA contact matrices generated by DNA-SPRITE at different hub-associated regions following ActD treatment (lower diagonal) or DMSO-control treatment (upper diagonal). (Left) Weighted contact matrixes from SPRITE clusters of size 2-10K reads for chromosomes 12-19. Raw contact frequencies were rescaled to the mean intra-chromosomal contact frequency (see **Methods**). (Right) Weighted contact matrixes from SPRITE clusters of size 2-1000 reads for a region on Chromosome 11 spanning two snRNA gene clusters. Raw contact frequencies were rescaled based on rank-ordering (see **Methods**). **(H)** IF stain for NPM1 (green), IF stain for Fibrillin (pink), nuclear stain with DAPI (blue) and overlayed images in DMSO-control treated cells (left) or ActD treated cells (right). Scalebar is 10μm. **(I)** (Left) Genome-wide,

weighted DNA-SPRITE contact frequencies in SPRITE clusters of size 2-1000 reads for ActD or DMSO-control treated samples. (Right) Weighted DNA-SPRITE contact frequencies on chromosome 2 in SPRITE clusters of size 2-1000 reads measured by DNA-SPRITE for ActD or DMSO-control treated samples.



Supplemental Figure 5. A widespread role for ncRNAs in shaping compartments throughout the nucleus that are associated with various nuclear functions. A model schematic of the localization of the different nuclear compartments within the nucleus and the molecular components contained within them. In each of these cases, an RNA seeds organization by achieving high concentration in spatial proximity to its transcriptional locus. This leads to the formation of nuclear compartments associated with RNA processing, heterochromatin assembly, and gene regulation.

4.7. SUPPLEMENTAL TABLE LEGENDS

Supplemental Table 1: Multi-way (k-mer) contact score statistics for RD-SPRITE. To access the significance of a multi-way interactions between RNAs within the RD-SPRITE dataset, we designed a mutli-way contact score analysis (see **Methods)**. Hubs were defined as higher-order, multi-way structures with many significant multi-way contacts.

Supplemental Table 2: Read alignment statistics for RD-SPRITE. Alignment statistics to the mouse genome for DPM-tagged (DNA) reads or RPM-tagged (RNA) reads from individual RD-SPRITE experiments and libraries. Unaligned or low MAPQ score aligned (ie multi-mapping) RPM-tagged reads (Repeat) were subsequently aligned to a custom reference genome of repeat RNA sequences (see **Methods**).

Supplemental Table 3: Barcode Identification statistics for RD-SPRITE and DNA-SPRITE. The complete barcode of each read is identified from read 2 (see **Supplemental Figure 1A).** The percentage of reads with 0, 1, 2 … up to n (where n is the maximum number of possible) tags is reported for each individual SPRITE library. This represents a quality metric and is included as an output in the processing pipeline for RD-SPRITE or DNA-SPRITE (see **Methods)**.

Supplemental Table 4: Template for calculating read depth for sequencing SPRITE libraries. To determine the amount of reads required to sequence each SPRITE library aliquot to saturation, we estimate the number of unique molecules (pre-PCR) using the final library concentrations. We typically sequence each library 1.5-2x coverage.

4.8. SUPPLEMENTAL NOTES

Supplemental Note 1: ***RD-SPRITE improves efficiency of RNA tagging.*** Although our previous version of SPRITE could map both RNA and DNA, it was limited primarily to detecting highly abundant RNA species (e.g. 45S pre-rRNA). In RD-SPRITE, we have improved detection of lower abundance RNAs by increasing yield through the following adaptations. (i) We increased the RNA ligation efficiency by utilizing a higher concentration of RPM, corresponding to ~2000 molar excess during RNA ligation. (ii) Adaptor dimers that are formed through residual purification on our magnetic beads lead to reduced efficiency because they preferentially amplify and preclude amplification of tagged RNAs. To reduce the number of adaptor dimers in library generation, we introduced an exonuclease digestion of excess reverse transcription (RT) primer that dramatically reduces the presence of the RT primer. (iii) Reverse transcription is used to add the barcode to the RNA molecule, yet when RT is performed on crosslinked material it will not efficiently reverse transcribe the entire RNA (because crosslinked proteins will act to sterically preclude RT). To address this, we performed a short RT in crosslinked samples followed by a second RT reaction after reverse crosslinking to copy the remainder of the RNA fragment. (iv) Because cDNA is single stranded, we need to ligate a second adaptor to enable PCR amplification. The efficiency of this reaction is critical for ensuring that we detect each RNA molecule. We significantly improved cDNA ligation efficiency by introducing a modified “splint” ligation. Specifically, a double stranded “splint” adaptor containing the Read1 Illumina priming region and a random 6mer overhang is ligated to the 3’end of the cDNA at high efficiency by performing a double stranded DNA ligation. This process is more efficient than the single stranded DNA-DNA ligation previously utilized[55]. (v) Finally, we found that nucleic acid purification performed after reverse crosslinking leads to major loss of complexity because we lose a percentage of the unique molecules during each cleanup. In the initial RNA-DNA SPRITE protocol there were several column (or bead) purifications utilized to remove enzymes and enable the next enzymatic reaction. We reduced these cleanups by introducing biotin modifications into the DPM and RPM adaptors that enable binding to streptavidin beads and for all subsequent

molecular biology steps to occur on the same beads. Together, these improvements enabled a dramatic improvement of our overall RNA recovery and enables generation of high complexity RNA/DNA structure maps.

Supplemental Note 2: ***The snRNA biogenesis hub may be similar to the Cajal body.*** We note that the snRNA biogenesis hub may be similar to Cajal bodies, which have been noted to contain snRNA genes and scaRNAs[80,82,84,102,103]. However, Cajal bodies are traditionally defined by the presence of Coilin foci in the nucleus[78,82,104] and based on this definition, our mES cells do not contain visible Cajal bodies with all three antibodies tested (**Supplemental Figure 3A**). Despite the absence of traditionally defined Cajal bodies, our data suggest that snRNA biogenesis hubs do indeed exist and form around snRNA gene loci, even in the absence of observable Coilin foci. Our data suggest that scaRNA localization more accurately defines snRNA processing bodies relative to Coilin. Consistent with this idea, scaRNAs have a clearly defined functional role in snRNA biogenesis whereas Coilin is dispensable for snRNA biogenesis[84]. It is also possible that these snRNA processing bodies are distinct from Cajal bodies, which may represent a different nuclear structure. For example, these might represent nuclear gems[105], which contain SMN protein, or "residual bodies," which are Coilin negative[78,106]. We note that we observe SMN foci in our mES cells and that some, but not all, scaRNAs colocalize with SMN protein in the nucleus (**Supplemental Figure 3A, D**). Additionally, we observed that scaRNAs also localize to histone gene clusters, form higher-order DNA interactions, and are adjacent to the HLB in the nucleus (**Figure 3A, Supplemental Figure 3C-E, J**). This is consistent with previous observations that HLBs and Cajal bodies are often found adjacent to each other in the nucleus[78,83].

Supplemental Note 3: ***RD-SPRITE measures the frequency at which RNAs are contacting chromatin.*** Although data from previous methods have reported that both lncRNAs and mRNAs are similarly enriched on chromatin at their transcriptional loci, we observed a striking difference in chromatin localization between these classes of RNA. The major reason for this is because RD-SPRITE measures RNA localization within all compartments of the cell, including in the nucleus and cytoplasm. Accordingly, we can

compute a chromatin enrichment score, which we define as the frequency at which a given RNA is localized on chromatin. Other RNA-DNA mapping methods such as hybridization (e.g. RAP, ChIRP) or proximity-ligation (e.g. GRID-Seq, Margi) methods exclusively measure RNA when they are present on chromatin and therefore cannot measure this differential localization frequency.

4.9. METHODS

Cell line generation, cell culture, and drug treatments

Cell lines used in this study. We used the following cell lines in this study: (i) Female ES cells (*pSM44* ES cell line) derived from a 129 × castaneous F1 mouse cross. These cells express Xist from the endogenous locus under control of a tetracycline-inducible promoter. The dox-inducible Xist gene is present on the 129 allele, enabling allele-specific analysis of Xist induction and X chromosome silencing. (ii) Female ES cells, where we replaced the endogenous Kcnq1ot1 promoter with a tetracycline-inducible promoter (*Kcnq1ot1-inducible* ES cell line). In the absence of Doxycycline, these cells do not express Kcnq1ot1; in the presence of Doxycycline, these cells express Kcnq1ot1. (iii) Female ES cells containing dCas9 fused to 4-copies of the SID transcriptional repression domain integrated into a single locus in the genome (dCas9-4XSID). (iv) HEK293T, a female human embryonic kidney cell line (ATCC Cat# CRL-3216, RRID:CVCL_0063).

Cell culture conditions. All mouse ES cell lines were cultured in serum-free 2i/LIF medium as previously described[55]. HEK293T cells were cultured in complete media consisting of DMEM (GIBCO, Life Technologies) supplemented with 10% FBS (Seradigm Premium Grade HI FBS, VWR), 1X penicillin-streptomycin (GIBCO, Life Technologies), 1X MEM non-essential amino acids (GIBCO, Life Technologies), 1 mM sodium pyruvate (GIBCO, Life Technologies) and maintained at 37°C under 5% CO_2. For maintenance, 800,000 cells were seeded into 10 mL of complete media every 3-4 days in 10 cm dishes. HEK293T cells were used for human-mouse mixing experiments to assess noise during the SPRITE procedure as well as for imaging Coilin foci.

Doxycycline Inducible Xist Cell Line Development. Female ES cells (F1 2-1 line, provided by K. Plath) were CRISPR-targeted (nicking gRNA pairs TGGGCGGGAGTCTTCTGGGCAGG and GGATTCTCCCAGGCCCAGGGCGG) to integrate the Tet transactivator (M2rtTA) into the Rosa26 locus using R26P-M2rtTA, a gift from Rudolf Jaenisch (Addgene plasmid #47381). This line was subsequently CRISPR-

targeted (nicking gRNA pairs GCTCGTTTCCCGTGGATGTG and GCACGCCTTTAACTGATCCG) to replace the endogenous Xist promoter with tetracycline response elements (TRE) and a minimal CMV promoter as previously described[107]. The promoter replacement insertion was verified by PCR amplification of the insertion locus and Sanger sequencing of the amplicon. SNPs within the amplicon allowed for allele identification of the insertion, confirming that the 129 allele was targeted and induced Xist expression. We routinely confirmed the presence of two X chromosomes within these cells by checking the presence of X-linked SNPs on the 129 and castaneous alleles.

Doxycycline induction. Xist and Kcnq1ot1 expression were induced in their respective cell lines by treating cells with 2 μg/mL doxycycline (Sigma D9891). Xist was induced for 24 hours prior to crosslinking and analysis. Kcnq1ot1 was induced for 12-16hrs prior to RNA harvesting for qRT-PCR or induced for 24hrs prior to cell crosslinking with 1% formaldehyde for ChIP-seq.

Actinomycin D (ActD) Treatment. ActD transcriptional inhibition was performed by culturing cells in 25 μg/mL ActD (Sigma A9415, 25 μL of 1 mg/mL stock added per 1 mL culture medium) or DMSO for 4 hours before cells were processed for RNA-FISH, IF or SPRITE. The concentrations for imaging and for SPRITE were the same and the same stocks were used for all experiments.

Antibodies

Antibodies. Primary antibodies used in the study: anti-Nucleolin (Abcam Cat# ab22758, RRID:AB_776878, 1:500); anti-NPAT (Abcam Cat# ab70595, RRID:AB_1269585, 1:100); anti-SMN (BD Biosciences Cat# 610646, RRID:AB_397973, 1:100); anti-HP1ß (Active Motif Cat# 39979, RRID:AB_2793416, 1:200); anti-Coilin (Abcam Cat # ab210785; Santa Cruz Biotechnology Cat# sc-55594, RRID:AB_1121780; Santa Cruz Biotechnology Cat# sc-56298, RRID:AB_1121778; 1:100); anti-Sharp (Bethyl Cat# A301-119A, RRID:AB_873132, 1:200); anti-Histone H3K27ac (Active Motif Cat# 39134,

RRID:AB_2722569); anti-NPM1 (Abcam Cat# ab10530, RRID:AB_297271; 1:200); anti-Fibrillarin (Abcam Cat# ab5821, RRID:AB_2105785; 1:200); anti-LaminB1 (Abcam Cat# ab16048, RRID:AB_10107828; 1:1000); For imaging studies, all antibodies were diluted in blocking solution.

RNA & DNA-SPRITE

RD-SPRITE is an adaptation of our initial SPRITE protocol[55] with significant improvements to the RNA molecular biology steps that enable generation of higher complexity RNA libraries. The approach was performed as follows:

Crosslinking, lysis, sonication, and chromatin digestion. Cells were lifted using trypsinization and were crosslinked in suspension at room temperature with 2 mM disuccinimidyl glutarate (DSG) for 45 minutes followed by 3% Formaldehyde for 10 minutes to preserve RNA and DNA interactions *in situ*. After crosslinking, the formaldehyde crosslinker was quenched with addition of 2.5M Glycine for final concentration of 0.5M for 5 minutes, cells were spun down, and resuspended in 1x PBS + 0.5% RNAse Free BSA (AmericanBio AB01243-00050) over three washes, 1x PBS + 0.5% RNAse Free BSA was removed, and flash frozen at -80C for storage. We found that RNAse Free BSA is critical to avoid RNA degradation. RNase Inhibitor (1:40, NEB Murine RNAse Inhibitor or Thermofisher Ribolock) was also added to all lysis buffers and subsequent steps to avoid RNA degradation. After lysis, cells were sonicated at 4-5W of power for 1 minute (pulses 0.7 second on, 3.3 seconds off) using the Branson Sonicator and chromatin was fragmented using DNAse digestion to obtain DNA of approximately ~150bp-1kb in length.

Estimating molarity. After DNase digestion, crosslinks were reversed on approximately 10 μL of lysate in 82 μL of 1X Proteinase K Buffer (20 mM Tris pH 7.5, 100 mM NaCl, 10 mM EDTA, 10 mM EGTA, 0.5% Triton-X, 0.2% SDS) with 8 μL Proteinase K (NEB) at 65°C for 1 hour. RNA and DNA were purified using Zymo RNA Clean and Concentrate columns per the manufacturer's specifications (>17nt protocol) with minor adaptations,

such as binding twice to the column with 2X volume RNA Binding Buffer combined with by 1X volume 100% EtOH to improve yield. Molarities of the RNA and DNA were calculated by measuring the RNA and DNA concentration using the Qubit Fluorometer (HS RNA kit, HS dsDNA kit) and the average RNA and DNA sizes were estimated using the RNA High Sensitivity Tapestation and Agilent Bioanalyzer (High Sensitivity DNA kit).

NHS bead coupling. We used the RNA and DNA molarity estimated in the lysate to calculate the total number of RNA and DNA molecules per microliter of crosslinked lysate. We coupled the lysate to ~10mL of NHS-activated magnetic beads (Pierce) in 1x PBS + 0.1% SDS combined with 1:40 dilution of NEB Murine RNase Inhibitor overnight at 4°C as previously described[55]. We coupled at a ratio of 0.25-0.5 molecules per bead to reduce the probability of simultaneously coupling multiple independent complexes to the same bead, which would lead to their association during the split-pool barcoding process. Because multiple molecules of DNA and RNA can be crosslinked in a single complex, this estimate is a more conservative estimate of the number of molecules to avoid collisions on individual beads. After NHS coupling overnight, the coupling was quenched in 0.5M Tris pH 7.5 and beads were washed post coupling as previously described.

Because the crosslinked complexes are immobilized on NHS magnetic beads, we can perform several enzymatic steps by adding buffers and enzymes directly to the beads and performing rapid buffer exchange between each step on a magnet. All enzymatic steps were performed with shaking at 1200-1600 rpm (Eppendorf Thermomixer) to avoid bead settling and aggregation. All enzymatic steps were inactivated either by adding 1 mL of SPRITE Wash buffer (20mM Tris-HCl pH 7.5, 50mM NaCl, 0.2% Triton-X, 0.2% NP-40, 0.2% Sodium deoxycholate) supplemented with 50 mM EDTA and 50 mM EGTA to the NHS beads or Modified RLT buffer (1x Buffer RLT supplied by Qiagen, 10mM Tris-HCl pH 7.5, 1mM EDTA, 1mM EGTA, 0.2% N-Lauroylsarcosine, 0.1% Triton-X, 0.1% NP-40).

DNA End Repair and dA-tailing. We then repair the DNA ends to enable ligation of tags to each molecule. Specifically, we blunt end and phosphorylate the 5′ ends of double-

stranded DNA using two enzymes. First, the NEBNext End Repair Enzyme cocktail (E6050L; containing T4 DNA Polymerase and T4 PNK) and 1x NEBNext End Repair Reaction Buffer is added to beads and incubated at 20°C for 1 hour, and inactivated and buffer exchanged as specified above. DNA was then dA-tailed using the Klenow fragment (5′-3′ exo-, NEBNext dA-tailing Module; E6053L) at 37°C for 1 hour, and inactivated and buffer exchanged as specified above. Note, we do not use the NEBNext Ultra End Repair/dA-tailing module as the temperatures in the protocol are not compatible with SPRITE as the higher temperature will reverse crosslinks. To prevent degradation of RNA, each enzymatic step is performed with the addition of 1:40 NEB Murine RNAse Inhibitor or Thermofisher Ribolock.

Ligation of the DNA Phosphate Modified ("DPM") Tag. After end repair and dA-tailing of DNA, we performed a pooled ligation with "DNA Phosphate Modified" (DPM) tag that contains certain modifications that we found to be critical for the success of RD-SPRITE. Specifically, (i) we incorporate a phosphothiorate modification into the DPM adaptor to prevent its enzymatic digestion by Exo1 in subsequent RNA steps and (ii) we integrated an internal biotin modification to facilitate an on-bead library preparation post reverse-crosslinking. The DPM adaptor also contains a 5'phosphorylated sticky end overhang to ligate tags during split-pool barcoding. Ligation was performed as previously described using Instant Sticky End Mastermix (NEB) except that all ligations were supplemented with 1:40 RNAse inhibitor (ThermoFisher Ribolock or NEB Murine RNase Inhibitor) to prevent RNA degradation. Because T4 DNA Ligase only ligates to double-stranded DNA, the unique DPM sequence enables accurate identification of DNA molecules after sequencing.

Ligation of the RNA Phosphate Modified ("RPM") Tag. To map RNA and DNA interactions simultaneously, we ligated an RNA adaptor to RNA that contains the same 7nt 5'phosphorylated sticky end overhang as the DPM adaptor to ligate tags to both RNA and DNA during split-pool barcoding. To do this, we first modify the 3'end of RNA to ensure that they all have a 3'OH that is compatible for ligation. Specifically, RNA overhangs are repaired with T4 Polynucleoide Kinase (NEB) with no ATP at 37°C for 20 min. RNA is

subsequently ligated with a "RNA Phosphate Modified" (RPM) adaptor as previously described using High Concentration T4 RNA Ligase I[108]. Because T4 RNA Ligase 1 only ligates to single-stranded RNA, the unique RPM sequence enables accurate identification of RNA and DNA molecules after sequencing. After RPM ligation, RNA was converted to cDNA using Superscript III at 42°C for 1 hour using the "RPM bottom" RT primer that contains an internal biotin to facilitate on-bead library construction (as above) and a 5'end sticky end to ligate tags during SPRITE. Excess primer is digested with Exonuclease 1 at 42°C for 10-15 min. All ligations were supplemented with 1:40 RNAse inhibitor (ThermoFisher Ribolock or NEB Murine RNase Inhibitor) to prevent RNA degradation.

Split-and-pool barcoding to identify RNA and DNA interactions. The beads were then repeatedly split-and-pool ligated over four rounds with a set of "Odd," "Even", and "Terminal" tags (see SPRITE Tag Design in Quinodoz et al. Cell 2018[55]). Both DPM and RPM contain the same 7 nucleotide sticky end that will ligate to all subsequent split-pool barcoding rounds. All split-pool ligation steps and reverse crosslinking were performed for 45min to 1 hour at 20°C as previously described. All ligations were supplemented with 1:40 RNAse inhibitor (ThermoFisher Ribolock or NEB Murine RNase Inhibitor) to prevent RNA degradation.

Reverse crosslinking. After multiple rounds of SPRITE split-and-pool barcoding, the tagged RNA and DNA molecules are eluted from NHS beads by reverse crosslinking overnight (~12-13 hours) at 50°C in NLS Elution Buffer (20mM Tris-HCl pH 7.5, 10mM EDTA, 2% N-Lauroylsarcosine, 50mM NaCl) with added 5M NaCl to 288 mM NaCl Final combined with 5 μL Proteinase K (NEB).

Post reverse-crosslinking library preparation. AEBSF (Gold Biotechnology CAS#30827-99-7) is added to the Proteinase K (NEB Proteinase K #P8107S; ProK) reactions to inactive the ProK prior to coupling to streptavidin beads. Biotinylated barcoded RNA and DNA are bound to Dynabeads™ MyOne™ Streptavidin C1 beads (ThermoFisher #65001). To improve recovery, the supernatant is bound again to 20μL of streptavidin beads and combined with the first capture. Beads are washed in 1X PBS + RNase inhibitor and then

resuspended in 1x First Strand buffer to prevent any melting of the RNA:cDNA hybrid. Beads were pre-incubated at 40C for 2 min to prevent any sticky barcodes from annealing and extending prior to adding the RT enzyme. A second reverse transcription is performed by adding Superscript III (Invitrogen #18080051) (without RT primer) to extend the cDNA through the areas which were previously crosslinked. The second RT ensures that cDNA recovery is maximal, particularly if RT terminated at a crosslinked site prior to reverse crosslinking. After generating cDNA, the RNA is degraded by addition of RNaseH (NEB # M0297) and RNase cocktail (Invitrogen #AM2288), and the 3'end of the resulting cDNA is ligated to attach an dsDNA oligo containing library amplification sequences for subsequent amplification.

Previously, we performed cDNA (ssDNA) to ssDNA primer ligation which relies on the two single stranded sequences coming together for conversion to a product that can then be amplified for library preparation. To improve the efficiency of cDNA molecules ligated with the Read1 Illumina priming sequence, we perform a "splint" ligation, which involves a chimeric ssDNA-dsDNA adaptor that contains a random 6mer that anneals to the 3' end of the cDNA and brings the 5' phosphorylated end of the cDNA adapter directly together with the cDNA via annealing. This ligation is performed with 1x Instant Sticky End Master Mix (NEB #M0370) at 20°C for 1 hour. This greatly improves the cDNA tagging and overall RNA yield.

Libraries were amplified using 2x Q5 Hot-Start Mastermix (NEB #M0494) with primers that add the indexed full Illumina adaptor sequences. After amplification, the libraries are cleaned up using 0.8X SPRI (AMPure XP) and then gel cut using the Zymo Gel Extraction Kit selecting for sizes between 280 bp - 1.3 kb. A calculator for estimating the number of reads required to reach a saturated signal depth for each library are provided in **Supplemental Table 4** as well as in Quinodoz et al. 2021 *Nature Protocols* (in press) describing the SPRITE method.

Sequencing. Sequencing was performed on an Illumina NovaSeq S4 paired-end 150x150 cycle run. For the mES RNA-DNA RD-SPRITE data in this experiment, 144 different

SPRITE libraries were generated from four technical replicate SPRITE experiments and were sequenced. The four experiments were generated using the same batch of crosslinked lysate processed on different days to NHS beads. Each SPRITE library corresponds to a distinct aliquot during the Proteinase K reverse crosslinking step which is separately amplified with a different barcoded primer, providing an additional round of SPRITE barcoding.

Primers Used for RPM, DPM, and Splint Ligation (IDT):

1. RPM top: /5Phos/rArUrCrArGrCrACTTAGCG TCAG/3SpC3/
2. RPM bottom (internal biotin): /5Phos/TGACTTGC/iBiodT/GACGCTAAGTGCTGAT
3. DPM Phosphorothioate top: /5Phos/AAGACCACCAGATCGGAAGAGCGTCGTG*T* A*G*G* /32MOErG/ *Denotes Phosphorothioate bonds
4. DPM bottom (internal biotin): /5Phos/TGACTTGTCATGTCT/iBioT/CCGATCTGGTGGTCTTT
5. 2Puni splint top: TACACGACGCTCTTCCGATCT NNNNNN/3SpC3/
6. 2Puni splint bottom: /5Phos/AGA TCG GAA GAG CGT CGT GTA/3SpC3/

Annealing of adaptors. A double-stranded DPM oligo and 2P universal "splint" oligo were generated by annealing the complementary top and bottom strands at equimolar concentrations. Specifically, all dsDNA SPRITE oligos were annealed in 1x Annealing Buffer (0.2 M $LiCl_2$, 10 mM Tris-HCl pH 7.5) by heating to 95°C and then slowly cooling to room temperature (-1°C every 10 sec) using a thermocycler.

Assessing molecule to bead ratio. We ensured that SPRITE clusters represent *bona fide* interactions that occur within a cell by mixing human and mouse cells and ensuring that virtually all SPRITE clusters (~99%) represent molecules exclusively from a single species. Specifically, we separately crosslinked HEK293T cells performed a human-mouse mixing RD-SPRITE experiment and identified conditions with low interspecies mixing

(molecules = RNA+DNA instead of DNA). Specifically, for SPRITE clusters containing 2-1000 reads, the percent of interspecies contacts is: 2 beads:molecule = 0.9% interspecies contacts, 4 beads:molecule = 1.1% interspecies contacts, 8 beads:molecule = 1.1% interspecies contacts. We used the 2 beads:molecule and 4 beads:molecule ratio for the RD-SPRITE data sets generated in this paper.

RD-SPRITE technical replicates. One of the RD-SPRITE replicate libraries was generated with a DPM lacking the phosphorothioate bond and 2'-O-methoxy-ethyl bases on the 3'end of the top adaptor. We found that this resulted in a lower number of DNA reads because the exonuclease step can degrade the single-stranded portion of the DPM oligo. As a result, this library has lower DNA-DNA and DNA-RNA pairs, but has more RNA-RNA contacts overall. This experiment was analyzed to generate higher-resolution RNA-RNA contact matrices, including contacts of lower abundance RNAs. The three other RD-SPRITE replicate libraries were generated with the same batch crosslinked lysate but were ligated with a DPM adaptor containing these modifications to prevent DNA degradation.

RD-SPRITE processing pipeline

Adapter trimming. Adapters were trimmed from raw paired-end fastq files using Trim Galore! v0.6.2 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and assessed with Fastqc v0.11.9. Subsequently, the DPM (GATCGGAAGAG) and RPM (ATCAGCACTTA) sequences are trimmed using Cutadapt v2.5[109] from the 5' end of Read 1 along with the 3' end DPM sequences that result from short reads being read through into the barcode (GGTGGTCTTT, GCCTCTTGTT, CCAGGTATTT, TAAGAGAGTT, TTCTCCTCTT, ACCCTCGATT). The additional trimming helps improve read mapping in the end-to-end alignment mode. The SPRITE barcodes of trimmed reads are identified with Barcode ID v1.2.0 (https://github.com/GuttmanLab/sprite2.0-pipeline) and the ligation efficiency is assessed. Reads with an RPM or a DPM barcode are split into two separate files, to process RNA and DNA reads individually downstream, respectively.

Ligation Efficiency Quality Control. We assessed the reproducibility and quality of an RD-SPRITE experiment by calculating the ligation efficiency, defined as the proportion of sequencing reads containing only 1, 2, 3… through n barcodes (where n is the number of rounds of split-pool barcoding). Across technical replicates, biological replicates, and multiple sequencing libraries, we have found highly similar ligation efficiencies, with ~60% or more of reads containing all 5 barcoding tags (see **Supplemental Table 3**).

Processing RNA reads. RNA reads were aligned to GRCm38.p6 with the Ensembl GRCm38 v95 gene model annotation using Hisat2 v2.1.0[110] with a high penalty for soft-clipping --sp 1000,1000. Unmapped and reads with a low MapQ score (samtools view -bq 20) were filtered out for downstream realignment. (see **Supplemental Table 2** for alignment statistics). Mapped reads were annotated for gene exons and introns with the featureCounts tool from the subread package v1.6.4 using Ensembl GRCm38 v95 gene model annotation and the Repeat and Transposable element annotation from the Hammel lab[111]. Filtered reads were subsequently realigned to our custom collection of repeat sequences using Bowtie v2.3.5[112], only keeping mapped and primary alignment reads.

Processing DNA reads. DNA reads were aligned to GRCm38.p6 using Bowtie2 v2.3.5 (see **Supplemental Table 2** for alignment statistics), filtering out unmapped and reads with a low MapQ score (samtools view -bq 20). Data generated in F1 hybrid cells (pSM44: 129 × castaneous) were assigned the allele of origin using SNPsplit v0.3.4[113]. RepeatMasker[114] defined regions with milliDev ≤ 140 along with blacklisted v2 regions were filtered out using Bedtools v2.29.0[115].

SPRITE cluster file generation. RNA and DNA reads were merged, and a cluster file was generated for all downstream analysis. MultiQC v1.6[116] was used to aggregate all reports.

Masked bins. In addition to known repeat containing bins, we manually masked the following bins (mm10 genomic regions: chr2:79490000-79500000, chr11:3119270-3192250, chr15:99734977-99736026, chr3:5173978-5175025, chr13:58176952-58178051) because we observed a major overrepresentation of reads in the input samples.

Microscopy imaging

Immunofluorescence (IF). Cells were grown on coverslips and rinsed with 1x PBS, fixed in 4% paraformaldehyde in PBS for 15 minutes at room temperature, rinsed in 1x PBS, and permeabilized with 0.5% Triton X-100 in PBS for 10 minutes at room temperature. Cells were either stored at -20°C in 70% ethanol or used directly for immunostaining and incubated in blocking solution (0.2% BSA in PBS) for at least 1 hour. If stored in 70% ethanol, cells were re-hydrated prior to staining by washing 3 times in 1xPBS and incubated in blocking solution (0.2% BSA in PBS) for at least 1 hour. Primary antibodies were diluted in blocking solution and added to coverslips for 3-5 hours at room temperature incubation. Cells were washed three times with 0.01% Triton X-100 in PBS for 5 minutes each and then incubated in blocking solution containing corresponding secondary antibodies labeled with Alexa fluorophores (Invitrogen) for 1 hour at room temperature. Next, cells were washed 3 times in 1xPBS for 5 minutes at room temperature and mounting was done in ProLong Gold with DAPI (Invitrogen, P36935). Images were collected on a LSM800 or LSM980 confocal microscope (Zeiss) with a 63× oil objective. *Z* sections were taken every 0.3 μm. Image visualization and analysis was performed with Icy software (http://icy.bioimageanalysis.org/) and ImageJ software (https://imagej.nih.gov/).

Immunofluorescence (IF) for ActD experiments. Cells were cultured in DMSO or ActD (Sigma A9415, 25μL of 1mg/mL stock added per 1ml culture medium) for 4 hours, then fixed and processed for IF using the anti-NPAT antibody, as described earlier. Images were acquired using the Zeiss LSM980 microscope with 63x oil objective and 16 Z-sections were taken with 0.3 μm increments. To count the number of NPAT spots, we generated the maximal projections, defined a binary mask by thresholding based on background intensity levels, and manually counted the number of spots for each nucleus.

RNA Fluorescence in situ Hybridization (RNA-FISH). RNA-FISH performed in this study was based on the ViewRNA ISH (Thermo Fisher Scientific, QVC0001) protocol

with minor modifications. Cells grown on coverslips were rinsed in 1xPBS, fixed in 4% paraformaldehyde in 1xPBS for 15 minutes at room temperature, permeabilized in 0.5% Triton-100 in the fixative for 10 minutes at room temperature, rinsed 3 times with 1xPBS, and stored at -20°C in 70% ethanol until hybridization steps. All the following steps were performed according to manufacturer's recommendations. Coverslips were mounted with ProLong Gold with DAPI (Invitrogen, P36935) and stored at 4°C until acquisition. For nuclear and nucleolar RNAs, cells were pre-extracted with 0.5% ice cold Triton-100 for 3 minutes to remove cytoplasmic background and fixed as described. All probes used in the study were custom made by Thermofisher (order numbers available upon request). To test their specificity, we either utilized RNAse treatment prior to RNA-FISH or two different probes targeting the same RNA. Images were acquired on Zeiss LSM800 or LSM980 confocal microscope with a 100x glycerol immersion objective lens and Z-sections were taken every 0.3 μm. Image visualization and analysis was performed with Icy software and ImageJ software.

RNA FISH for scaRNA and tRNAs were performed with a combined set of probes to increase the signal of lower abundance RNAs. Specifically, scaRNAs were visualized with two combined probes of scaRNA2 and scaRNA17. tRNAs were visualized using probes targeting tRNA-Arg-TCG-4-1, tRNA-Leu-AAG-3-1, tRNA-Ile-AAT-1-8, tRNA-Arg-TCT-5-1, tRNA-Leu-CAA-2-1, tRNA-Ile-TAT-2-1, tRNA-Tyr-GTA-1-1. tRNA sequences were obtained using the GtRNAdb GRCm38/mm10 predictions (Lowe Lab, UCSC)[117,118].

Combined RNA-FISH and IF. For immunostaining combined with *in situ* RNA visualization, we used the ViewRNA Cell Plus (Thermo Fisher Scientific, 88-19000-99) kit per the manufacturer's protocol with minor modifications. Immunostaining was performed as described above, but all incubations were performed in blocking buffer with addition of RNAse inhibitor and all the wash steps were performed in RNAse free 1x PBS with RNAse inhibitor. Blocking buffer, PBS, RNAse inhibitors are provided in a kit. After the last wash in 1x PBS, cells underwent post-fixation in 2% paraformaldehyde on 1x PBS for 10min at room temperature, were washed 3 times in 1x PBS, and then RNA-FISH

protocol was followed as described above. Images were acquired on the Zeiss LSM800 or LSM980 confocal microscope with a 100x glycerol immersion objective lens and z-sections were taken every 0.3 μm. Image visualization and analysis was performed with Icy software and ImageJ software.

DNA-FISH. DNA-FISH was performed as previously described[119] with modifications. Cells grown on coverslips were rinsed with 1x PBS, fixed in 4% paraformaldehyde in 1x PBS for 15 minutes at room temperature, permeabilized in 0.5% Triton-100 in the fixative for 10 minutes at room temperature, rinsed 3 times with 1x PBS and stored at -20˚C in 70% ethanol until hybridization steps. Pre-hybridization cells were dehydrated in 100% ethanol and dried for 5 minutes at room temperature. 4 μL drop of hybridization mix with probes was spotted on a glass slide and dried coverslips were placed on the drop. Coverslips were sealed with rubber cement, slides were incubated for 5 minutes at 85˚C, and then incubated overnight at 37˚C in humid atmosphere. After hybridization and three washes with 2x SSC, 0.05% Triton-100 and 1mg/mL PVP in PBS at 50˚C for 10 minutes, cells were rinsed in 1x PBS and mounted with ProLong Gold with DAPI (Invitrogen, P36935).

Hybridization buffer consisted of 50% formamide, 10% dextran sulphate, 2xSSC, 1 mg/mL polyvinyl pyrrolidone (PVP), 0.05% Triton X-100, 0.5 mg/mL BSA. 1 mM short oligonucleotides labeled with Cy5 ([CY5]ttttctcgccatattccaggtc) were used as probes against Major Satellites and full-length minor satellite repeat sequence was used as probes against Minor Satellites. Minor satellite sequence was firstly cloned to pGEM plasmid and then labeled by PCR reaction with self-made TAMRA dATPs for minor satellites. Labeled PCR product was purified with a QIAquick PCR Purification Kit (QIAGEN), and 50 ng was mixed with hybridization buffer. Images were acquired on Zeiss LSM800 or LSM980 confocal microscope with a 63x glycerol immersion objective lens and Z-sections were taken every 0.3 μm. Image visualization and analysis was performed with Icy software and ImageJ software.

Analysis of RNA-DNA contacts

Generating contact profiles. To map the genome-wide localization profile of a specific RNA, we calculated the contact frequency between the RNA transcript and each region of the genome binned at various resolutions (1Mb, 100kb and 10kb). Raw contact frequencies were computed by counting the number of SPRITE clusters in which an RNA transcript and a genomic bin co-occur. We normalized these raw contacts by weighting each contact by a scaling factor based on the size of its corresponding SPRITE cluster. Specifically, we enumerate all pairwise contacts within a SPRITE cluster and weight each contact by 2/n, where n is the total number of reads within a cluster.

RNA and cluster sizes. RNA-DNA contacts were computed for a range of SPRITE cluster sizes, such as 2-10, 11-100, and 101-1000, ≥1001 reads. We found that different RNAs tend to be most represented in different clusters sizes – likely reflecting the size of the nuclear compartment that they occupy. For example, 45S and snoRNAs are most represented in large clusters, while Malat1, snRNAs, and other ncRNAs tend to be represented in smaller SPRITE clusters. For analyses in this paper, we utilized clusters containing 2-1000 reads unless otherwise noted.

Visualizing contact profiles. These methods produce a one-dimensional vector of DNA contact frequencies for each RNA transcript that we output in bedgraph format and visualize with IGV[120]. To compare DNA contact profiles between RNA transcripts, we calculated a Pearson correlation coefficient between the one-dimensional DNA contact vectors for all pairs of RNA transcripts.

Aggregate analysis of RNA-DNA contacts. To map RNA-DNA localization across chromosomes with respect to centromeres and telomeres (e.g. Terc and satellite ncRNAs), we computed an average localization profile as a function of distance from the centromere of each chromosomes. To do this, we converted each 1Mb genomic bin into a percentile bin from 0 to 100 based on its relative position on its chromosome (from 5’ to 3’ ends).

We then calculated the average contact frequency for a given RNA with each percentile bin across all chromosomes.

Allele specific analysis. To map localization to different alleles, we identified all clusters containing a given RNA (as above) and quantified the number of DNA reads uniquely mapping to each allele using allele specific alignments. Allele specific RNA-DNA contact frequencies were normalized by overall genomic read coverage for each allele to account for differences in coverage for each allele.

Nucleolar hub RNA-DNA contacts. We observe enrichment of pre-rRNAs and other nucleolar hub RNAs on chromosomes containing 45S ribosomal DNA (rDNA). Specifically, rDNA genes are contained on the centromere-proximal regions of chromosomes 12, 15, 16, 18, and 19 in mouse ES cells. We previously showed that regions on these chromosomes organize around nucleoli in the majority of cells imaged with DNA FISH combined with immunofluorescence for Nucleolin[55]. We also observed nucleolar hub RNAs enriched on other genomic regions corresponding to centromere-proximal DNA and transcriptionally inactive, gene poor regions. We previously showed that these genomic regions are organized proximal to the nucleolus using SPRITE and microscopy[55].

Splicing RNA concentration relative to nuclear speckle distance. We observed that snRNAs are enriched over genomic regions with high gene-density, which we have previously shown organize around the nuclear speckle[55]. To explore whether splicing RNA concentration is related to genomic DNA distance to nuclear speckles, we computed the RNA-DNA contact profile for U1 snRNA in 10 kb bins across the genome, weighted by cluster size. For the same 10 kb bins, we calculated the RNA expression levels (the number of clusters containing the pre-mRNA) and filtered for bins with RNA counts > 100. In our dataset, this filter selects for genomic regions with high gene expression levels regardless of speckle distance. We then generated a "distance to speckle" metric for each genomic bin using DNA-DNA SPRITE measurements. This "distance" is defined as the average inter-chromosomal contact frequency between a given bin and genomic bins corresponding to the "active" hub (i.e. "speckle" hub). A larger contact frequency value is considered "close

to the speckle" while a smaller value is "far from the speckle". We grouped the 10 kb bins into 5 groups based on the "distance to speckle" metric and focused our subsequent analysis on the "closest" and "farthest" groups. Closest regions contained a normalized speckle distance score between 0.4-0.5 and farthest contained a score from 0-0.1. We then compared the distribution of U1 density over genes close to or far from the nuclear speckle.

Analysis of RNA-RNA contacts

RNA-RNA contact matrices. We computed the contact frequency between each RNA-RNA pair by counting the number of SPRITE clusters containing two different RNAs. To account for coverage differences in individual RNAs, we normalized this matrix using a matrix balancing normalization approach as previously described[121]. Briefly, this approach works by ensuring the rows and columns of a symmetric matrix add up to 1. In this way, RNA abundance does not dominate the overall strength of the contact matrix. For multi-copy RNAs (e.g. repeat-encoded RNAs, ribosomal RNA, tRNAs), all reads mapping to a given RNA were collapsed. Specifically, multi-copy RNA reads mapping to either the mm10 genome annotated using repeat masker or a custom repeat genome consensus were collapsed.

RNA Hubs. Groups of pairwise interacting RNAs were first identified using hierarchical clustering of the pairwise RNA-RNA contact matrix. Groups were defined as sets of pairwise interacting RNAs that showed high pairwise contact frequencies with other RNAs within the same group, but low contact frequency with RNAs in other groups. We next explored the multiway contacts of the RNAs within these groups using our multi-way contact score (details below). The term "hub" is used to refer to these higher-order, multi-way interacting group of RNAs.

Multi-way Contact Score (k-mer analysis). To assess the significance of multiple RNAs co-occurring within the same SPRITE cluster, we computed a multi-way contact score. Specifically, we compared the observed number of SPRITE clusters containing a specific

multi-way contact to the "expected" number of SPRITE clusters containing the multi-way contacts if the components were randomly distributed. To account for the fact that higher-order structures (i.e. k-mers) might be more frequent than expected at random because only a subset of the RNAs, but not all components, specifically interact, we calculated the "expected" count for a given k-mer from permutations where we fixed the frequency and structure of each (k-1)-mer subsets and permuted the remaining RNAs in a cluster based on its observed RNA frequency in the dataset. We then computed the frequency that we observe the full k-mer structure at random. More concretely, consider the 3-way simultaneous contact between RNAs A, B, and C (A-B-C). First, we generate the permuted dataset to estimate the frequency of this interaction occurring randomly. We focus on only clusters in the RD-SPRITE dataset containing a sub-fragment of the interaction (clusters with A-B) and reassign the other members of the cluster using the fractional abundances of RNAs within the complete RD-SPRITE dataset. We then count the number of occurrences of A-B-C within the permuted dataset. We repeated these permutations 100 times to generate an "expected" distribution and used this distribution to compute a p-value (how frequently do we randomly generate a value greater than or equal to the observed frequency) and z-score (the observed frequency minus average frequency of permuted values divided by the permuted distribution standard deviation). For a given multi-way k-mer, we report the maximum statistics of all possible paths to assembling the k-mer (e.g. max(A-B|C, B-C|A, A-C|B)). In this way, if only the interaction of a k-mer subset, for instance B-C, occurs more frequently than by random chance, but the addition of A to the B-C k-mer does not occur more frequently than by random chance, the full multi-way interaction would not be significant.

Mapping intron versus exon RNA-RNA contacts. To explore the differential RNA contacts that occur within nascent pre-mRNA and mature mRNAs, we focused on the intronic regions and exonic regions of mRNAs respectively. We retained all intronic or exonic regions that were contained in at least 100 independent SPRITE clusters. We then generate contact matrices between splicing non-coding RNAs (U1, U2, U4, U5, U6) and translation non-coding RNAs (18S, 28S, 5S, 5.8S) and these mRNA exons, and introns.

We performed a matrix balancing normalization (ICE normalization[121]) on this symmetric contact matrix and plotted splicing RNAs and translation RNAs (columns) versus mRNA exons and introns (rows).

Identifying unannotated scaRNAs. We calculated the weighted contact frequency of how often a given RNA contacts scaRNA2. Many of the top hits correspond to Mus musculus (mm10) annotated scaRNAs (e.g. scaRNA9, scaRNA10, scaRNA6, scaRNA7, scaRNA1, scaRNA17, and scaRNA13). Other hits include regions within mRNA introns. We performed BLAST-like Alignment Tool (BLAT, https://genome.ucsc.edu/cgi-bin/hgBlat) on other top hits contacting scaRNA2, including the Trrap intron region and Gon4l1 intron region and found they are homologous to human scaRNA28 and scaRNA26A, respectively. Specifically, the Trrap region in mm10 homologous to scaRNA28 is chr5:144771339-144771531 and the Gon4l region in mm10 homologous to scaRNA26A is chr3:88880319-88880467.

Analysis of multi-way RNA and DNA SPRITE contacts

Generating RNA-DNA-DNA Contact Matrices for SPRITE clusters containing an individual or multiple RNAs. To analyze higher-order RNA and DNA contacts in the SPRITE clusters, we generated DNA-DNA contact frequency maps in the presence of specific sets of RNA transcripts. To generate these DNA-DNA contact maps, we first obtained the subset of SPRITE clusters that contained an RNA transcript or multiple transcripts of interest (e.g., nucleolar RNAs, spliceosomal RNAs, scaRNAs satellite RNAs, lncRNA). We then calculated DNA-DNA contact maps for each subset of SPRITE clusters at 100kb and 1Mb resolution by determining the number of clusters in which each pair of genomic bins co-occur. Raw contacts were normalized by SPRITE cluster size by dividing each contact by the total number of reads in the corresponding SPRITE cluster. Specifically, we enumerate all pairwise contacts within a SPRITE cluster and weight each contact by 2/n, where n is the total number of reads within a cluster. This resulted in

genome-wide DNA-DNA contact frequency maps for each set of RNA transcripts of interest.

Aggregate DNA-DNA inter-chromosomal maps for SPRITE clusters containing an individual or multiple RNAs. For satellite-derived ncRNAs, we also calculated a mean inter-chromosomal DNA-DNA contact frequency map. To do this, we converted each 1Mb genomic bin into a percentile bin from 0 to 100 based on its chromosomal position, where the 5' end is 0 and the 3' end is 100. We then calculated the DNA contact frequency between all pairs of percentile bins for all pairs of chromosomes. We used these values to calculate a mean inter-chromosomal contact frequency map, which reflects the average contact frequency between each pair of percentile bins between all pairs of chromosomes.

Actinomycin D RNA-DNA SPRITE and DNA SPRITE

DNA SPRITE. DNA SPRITE was performed on three biological replicates of ActD-treated or control DMSO treated cells following the protocol described in our previous work (Quinodoz, et al. *Nature Protocols* - In Press). The individual samples were processed in parallel during crosslinking, cell lysis, sonication, and chromatin fragmentation. DNase treatment conditions were independently optimized for cell lysates of ActD or DMSO-treated samples. Samples were then separately coupled to NHS-beads and the DNA fragments end-repaired and phosphorylated. For DPM adaptor ligation, a unique set of DPM adaptors was used for each treatment condition and replicate, allowing us to distinguish the subsequently sequenced DNA reads corresponding to each sample based on the identity of the DPM adaptor. Following DPM ligation, the six samples (three biological replicates of ActD and three biological replicates of DMSO) were pooled and taken through four rounds of split-pool barcoding.

RNA & DNA SPRITE. RD-SPRITE was performed on ActD or DMSO treated cells following the protocol detailed above. Similar to the DNA-SPRITE experiment, the individual replicates were processed in parallel for the first steps of the protocol and pooled

after the first round of split-pool barcoding. In DNA-SPRITE, there are 96 possible DPM adaptors and we could therefore use the identity of the DPM adaptor to distinguish reads from the individual samples. In RD-SPRITE, there is a single DPM adaptor and we instead use the first round of split-pool barcoding to distinguish individual samples. Therefore, the samples were only pooled after the first round of barcoding and each sample ligated with a unique subset of ODD adaptors for the first round.

Sequencing. Sequencing was performed on an Illumina NovaSeq S4 paired-end 150x150 cycle run. For the DNA-SPRITE data, 16 different SPRITE libraries were generated and sequenced. For the RD-SPRITE data, 16 different SPRITE libraries were generated and sequenced. In both cases, the individual libraries contained data from all three biological replicates of ActD-treated and all three biological replicates of DMSO-control treated samples.

DNA SPRITE processing pipeline. DNA-SPRITE data for ActD-treated and control DMSO-treated samples was processed using a pipeline we have previously described (Quinodoz, et al. *Nature Protocols* - In Press). To distinguish clusters corresponding to each sample, the identity of the DPM tag was used.

RNA-DNA SPRITE processing pipeline. RNA-DNA SPRITE data for ActD-treated and control DMSO-treated samples was processed as previously described, with minor modifications. For instance, updated versions of gene annotations (Gencode release M25 annotations for GRCm38.p6) and our custom collection of repeat RNA sequences were used to annotate RNA reads. To distinguish clusters corresponding to each sample, the identity of the first ODD barcode was used.

Sample replicates. Biological replicates of ActD-treated and control DMSO-treated samples were prepared in triplicate for both DNA-SPRITE and RNA-DNA SPRITE experiments. As described, the individual replicates were processed in parallel for the initial steps of the protocols and merged for the split-pool barcoding and sequencing steps of the protocols. Following cluster generation, the three replicates for each treatment

condition were merged into a single cluster file. All subsequent contact analysis was performed on the aggregated datasets. Various metrics, such as ligation efficiency, alignment rates, RNA expression, and cluster sizes, were comparable across the biological replicates.

Sample and cluster sizes. The cluster size distribution was computed for each sample and each replicate independently. In both RD-SPRITE and DNA-SPRITE, the cluster size distribution for different technical replicates of a single treatment condition was nearly identical. Between the ActD and DMSO conditions, we found that the ActD and DMSO overall cluster sizes (all clusters) were comparable. However, specifically within the clusters containing DNA reads, ActD treated samples and control DMSO treated samples had different cluster size distribution profiles, with ActD samples favoring larger DNA cluster sizes.

When comparing DNA-DNA contacts or RNA-DNA contacts for specific hub RNAs, we focused on the cluster size ranges we found reflected certain nuclear compartments in the untreated samples. Specifically, the nucleolar hub is best seen in larger cluster sizes (2-10,000 reads/cluster for DNA-SPRITE while the scaRNA hub or HLB hub is seen in smaller cluster sizes (2-1000 reads/cluster). In addition, we found that snoRNAs shifted from their typical localization in larger SPRITE clusters in control-DMSO samples[55], to smaller clusters in ActD treated samples, likely due to a loss of localization to the nucleolus. For analysis involving snoRNA-DNA contacts for DMSO and ActD treatment, we focused on larger cluster sizes (1001-10K).

Quantification of RNA abundance. RNA abundance was calculated by counting the number of annotated RNA reads within all SPRITE clusters of size 2-1000. To account for differences in read coverage between samples, we normalized expression to the number of counted reads for 28S rRNA. For classes of RNA corresponding to different hubs (snoRNAs, scaRNAs, tRNAs), we summed the total number of reads annotated with genes in this class. For intron reads, we only considered protein-coding transcripts and, for 45S rRNA, we considered reads mapped to ITS1, ITS2 or the 3’ end. Finally, to visualize the

changes for RNAs with vastly different expression levels, we set the normalized expression value of DMSO samples to one and rescaled the ACTD values accordingly.

DNA-DNA contact matrices. Cluster size weighted DNA-DNA contact matrices were generated at various resolutions (1Mb, 100kb, 50kb, etc.) from DNA-SPRITE data as previously described. In brief, raw contact frequencies were calculated by counting the number of clusters containing reads from both genomic bins. We weighted each contact by a scaling factor related to the cluster size, specifically, n/2 where n is the number of reads in each cluster. The weighted contact matrices were normalized using iterative correction and eigenvector decomposition (ICE), a matrix balancing normalization approach, as previously described[121].

To compare nucleolar-hub DNA-DNA contact profiles, we scaled the DNA-DNA matrices to the mean intra-chromosomal contact frequency. Specifically, to compute this re-scaling factor, we defined 20-bin windows for each chromosome and then calculated the average pairwise contacts within these 20-bin windows, excluding self-contacts, across the genome. This way, we can visualize changes in the inter-chromosomal vs intra-chromosomal contact frequency. We defined the genomic regions corresponding to the nucleolar hub based on previous SPRITE data[55].

Because the two samples contained slightly different read depths and cluster sizes, we wanted to ensure that observed differences could not simply be explained by these differences. Therefore, to compare DNA-DNA contact profiles at histone gene clusters or snRNA gene clusters between the ActD and DMSO treatment conditions and account for different read depths, we rank-order rescaled the DNA-DNA matrices. This normalization allows us to determine if the overall structure of the two matrices are similar, even if the exact order of magnitude of individual interactions might differ. To do this, we first computed the pairwise contact frequencies in both samples. Then we rank ordered the contact frequencies in a specific region for DMSO and ActD samples independently and computed the average rank ordered contact frequency. Finally, we remapped the matrix values for each sample to the average value based on rank position. After rescaling, the

DNA-DNA contact matrices for each sample share the same distribution and can be visually compared. We note that we observe comparable differences at the reported structures regardless of the precise method of normalization.

RNA-RNA contact matrices. We computed contact frequencies between pairs of RNAs by counting the number of SPRITE clusters containing both RNAs. To account for differences in RNA abundance in each sample, we normalized the contact frequency of a given pair to the number of clusters containing either RNA. Specifically, we computed a normalized score by dividing the number of SPRITE clusters containing A and B by the number of clusters containing A or B.

RNA-DNA contact bedgraphs. To compare changes in RNA localization on chromatin following ActD treatment, we plotted weighted DNA-contact profile bedgraphs for various hub RNAs. Specifically, to generate a DNA-contact profile, we computed the number of clusters containing the RNA and a genomic bin. Identical to DNA-DNA contact profiles, the raw RNA-DNA contacts were weighted by a n/2 scaling factor corresponding to cluster size, where n corresponds to the number of reads in each cluster. We then normalized the weighted bedgraph by dividing each contact frequency by the read count of a given RNA. This normalization allows us to account for differences in abundance of a given RNA.

4.10. REFERENCES

1. Strom, A.R., and Brangwynne, C.P. (2019). The liquid nucleome - phase transitions in the nucleus at a glance. Journal of cell science. 10.1242/jcs.235093.

2. Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nature Reviews Molecular Cell Biology *16*, 245–257. 10.1038/nrm3965.

3. Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nature Reviews Genetics *17*, 772–772. 10.1038/nrg.2016.147.

4. Gibcus, J.H., and Dekker, J. (2013). The Hierarchy of the 3D Genome. Molecular Cell. 10.1016/j.molcel.2013.02.011.

5. Meshorer, E., and Misteli, T. (2006). Chromatin in pluripotent embryonic stem cells and differentiation. Nature Reviews Molecular Cell Biology *7*, 540–546. 10.1038/nrm1938.

6. Dundr, M., and Misteli, T. (2010). Biogenesis of nuclear bodies. Cold Spring Harbor perspectives in biology *2*. 10.1101/cshperspect.a000711.

7. Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nature Reviews Molecular Cell Biology *16*, 245–257. 10.1038/nrm3965.

8. 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., et al. (2017). The 4D nucleome project. Nature. 10.1038/nature23884.

9. Phillips-Cremins, J.E. (2014). Unraveling architecture of the pluripotent genome. Current Opinion in Cell Biology *28*, 96–104. 10.1016/j.ceb.2014.04.006.

10. Pederson, T. (2011). The nucleolus. Cold Spring Harbor Perspectives in Biology *3*, 1–15. 10.1101/cshperspect.a000638.

11. Spector, D.L., and Lamond, A.I. (2011). Nuclear speckles. Cold Spring Harbor Perspectives in Biology *3*, 1–12. 10.1101/cshperspect.a000646.

12. Guo, Y.E., Manteiga, J.C., Henninger, J.E., Sabari, B.R., Dall'Agnese, A., Hannett, N.M., Spille, J.H., Afeyan, L.K., Zamudio, A. V., Shrinivas, K., et al. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature. 10.1038/s41586-019-1464-0.

13. Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I.I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science. 10.1126/science.aar4199.

14. Nickerson, J.A., Krochmalnic, G., Wan, K.M., and Penman, S. (1989). Chromatin architecture and nuclear RNA. Proceedings of the National Academy of Sciences of the United States of America *86*, 177–181. 10.1073/pnas.86.1.177.

15. Melé, M., and Rinn, J.L. (2016). "Cat's Cradling" the 3D Genome by the Act of LncRNA Transcription. Molecular Cell *62*, 657–664. 10.1016/j.molcel.2016.05.011.

16. Rinn, J.L., and Guttman, M. (2014). RNA and dynamic nuclear organization. Science *345*, 1240–1241. doi: 10.1126/science.1252966.

17. Quinodoz, S., and Guttman, M. (2014). Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends in cell biology *24*, 651–663. 10.1016/j.tcb.2014.08.009.

18. Hall, L.L., and Lawrence, J.B. (2016). RNA as a fundamental component of interphase chromosomes: Could repeats prove key? Current Opinion in Genetics and Development. 10.1016/j.gde.2016.04.005.

19. Nozawa, R.S., and Gilbert, N. (2019). RNA: Nuclear Glue for Folding the Genome. Trends in Cell Biology. 10.1016/j.tcb.2018.12.003.

20. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 10.1038/nature07672.

21. Cabili, M., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes and Development. 10.1101/gad.17446611.

22. Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research. 10.1093/nar/gky955.

23. Cech, T.R., and Steitz, J.A. (2014). The noncoding RNA revolution - Trashing old rules to forge new ones. Cell. 10.1016/j.cell.2014.03.008.

24. Rinn, J.L., and Chang, H.Y. (2012). Genome Regulation by Long Noncoding RNAs. Annual Review of Biochemistry. 10.1146/annurev-biochem-051410-092902.

25. Black, D.L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. Annual Review of Biochemistry. 10.1146/annurev.biochem.72.121801.161720.

26. Nilsen, T.W., and Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. Nature. 10.1038/nature08909.

27. Watkins, N.J., and Bohnsack, M.T. (2012). The box C/D and H/ACA snoRNPs: Key players in the modification, processing and the dynamic folding of ribosomal RNA. Wiley Interdisciplinary Reviews: RNA. 10.1002/wrna.117.

28. Kiss-László, Z., Henry, Y., Bachellerie, J.P., Caizergues-Ferrer, M., and Kiss, T. (1996). Site-specific ribose methylation of preribosomal RNA: A novel function for small nucleolar RNAs. Cell. 10.1016/S0092-8674(00)81308-2.

29. Ni, J., Tien, A.L., and Fournier, M.J. (1997). Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. Cell. 10.1016/S0092-8674(00)80238-X.

30. Spycher, C., Streit, A., Stefanovic, B., Albrecht, D., Koning, T.H. wittop, and Schümperli, D. (1994). 3′ end processing of mouse histone pre-mRNA: Evidence for additional base-pairing between U7 snRNA and pre-mRNA. Nucleic Acids Research. 10.1093/nar/22.20.4023.

31. Mowry, K.L., and Steitz, J.A. (1987). Identification of the human U7 snRNP as one of several factors involved in the 3′ end maturation of histone premessenger RNA's. Science. 10.1126/science.2825355.

32. Salzler, H.R., Tatomer, D.C., Malek, P.Y., McDaniel, S.L., Orlando, A.N., Marzluff, W.F., and Duronio, R.J. (2013). A Sequence in the Drosophila H3-H4 Promoter Triggers Histone Locus Body Assembly and Biosynthesis of Replication-Coupled Histone mRNAs. Developmental Cell. 10.1016/j.devcel.2013.02.014.

33. Kolev, N.G., and Steitz, J.A. (2005). Symplekin and multiple other polyadenylation factors participate in 3′-end maturation of histone mRNAs. Genes and Development. 10.1101/gad.1371105.

34. Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-DiNardo, D., and Kanduri, C. (2008). Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation. Molecular Cell. 10.1016/j.molcel.2008.08.022.

35. Mancini-DiNardo, D., Steele, S.J.S., Levorse, J.M., Ingram, R.S., and Tilghman, S.M. (2006). Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. Genes and Development. 10.1101/gad.1416906.

36. Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A., and Panning, B. (2002). Xist RNA and the Mechanism of X Chromosome Inactivation . Annual Review of Genetics. 10.1146/annurev.genet.36.042902.092433.

37. Quaresma, A.J.C., Bugai, A., and Barboric, M. (2016). Cracking the control of RNA polymerase II elongation by 7SK snRNP and P-TEFb. Nucleic Acids Research. 10.1093/nar/gkw585.

38. Zhou, Q., Li, T., and Price, D.H. (2012). RNA Polymerase II Elongation Control. Annual Review of Biochemistry. 10.1146/annurev-biochem-052610-095910.

39. Egloff, S., Studniarek, C., and Kiss, T. (2018). 7SK small nuclear RNA, a multifunctional transcriptional regulatory RNA with gene-specific features. Transcription. 10.1080/21541264.2017.1344346.

40. Carmo-Fonseca, M., and Rino, J. (2011). RNA seeds nuclear bodies. Nature Cell Biology. 10.1038/ncb0211-110.

41. Shevtsov, S.P., and Dundr, M. (2011). Nucleation of nuclear bodies by RNA. Nature Cell Biology. 10.1038/ncb2157.

42. Andersen, J.S., Lam, Y.W., Leung, A.K.L., Ong, S.E., Lyon, C.E., Lamond, A.I., and Mann, M. (2005). Nucleolar proteome dynamics. Nature. 10.1038/nature03207.

43. Boisvert, F.-M., van Koningsbruggen, S., Navascués, J., and Lamond, A.I. (2007). The multifunctional nucleolus. Nature Reviews Molecular Cell Biology *8*, 574–585. 10.1038/nrm2184.

44. Kresoja-Rakic, J., and Santoro, R. (2019). Nucleolus and rRNA Gene Chromatin in Early Embryo Development. Trends in Genetics. 10.1016/j.tig.2019.06.005.

45. Plath, K., Mlynarczyk-Evans, S., Nusinow, D.A., and Panning, B. (2002). Xist RNA and the Mechanism of X Chromosome Inactivation . Annual Review of Genetics. 10.1146/annurev.genet.36.042902.092433.

46. Wutz, A., and Jaenisch, R. (2000). A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. Molecular cell *5*, 695–705. 10.1016/s1097-2765(00)80248-8.

47. Chaumeil, J., Le Baccon, P., Wutz, A., and Heard, E. (2006). A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. Genes and Development *20*, 2223–2237. 10.1101/gad.380906.

48. Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. Science *341*, 1237973–1237973. 10.1126/science.1237973.

49. 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., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Molecular Cell. 10.1016/j.molcel.2010.08.011.

50. Bell, J.C., Jukam, D., Teran, N.A., Risca, V.I., Smith, O.K., Johnson, W.L., Skotheim, J.M., Greenleaf, W.J., and Straight, A.F. (2018). Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. eLife. 10.7554/eLife.27024.

51. Li, X., Zhou, B., Chen, L., Gou, L.T., Li, H., and Fu, X.D. (2017). GRID-seq reveals the global RNA-chromatin interactome. Nature Biotechnology. 10.1038/nbt.3968.

52. Yan, Z., Huang, N., Wu, W., Chen, W., Jiang, Y., Chen, J., Huang, X., Wen, X., Xu, J., Jin, Q., et al. (2019). Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs. Proceedings of the National Academy of Sciences of the United States of America. 10.1073/pnas.1819788116.

53. Sridhar, B., Rivas-Astroza, M., Nguyen, T.C., Chen, W., Yan, Z., Cao, X., Hebert, L., and Zhong, S. (2017). Systematic Mapping of RNA-Chromatin Interactions In Vivo. Current Biology. 10.1016/j.cub.2017.01.011.

54. Bonetti, A., Agostini, F., Suzuki, A.M., Hashimoto, K., Pascarella, G., Gimenez, J., Roos, L., Nash, A.J., Ghilotti, M., Cameron, C.J., et al. (2019). RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. bioRxiv. 10.1101/681924.

55. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744-757.e24. 10.1016/j.cell.2018.05.024.

56. Simon, M.D., Pinter, S.F., Fang, R., Sarma, K., Rutenberg-Schoenberg, M., Bowman, S.K., Kesner, B.A., Maier, V.K., Kingston, R.E., and Lee, J.T. (2013). High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation. Nature. 10.1038/nature12719.

57. Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. Science *341*, 1237973–1237973. 10.1126/science.1237973.

58. West, J.A., Davis, C.P., Sunwoo, H., Simon, M.D., Sadreyev, R.I., Wang, P.I., Tolstorukov, M.Y., and Kingston, R.E. (2014). The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites. Molecular Cell. 10.1016/j.molcel.2014.07.012.

59. Engreitz, J.M., Sirokman, K., McDonel, P., Shishkin, A.A., Surka, C., Russell, P., Grossman, S.R., Chow, A.Y., Guttman, M., and Lander, E.S. (2014). RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. Cell *159*, 188–199. 10.1016/j.cell.2014.08.018.

60. Schoeftner, S., and Blasco, M.A. (2008). Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. Nature Cell Biology. 10.1038/ncb1685.

61. Mumbach, M.R., Granja, J.M., Flynn, R.A., Roake, C.M., Satpathy, A.T., Rubin, A.J., Qi, Y., Jiang, Z., Shams, S., Louie, B.H., et al. (2019). HiChIRP reveals RNA-associated chromosome conformation. Nature Methods. 10.1038/s41592-019-0407-x.

62. Decker, C.J., and Parker, R. (2012). P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. Cold Spring Harbor perspectives in biology *4*. 10.1101/cshperspect.a012286.

63. Wolozin, B., and Ivanov, P. (2019). Stress granules and neurodegeneration. Nature Reviews Neuroscience. 10.1038/s41583-019-0222-5.

64. Banani, S.F., Lee, H.O., Hyman, A.A., and Rosen, M.K. (2017). Biomolecular condensates: Organizers of cellular biochemistry. Nature Reviews Molecular Cell Biology. 10.1038/nrm.2017.7.

65. Goldfarb, K.C., and Cech, T.R. (2017). Targeted CRISPR disruption reveals a role for RNase MRP RNA in human preribosomal RNA processing. Genes and Development. 10.1101/gad.286963.116.

66. Dragon, F., Compagnone-Post, P.A., Mitchell, B.M., Porwancher, K.A., Wehner, K.A., Wormsley, S., Settlage, R.E., Shabanowitz, J., Osheim, Y., Beyer, A.L., et al. (2002). A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. Nature. 10.1038/nature00769.

67. Baßler, J., and Hurt, E. (2019). Eukaryotic Ribosome Assembly. Annual Review of Biochemistry. 10.1146/annurev-biochem-013118-110817.

68. Lee, Y., and Rio, D.C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. Annual Review of Biochemistry. 10.1146/annurev-biochem-060614-034316.

69. Neugebauer, K.M. (2002). On the importance of being co-transcriptional. Journal of Cell Science. 10.1242/jcs.00073.

70. Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nature Reviews Genetics *15*, 163–175. 10.1038/nrg3662.

71. McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., Patterson, S.D., Wickens, M., and Bentley, D.L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature. 10.1038/385357a0.

72. Martin, R.M., Rino, J., Carvalho, C., Kirchhausen, T., and Carmo-Fonseca, M. (2013). Live-Cell Visualization of Pre-mRNA Splicing with Single-Molecule Sensitivity. Cell Reports. 10.1016/j.celrep.2013.08.013.

73. Herzel, L., Ottoz, D.S.M., Alpert, T., and Neugebauer, K.M. (2017). Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. Nature Reviews Molecular Cell Biology. 10.1038/nrm.2017.63.

74. Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA. 10.1261/rna.1714509.

75. Maden, B.E.H. (1990). The Numerous Modified Nucleotides in Eukaryotic Ribosomal RNA. Progress in Nucleic Acid Research and Molecular Biology. 10.1016/S0079-6603(08)60629-7.

76. Reddy, R., and Busch, H. (1988). Small Nuclear RNAs: RNA Sequences, Structure, and Modifications. In Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles 10.1007/978-3-642-73020-7_1.

77. Tycowski, K.T., You, Z.H., Graham, P.J., and Steitz, J.A. (1998). Modification of U6 spliceosomal RNA is guided by other small RNAs. Molecular Cell. 10.1016/S1097-2765(00)80161-6.

78. Nizami, Z., Deryusheva, S., and Gall, J.G. (2010). The Cajal body and histone locus body. Cold Spring Harbor perspectives in biology *2*. 10.1101/cshperspect.a000653.

79. Darzacq, X., Jády, B.E., Verheggen, C., Kiss, A.M., Bertrand, E., and Kiss, T. (2002). Cajal body-specific small nuclear RNAs: A novel class of 2′-O-methylation and pseudouridylation guide RNAs. EMBO Journal. 10.1093/emboj/21.11.2746.

80. Richard, P., Darzacq, X., Bertrand, E., Jády, B.E., Verheggen, C., and Kiss, T. (2003). A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs. EMBO Journal. 10.1093/emboj/cdg394.

81. Karijolich, J., and Yu, Y.T. (2010). Spliceosomal snRNA modifications and their function. RNA Biology. 10.4161/rna.7.2.11207.

82. Machyna, M., Neugebauer, K.M., and Staněk, D. (2015). Coilin: The first 25 years. RNA Biology. 10.1080/15476286.2015.1034923.

83. Machyna, M., Heyn, P., and Neugebauer, K.M. (2013). Cajal bodies: Where form meets function. Wiley Interdisciplinary Reviews: RNA. 10.1002/wrna.1139.

84. Deryusheva, S., and Gall, J.G. (2009). Small Cajal body-specific RNAs of Drosophila function in the absence of Cajal bodies. Molecular Biology of the Cell. 10.1091/mbc.E09-09-0777.

85. Smith, K.P., Carter, K.C., Johnson, C. V., and Lawrence, J.B. (1995). U2 and U1 snRNA gene loci associate with coiled bodies. Journal of Cellular Biochemistry. 10.1002/jcb.240590408.

86. Marzluff, W.F., Wagner, E.J., and Duronio, R.J. (2008). Metabolism and regulation of canonical histone mRNAs: Life without a poly(A) tail. Nature Reviews Genetics. 10.1038/nrg2438.

87. Marzluff, W.F., and Koreski, K.P. (2017). Birth and Death of Histone mRNAs. Trends in Genetics. 10.1016/j.tig.2017.07.014.

88. Nizami, Z., Deryusheva, S., and Gall, J.G. (2010). The Cajal body and histone locus body. Cold Spring Harbor perspectives in biology. 10.1101/cshperspect.a000653.

89. Calvet, J.P., and Pederson, T. (1981). Base-pairing interactions between small nuclear RNAs and nuclear RNA precursors as revealed by psoralen cross-linking in vivo. Cell. 10.1016/0092-8674(81)90205-1.

90. Maxwell, E.S., and Fournier, M.J. (1995). The small nucleolar RNAs. Annual review of biochemistry *64*, 897–934. 10.1146/annurev.bi.64.070195.004341.

91. Mowry, K.L., and Steitz, J.A. (1987). Both conserved signals on mammalian histone pre-mRNAs associate with small nuclear ribonucleoproteins during 3’ end formation in vitro. Molecular and Cellular Biology *7*, 1663–1672. 10.1128/mcb.7.5.1663-1672.1987.

92. Schaufele, F., Gilmartin, G.M., Bannwarth, W., and Birnstiel, M.L. (1986). Compensatory mutations suggest that base-pairing with a small nuclear RNA is required to form the 3′ end of H3 messenger RNA. Nature *323*, 777–781. 10.1038/323777a0.

93. Dominski, Z., and Marzluff, W.F. (2007). Formation of the 3′ end of histone mRNA: Getting closer to the end. Gene *396*, 373–390. 10.1016/j.gene.2007.04.021.

94. Bensaude, O. (2011). Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? Transcription *2*, 103–108. 10.4161/trns.2.3.16172.

95. Kaiser, T.E., Intine, R. V., and Dundr, M. (2008). De novo formation of a subnuclear body. Science. 10.1126/science.1165216.

96. Chao, S.H., and Price, D.H. (2001). Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo. Journal of Biological Chemistry. 10.1074/jbc.M102306200.

97. 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., et al. (2017). SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin-Associated RNAs. Cell. 10.1016/j.cell.2017.05.029.

98. Nozawa, R.S., and Gilbert, N. (2019). RNA: Nuclear Glue for Folding the Genome. Trends in Cell Biology. 10.1016/j.tcb.2018.12.003.

99. Guo, Y.E., Manteiga, J.C., Henninger, J.E., Sabari, B.R., Dall'Agnese, A., Hannett, N.M., Spille, J.H., Afeyan, L.K., Zamudio, A. V., Shrinivas, K., et al. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. Nature. 10.1038/s41586-019-1464-0.

100. Cho, W.K., Spille, J.H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I.I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. Science. 10.1126/science.aar4199.

101. Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. Cell. 10.1016/j.cell.2017.02.007.

102. Gall, J.G. (2000). Cajal Bodies: The First 100 Years. Annual Review of Cell and Developmental Biology. 10.1146/annurev.cellbio.16.1.273.

103. Jády, B.E., and Kiss, T. (2001). A small nucleolar guide RNA functions both in 2′-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. EMBO Journal. 10.1093/emboj/20.3.541.

104. Ogg, S.C., and Lamond, A.I. (2002). Cajal bodies and coilin - Moving towards function. Journal of Cell Biology. 10.1083/jcb.200206111.

105. Matera, A.G., and Frey, M.R. (1998). Coiled bodies and gems: Janus or gemini? American Journal of Human Genetics. 10.1086/301992.

106. Tucker, K.E., Berciano, M.T., Jacobs, E.Y., LePage, D.F., Shpargel, K.B., Rossire, J.J., Chan, E.K.L., Lafarga, M., Conlon, R.A., and Gregory Matera, A. (2001). Residual Cajal bodies in coilin knockout mice fail to recruit Sm snRNPs and SMN, the spinal muscular atrophy gene product. Journal of Cell Biology. 10.1083/jcb.200104083.

107. Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA Exploits

Three-Dimensional Genome Architecture to Spread Across the X Chromosome. Science *341*, 1237973–1237973. 10.1126/science.1237973.

108. Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J., Bhattacharyya, R.P., Rudy, R.F., Patel, M.M., et al. (2015). Simultaneous generation of many RNA-seq libraries in a single reaction. Nature Methods *12*, 323–325. 10.1038/nmeth.3313.

109. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 10.14806/ej.17.1.200.

110. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: A fast spliced aligner with low memory requirements. Nature Methods. 10.1038/nmeth.3317.

111. Jin, Y., Tam, O.H., Paniagua, E., and Hammell, M. (2015). TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. 10.1093/bioinformatics/btv422.

112. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods *9*, 357–359. 10.1038/nmeth.1923.

113. Krueger, F., and Andrews, S.R. (2016). SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. F1000Research. 10.12688/f1000research.9037.1.

114. Smit, A., Hubley, R., and Grenn, P. (2015). RepeatMasker Open-4.0. RepeatMasker Open-4.0.7.

115. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 10.1093/bioinformatics/btq033.

116. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 10.1093/bioinformatics/btw354.

117. Chan, P.P., and Lowe, T.M. (2009). GtRNAdb: A database of transfer RNA genes detected in genomic sequence. Nucleic Acids Research. 10.1093/nar/gkn787.

118. Chan, P.P., and Lowe, T.M. (2016). GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Research. 10.1093/nar/gkv1309.

119. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M.R., et al. (2005). Three-dimensional maps of all

chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biology. 10.1371/journal.pbio.0030157.

120. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nature Biotechnology. 10.1038/nbt.1754.

121. Imakaev, M., Fudenberg, G., McCord, R.P., Naumova, N., Goloborodko, A., Lajoie, B.R., Dekker, J., and Mirny, L.A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods. 10.1038/nmeth.2148.

Chapter 5

CONCLUSION AND FUTURE PERSPECTIVES

Gene regulation involves the coordinated actions of many regulatory factors to direct spatial- and temporal-specific gene expression programs within each cell. Propelled by rapid advancements in sequencing technologies, researchers have uncovered the multifactorial nature of regulation, implicating critical roles for thousands of components including cis-regulatory elements, epigenomic modifications, sequence-specific transcription factors, DNA modifying enzymes, chromatin-regulatory proteins, non-coding RNAs and 3D genomic organization. Despite a wealth of consortium-generated data, how a single genome guides distinctive, cell-type specific expression patterns for hundreds of individual cell types remains elusive.

Learning the rules of gene regulation has been challenging. This is because current experimental methods are limited in their ability to elucidate the functional relationships between complex assemblies of regulatory factors and transcriptional output. More specifically, most sequencing-based genomics approaches only measure pairwise interactions (e.g., protein-to-DNA or DNA-to-DNA) and, for regulatory proteins, are technically limited to studying one protein at a time. Previous studies using such methods have demonstrated correlations between transcription and single regulatory proteins, non-coding RNAs, epigenomic signatures or genome structure. However, whether these regulatory factors are merely coincidental with or necessary for transcription activity/inactivity has remained unknown. Pairwise interaction measurements simply cannot appreciate the temporal and locus-specific complexity of gene regulation nor detect higher-order regulatory structures or logic circuits.

Advancing our understanding of how the genome programs gene expression will require the development of new experimental methods, data visualization strategies and computational models that account for the multifactorial, context-dependent nature of gene regulation. Specifically, we need novel strategies to increase the scale and diversity of measurements and directly measure the associations of regulatory factors with each other and with transcription. Here, we propose SPRITE as an enabling technology because it can

capture ‘snapshots’ of the molecular events within the nucleus. SPRITE detects the multiway, multimodal interactions occurring between many individual molecules within a single cell at a single point in time. Previous adaptations of SPRITE have focused on mapping nucleic acid interactions (e.g., DNA-DNA, RNA-DNA, RNA-RNA); to build a comprehensive model for gene regulation, future adaptions need to incorporate protein detection.

In this thesis, I describe RNA-DNA SPRITE, a method for simultaneous measurement of multiway RNA and DNA interactions, and ChIP-DIP, a highly multiplexed method for mapping regulatory proteins to DNA. ChIP-DIP provides a drastic increase in scale for cell-type specific regulatory protein maps and was used to demonstrate that additional information is encoded in the quantitative combinations of histone modifications at regulatory elements. RNA-DNA SPRITE provides a direct readout of higher order organizations of RNA and DNA and was used to demonstrate the bidirectional, functional relationship between nascent transcription and genome organization. Combining these two methods would provide locus-specific, multi-modal information (e.g., co-localizing ncRNAs, 3D genomic structure, nascent transcription) for many individual regulatory proteins. Alternatively, strategies that allow for multi-way detection of proteins could also be designed. For instance, first ChIP-DIP could be used to rapidly screen and identify high-quality protein affinity reagents; following, these antibodies could be covalently modified with unique identifying sequences and used for a pooled immunoprecipitation; finally, the sample would be processed through the standard RNA-DNA SPRITE workup. This workflow would be highly modular and allow for multi-way detection of interactions between proteins, RNAs, and DNA.

In summary, the fields of genomics and epigenomics stand at a turning point. Previous research has produced a wealth of regulatory candidates but a unified model for gene regulation remains elusive. New experimental methods capable of measuring the inter-connections of regulatory factors at scale and the corresponding analytical and theoretical

frameworks will be required to define the principles controlling gene expression and understand the mechanisms involved in human health and disease etiology.