On the Role of Three-Dimensional Genome Organization in Gene Regulation and mRNA Splicing

Thesis by
Prashant Bhat

In Partial Fulfillment of the Requirements for the degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2023
(Defended May 3, 2023)

Prashant Bhat
ORCID: 0000-0003-3832-4871

ACKNOWLEDGMENTS

I have had an incredible time at Caltech. Caltech's small size and academically stimulating atmosphere resemble that of an intellectual monastery, which fosters an environment where great discoveries require time and careful cultivation. There are many people I must thank.

Professor Mitchell "Mitch" Guttman has been an outstanding mentor during my time in the lab, and I've gained invaluable life lessons from him. I liked how invested Mitch was in the details of my project. I had the opportunity to spend countless hours studying experimental data, analytical tools, and learning Mitch's thought processes. Furthermore, I was able to improve my ability to give effective presentations, write grants, and manage larger teams. Mitch was also supportive of my desire to learn new methodologies and tackle the field of splicing, which no one in the lab had previously pursued. He not only allowed me to take a substantial risk, but also encouraged me to dream big. I am lucky to call him a friend and an everlasting source of support.

Within the lab there are many people I wish to thank. Most importantly, Sofia "Sofi" Quinodoz trained me in genomics in the initial part of my PhD and we have spent countless hours talking science, pipetting, and nervously awaiting traces on the BioAnalyzer late in the night. She continues to be an incredible colleague, mentor, and friend. Mackenzie Strehle has been my closest friend in the lab and will miss our daily chats between experiments and meetings. Further, much of what I have learned about microscopy is from her. Amy Chow has helped me a great deal throughout my PhD and has answered my many questions. In addition to that, she is an incredible manager of the lab. The lab would not function without Amy. I also thank Mario Blanco and Joanna Jachowicz, both of whom have given me invaluable personal and professional advice. I thank Inna-Marie Strazhnik for her incredible ability to bring our data to life through her illustrations. I also thank Shawna Hiley for her help with editing and refining the language of our manuscripts. There are many other

lab members I have had the pleasure of working with: Olivia Ettlin, Elizabeth Soehalim, Allen Chen, Wesley Huang, Drew Honson, Isabel Goronzy, Chris Chen, Vickie Trinh, Elizabeth Detmar, Ali Palla, and Parham Peyda.

I also would like to thank my collaborators to whom I owe a great deal of gratitude for their time and shared passion for science. Special thanks to Benjamin "Ben" Emert, who provided expertise in single molecule FISH and advice about a career as an MD/PhD. Yodai Takei is one of the most talented scientists and most genuine people I know; it has been a pleasure collaborating with him. I also thank Igor Antoshechkin for being so accommodating with our last-minute sequencing runs.

I also thank several faculty mentors. I thank David Baltimore, Doug Black, Howard Chang, and Michael Elowitz for providing valuable advice and guidance on my thesis. Doug has been especially helpful with his encyclopedic knowledge of splicing. I thank Lior Pachter, who has been an incredible mentor and friend – he has taught me how to simultaneously be a caring person and exceptional scientist. I thank Barbara Wold for providing myocytes for us to extend our speckle findings to other cell types. I also thank Magdalena Zernicka-Goetz, Mary Ann Kennedy, Shasha Chong, and Fangyuan Ding for insightful conversations and further guidance. A huge amount of gratitude to my undergraduate research mentors, Jennifer Doudna and Sam Sternberg, where the discovery of CRISPR was the launch of my interest in understanding the fundamentals of RNA.

I thank Liz Ayala, Raina Beaven, and Lauren Breeyear for being the engine that powers the division of biology and biological engineering.

I thank my friends who brought my Caltech experience to life. Luke, who is the original "splicing dude" and introduced me to splicing and its relevance to disease. I also appreciate his friendship and advice throughout the years. Many thanks to Tara Chari, Ben Riviere, Kent Leslie, Levi

Palmer, Patrick Nutter, Lev Tsypin, Alex Winnett, Michael Porter, Adrian Bruckner, and Dustin Buccino, with whom I have cultivated a great support system and friendships.

My family has been a source of both inspiration and support. My mom and dad have provided an endless amount of support and advice my whole life. Their experiences as scientists have inspired me in many ways. I have tried to emulate my mom's incredible notetaking and attention to detail and my dad's ability to transcend divergent fields – having taught himself biology as a medicinal chemist. My parents have given me so much and have been the biggest cheerleaders for my scientific passions. I thank my sister, Priya Bhat-Patel who has taught me confidence and to always stand up for what is right, despite the barriers. She has cultivated in me the mindset of, "If not me, then who?" I also am grateful to my brother-in-law, Tej Patel, for support and my dearest nephew, Jahan, who is a bundle of joy and is growing up to be a very intelligent and caring boy.

Finally, I thank my best friend and life partner Kishan Patel, for being the person who supports me every day with love and encouragement. His boundless optimism and radiating happiness have illuminated my life and made me excited for the years to come.

ABSTRACT

The nucleus is spatially organized such that DNA, RNA, and protein molecules involved in shared functional and regulatory processes are compartmentalized in three-dimensional (3D) structures. These structures are emerging as a paradigm for gene regulation, a highly complex process that requires the dynamic coordination of hundreds of regulatory factors around precise targets in different cell states. We describe the discovery of hundreds of RNA-DNA hubs throughout the nucleus that are organized around essential nuclear functions such as RNA processing, centromeric heterochromatin organization, and gene regulation. Focusing on RNA processing, specifically co-transcriptional splicing, we find that genome-wide organization of active genes near nuclear speckles drives the efficiency of pre-mRNA splicing in a cell-type specific manner. The results of this thesis illustrate how spatial compartmentalization of biomolecules increases the local concentration of reactants and enzymes such that greater efficiency is achieved in scenarios where rapid responses are required for cell survival.

PUBLISHED CONTENT AND CONTRIBUTIONS

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

   P.B. wrote the review with D.H. and M.G.

2. Quinodoz, S. A., Bhat, P.*, Jachowicz, J.*, Ollikainen, N.*, *et al.* RNA promotes the formation of spatial compartments in the nucleus. *Cell* **184**, 5775-5790.e30 (2021).
   *Equal Contribution

   P.B. developed and optimized the RD-SPRITE protocol; performed SPRITE experiments; analyzed and interpreted data; contributed to data visualization, figure presentation, and model schematics/illustrations; and wrote the paper with S.A.Q and M.G.

3. Bhat, P. *et al.* 3D genome organization around nuclear speckles drives mRNA splicing efficiency. bioRxiv doi: https://doi.org/10.1101/2023.01.04.522632 (In Revision)

   P.B. conceived and led the study, designed experiments, performed experiments, analyzed data, interpreted results, and wrote the manuscript with input and support from M.G. and other authors.

TABLE OF CONTENTS

Chapter 1: Introduction

NUCLEAR COMPARTMENTALIZATION AS A MECHANISM OF QUANTITATIVE CONTROL OF GENE EXPRESSION

Prashant Bhat, Drew Honson, and Mitchell Guttman

A modified version of this chapter was published as: Bhat, P., Honson, D., & Guttman, M. Nuclear compartmentalization as a mechanism of quantitative control of gene expression. *Nat. Rev. Mol. Cell Biol.* **22**, 653–670 (2021). https://doi.org/10.1038/s41580-021-00387-1

"Inside the nucleus, genes recline along chromosomes, a word that means 'colored bodies.' They coil and uncoil, winding and unwinding, gathering and releasing information. Proteins bind to them, like readers holding an ancient text. And in this crowded, three-dimensional city of cells, genes and proteins meet, collide, separate, and merge in a brilliant choreography. It is here, in this dance of molecules, that life finds its rhythm and its meaning." – Siddhartha Mukherjee in The Gene: An Intimate History

1.1 ABSTRACT

Gene regulation requires the dynamic coordination of hundreds of regulatory factors at precise targets. While many regulatory factors have specific affinity for their nucleic acid targets, molecular diffusion and affinity models alone cannot explain many of the quantitative features of these processes in the nucleus. One emerging explanation for these non-linear regulatory properties is that DNA, RNA, and proteins organize within precise, three-dimensional (3D) compartments in the nucleus to concentrate groups of functionally related molecules. Recently, nucleic acids and proteins involved in many critical nuclear processes have been shown to engage in cooperative interactions that lead to the formation of condensates that act to partition molecular components. In this introductory chapter, we discuss an emerging perspective in gene regulation that moves away from classic stoichiometric models towards an understanding of how spatial compartmentalization can lead to non-stoichiometric molecular interactions and non-linear regulatory behaviors. We describe key features of nuclear compartmentalization and their importance in controlling transcriptional regulation, higher-order chromatin structure, and RNA processing and more generally explore how these properties may explain critical quantitative aspects of gene regulation.

1.2 INTRODUCTION

Gene regulation is a highly complex process that requires the dynamic coordination of hundreds of regulatory factors – including chromatin and transcriptional regulators, mRNA processing and splicing factors – at precise molecular targets in different cell states. While many regulatory factors have specific affinity for their targets (e.g.. DNA sequences, RNA structures), molecular diffusion and affinity models alone cannot explain many of the observed quantitative properties of these regulatory processes in the nucleus. For example, the rate at which transcription factors bind to their targets is >1000-fold faster than is predicted from diffusion and affinity alone[1,2].

Recent advances in our understanding of nuclear organization, driven by advances in genomic and microscopy methods (**Box 1**), have begun to elucidate a new paradigm that may explain many of these regulatory properties. Specifically, DNA, RNA, and protein molecules can organize within precise three-dimensional territories in the nucleus to concentrate groups of functionally related molecules[3–6]. For example, genomic DNA is dynamically organized to promote enhancer-promoter interactions[7], topological association of sets of co-regulated genes[8], and recruitment of DNA and pre-mRNA to different nuclear compartments[9,10]. Moreover, many proteins involved in transcriptional regulation (e.g., RNA Pol II[11], TFs[12]), enhancer-promoter interactions (e.g., Mediator[11]), heterochromatin formation and maintenance (e.g., HP1[13,14], SAFB[15]) and mRNA splicing (e.g., FUS[16]) have been shown to be enriched within high-concentration territories within the nucleus. A central tenet of this model is that formation of nuclear compartments can act to partition molecular components and biochemical functions[17–19] (**Figure 1A**).

In this introductory chapter, we explore an emerging paradigm for gene regulation that moves away from classic stoichiometric models and describe how spatial compartmentalization can lead to non-stoichiometric molecular interactions and non-linear regulatory behaviors. We discuss the molecular mechanisms by which compartmentalization is achieved

and the role of compartmentalization in spatially organizing enhancers, promoters, and transcription factors to drive transcriptional initiation. In addition, we explore the role of compartmentalization in controlling higher-order chromatin organization and regulation, and co-transcriptional RNA processing. In each of these cases, we discuss the key features of nuclear compartmentalization that enable the non-linear behaviors that quantitatively control critical aspects of gene regulation.

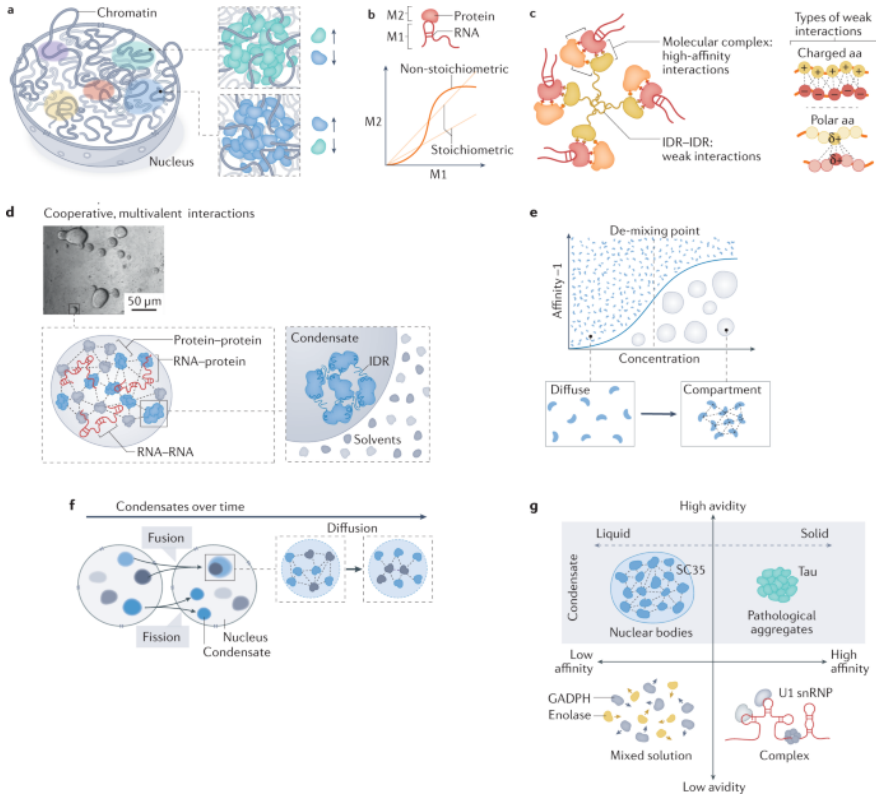


Figure 1 *| a | Nuclear compartments contain high local concentration of specific molecules in 3D space. Upward arrows represent higher concentration and downward arrows represent lower concentration. b | Stoichiometric interactions between two molecules, M1 (an RNA) and M2 (a protein), show a linear increase in the number of M2 molecules bound per unit increase in M1. By contrast, non-stoichiometric interactions lead to an exponential increase in binding of the target molecule (M2) as M1 increases. c | Stoichiometric*

interactions between molecules and cooperative interactions between multivalent intrinsically disordered regions (IDRs) of proteins can be found within a sample of complexes. The low-affinity interactions between IDRs are often mediated by stretches of charged and polar amino acids (aa). d | Image of polypyrimidine tract-binding protein 1 (PTBP1) molecules undergoing concentration-dependent condensate formation with RNA in vitro (top). When protein concentration is high, multivalent interactions can promote 'de-mixing' (bottom). e | Phase diagram describing whether molecules will be present in a diffusive state or a compartmentalized state. Formation of these structures is controlled by local concentration leading to sharp transition behaviors (blue line). f | Condensates can exhibit properties of liquids, including the ability to split (fission), merge (fusion) and undergo rapid internal rearrangement (diffusion). g | Cartesian plane highlighting the relationship between affinity, avidity and physical manifestations of different types of molecular associations. Well-mixed solution of two soluble proteins with little to no affinity for one another, such as GAPDH and enolase (bottom left). U1 small nuclear ribonucleoprotein (U1 snRNP) is an example of a macromolecular complex composed of high-affinity interactions between a fixed stoichiometric ratio of interacting component molecules (bottom right). Nuclear bodies are composed of self-interacting proteins, for example, SC35 (also known as SRSF2) proteins in nuclear speckles (top left). Solid aggregates of tau are observed in neurodegenerative diseases (top right). Image courtesy of Inna-Marie Strazhnik/Caltech. Part d adapted from ref.47, Springer Nature Limited.

1.3 MECHANISMS OF NUCLEAR COMPARTMENT FORMATION

Macromolecular complex formation between protein, DNA, and RNA molecules has traditionally been viewed through the lens of stoichiometric molecular interactions. These interactions generally occur through well-structured domains that form high-affinity contacts to form complexes containing fixed molecular ratios (**Figure 1B**). Recently, many critical regulatory proteins have also been shown to form low affinity interactions with other protein, DNA, or RNA molecules – primarily through multivalent interactions between unstructured, low complexity domains or intrinsically disordered regions (IDRs). While individually these weak associations might not enable specific interactions, cooperative association of molecules at high concentrations can facilitate the formation of biomolecular condensates (reviewed extensively[17,19,20]). Here, we describe how these two molecular mechanisms – stoichiometric molecular

interactions and condensate formation – can lead to nuclear compartmentalization.

1.3.1 MOLECULAR INTERACTIONS DRIVE FORMATION OF COMPLEXES AT FIXED STOICHIOMETRIC RATIOS

Macromolecular complexes are formed through interactions that occur between molecules at precise stoichiometric ratios (**Figure 1B**). For example, a single RNA Pol II holoenzyme is formed by high affinity protein-protein interactions between RNA Pol II and general transcription factors such as TFIIF[21]. Transcription factors can also bind with high affinity to specific DNA sequences within the major groove of the double helix through their DNA binding domains (e.g., zinc finger domains and leucine zippers)[22]. Similarly, high affinity protein-RNA and RNA-RNA interactions contribute to a variety of macromolecular complexes in the nucleus. For example, various proteins containing RNA binding domains (i.e., RRMs[23] and KH[24] domains) bind directly to specific RNA sequences or structures and specific RNAs directly hybridize with other RNA molecules to form macromolecular complexes. A single U4/U6.U5 tri-snRNP spliceosomal complex is assembled via extensive base pairing between U4 and U6 snRNAs and high affinity interactions between more than 30 proteins that interact with each other and snRNAs[25].

To date, the vast majority of studied transcriptional and post-transcriptional regulators represent high affinity macromolecular complexes **(Figure 1C)**. This is in large part because traditional methods for exploring the structure of macromolecular complexes (e.g., X-ray crystallography) require formation of stable, high-affinity conformations. Despite the critical importance of such macromolecular complexes in nuclear functions, many proteins involved in transcriptional regulation (e.g., Mediator, RNA Pol II, and cell-type specific transcription factors), chromatin regulation (e.g., Polycomb, HP1), and RNA processing (e.g., SRSF1, FUS, PTBP1) also contain large IDRs that do not form well-

structured domains and have traditionally been excluded from expressed proteins when attempting to solve the structure of macromolecular complexes[26–28].

1.3.2 COOPERATIVE, MULTIVALENT INTERACTIONS CAN DRIVE FORMATION OF MOLECULAR CONDENSATES

In contrast to macromolecular complexes that occur at fixed stoichiometries, molecules can also form condensates that contain multiple interacting molecular components that can associate at variable stoichiometries where the molecular components are spatially enriched relative to the surrounding cellular environment[17–19] (**Figure 1D, left**). Condensate formation is a concentration-dependent process that is often driven by molecules that engage in cooperative, multivalent, low affinity, interactions[29].

Chemistry and soft-matter physics have provided insights into the thermodynamics underlying condensate formation in cells[30] (**Figure 1E**). A useful way to conceptualize this is to consider a set of molecules (e.g., multiple copies of a particular protein, "A") mixed with a collection of other molecules (e.g., nucleoplasm, "B"). If the "A" molecules and "B" molecules are attracted to one another, a well-mixed solution will form to maximize the entropy of the system. However, if the "A" molecules exhibit preferential molecular attraction to other "A" molecules compared to "B" molecules, formation of these preferential "A" interactions will be more energetically favorable than random mixing of "A" and "B" (**Figure 1E**). While molecules that achieve weak affinity of individual interactions (e.g., μM binders) might be insufficient to promote energetically favorable association, interactions that occur at multiple independent sites (multivalency) lead to a large increase in the overall affinity – referred to as avidity[31]– between molecules. When the concentration of "A" molecules achieves a critical concentration threshold, the self-interacting "A" molecules will "demix" from the other "B" molecules (**Figure 1E**), a

process referred to as concentration-dependent phase separation. The thermodynamics of multivalency[19,31], physics of phase separation[18,30], and their features and limitations for thinking about biological processes[18–20,32,33] are explored in more detail in several reviews. In this Review, we use the term condensate to refer to molecular assemblies that are formed through concentration-dependent, multivalent associations regardless of whether they undergo phase separation.

Many of the proteins that are important for promoting condensate formation contain large regions that do not form well-structured domains (referred to as IDRs). These IDRs are often composed of charged and polar amino acids which can facilitate low affinity interactions[34] (**Figure 1C**). The charge properties of these IDRs combined with their length allow them to engage in multivalent, low affinity, homotypic and heterotypic interactions. Because multivalent binding can induce a multiplicative increase in the overall avidity between two molecules, systematically increasing the number of possible interactions that a given molecule can form ("valency") enables association and demixing to occur at lower overall molecular concentrations[18,35] (**Figure 1D, right**). Importantly, such concentration-dependent assemblies can form through homotypic or heterotypic multivalent associations of RNA, DNA and protein molecules[36,37].

1.3.3 CONDENSATES WITH PROPERTIES OF LIQUIDS, GELS, OR SOLIDS CAN FORM VIA PHASE SEPARATION

Molecules can undergo phase separation to form condensates that exist within different physical states including liquid, solid, and intermediate ("gel-like") states (**Figure 1E**). The key to formation of the liquid state – which involves rapid molecular exchange within a condensate – is that each individual interaction is weak. Because of the low affinity of individual interactions, proteins can associate and disassociate within the condensate. In contrast, if there is a strong affinity between individual

interacting components, high avidity can still be achieved but it will restrict diffusion within the condensate and result in a more rigid, solid-like structure. Because affinity is a continuous property, the precise physical states of a condensate can be thought of as a continuum across the liquid to solid dimension (**Figure 1F)**.

The specific phenomenon of phase separation to form liquid-like condensates is known as liquid-liquid phase separation (LLPS)[18]. One of the first discovered examples of LLPS-mediated condensate formation was P granules in *C. elegans*[38]. P granules were shown to form condensates that exhibit several liquid-like properties including (i) the ability to fuse with other compartments (fusion of droplets), (ii) the ability to split into distinct compartments (fission of droplets), (iii) the ability to dynamically rearrange the molecules within the compartment[38] (**Figure 1E**), and (iv) these compartments displayed a spherical shape that reflects its surface tension with the surrounding cytoplasm. By virtue of its liquid-like properties, the molecules contained within an LLPS compartment are expected to be well-mixed and uniformly distributed[32].

Several membraneless compartments in the nucleus have been shown to form liquid-like structures, including the nucleolus[39]. In addition, many proteins involved in gene regulation have been proposed to undergo LLPS. These include chromatin regulatory proteins (HP1[13,14] and Polycomb[40–42]), transcriptional machinery (Mediator[11,43], RNA Pol II[11,43], TAZ[44], BRD4[45]), transcription factors (OCT4[12], TAF15[46]), splicing factors (SRSF1[43], SRSF2[43]) and RNA processing factors (PTBP1[35,47]).

1.3.4 SPATIALLY CONSTRAINED NUCLEIC ACIDS CAN SEED COMPARTMENTALIZATION IN THE NUCLEUS

In order to form a nuclear compartment, molecules need to achieve high concentrations within a spatially constrained territory of the nucleus. Because nuclear proteins are translated in the cytoplasm and then

trafficked into the nucleus, they are intrinsically diffusible; they need to diffuse through the nucleus to associate with their targets. In theory, proteins that form condensates could stochastically come together via molecular diffusion at the concentrations needed to undergo condensate formation. Yet, this is likely a rare event because the overall concentrations of individual nuclear proteins are often too low to allow for the simultaneous association of the multiple proteins required[35]. Instead, many nuclear compartments are seeded through high-affinity interactions with spatially constrained molecules within the nucleus. These can include interactions with nucleic acids, histone modifications, or existing nuclear structures (e.g., nuclear lamina)[48,49]. Specifically, these spatially constrained molecules can bind to various diffusible molecules and recruit them to specific nuclear territories to seed formation of nuclear compartments (**Figure 2A**).

Many nuclear structures have been shown to form via interactions with non-coding RNAs (ncRNAs). For example, formation of the nucleolus is seeded by transcription of pre-ribosomal RNA[50,51], the paraspeckle is seeded by the Neat1 lncRNA[52], histone locus bodies are seeded around transcription of histone pre-mRNAs[53], and the Barr body (inactive X) is seeded by the Xist lncRNA[8,54]. Moreover, many chromatin regulators form spatial compartments in the nucleus that have been shown to be dependent on RNA (e.g., HP1[55–57], SHARP[56], and Polycomb[57,58]). For example, live cell imaging of SHARP (an RNA binding protein that recruits HDAC3) revealed dozens of condensate-like structures throughout the nucleus that become diffuse throughout the nucleus upon deletion of its RNA binding domain[56] (**Figure 2B**).

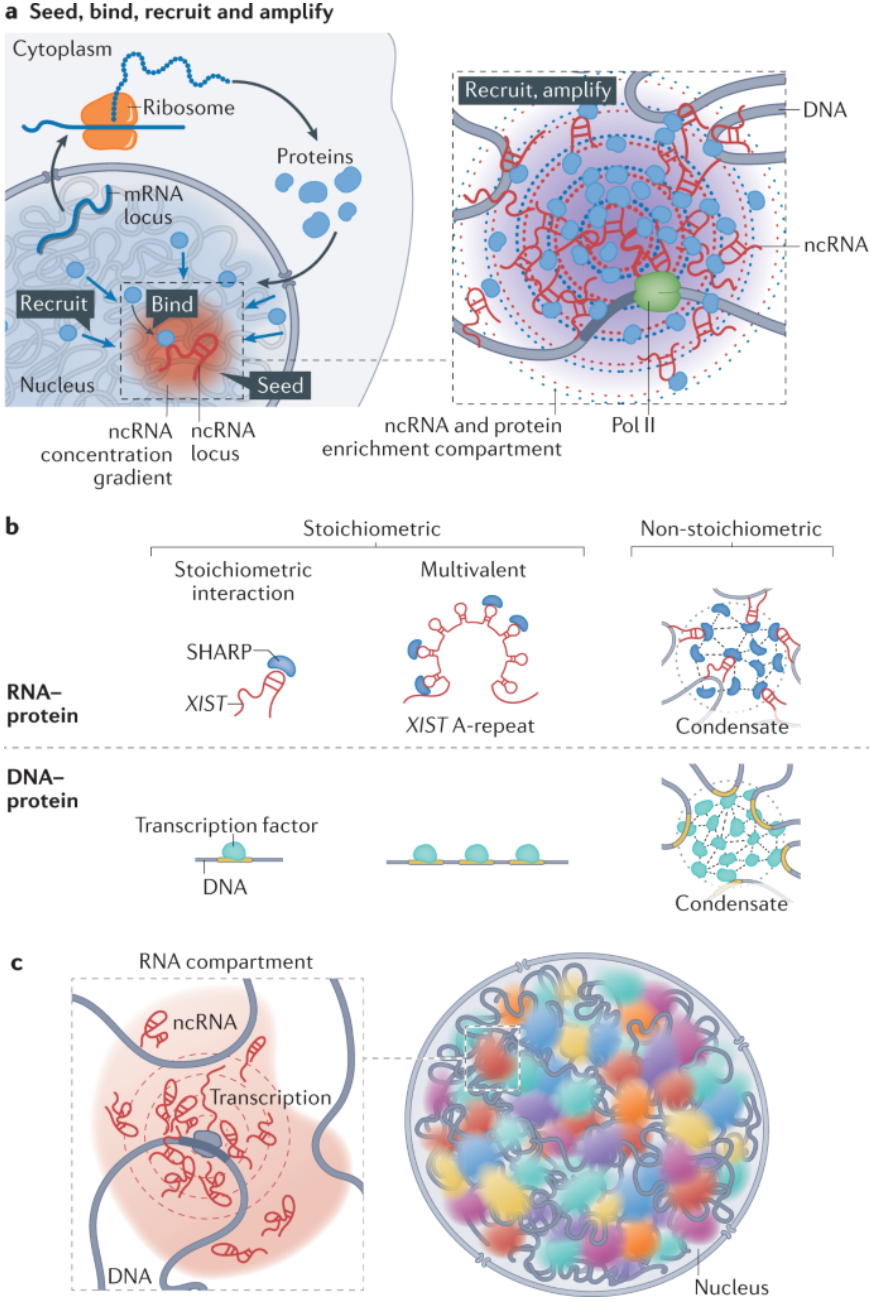

a Seed, bind, recruit and amplify
Cytoplasm
Ribosome
Proteins
mRNA locus
Recruit
Bind
Seed
Nucleus
ncRNA concentration gradient
ncRNA locus
Recruit, amplify
DNA
ncRNA
ncRNA and protein enrichment compartment
Pol II
b
Stoichiometric
Non-stoichiometric
Stoichiometric interaction
Multivalent
RNA–protein
SHARP
XIST
XIST A-repeat
Condensate
DNA–protein
Transcription factor
DNA
Condensate
c
RNA compartment
ncRNA
Transcription
DNA
Nucleus

Figure 2 *| a | Model of how non-coding RNAs (ncRNAs) can drive compartmentalization. Sites of ncRNA transcription can support high ncRNA concentrations and, thus, seed the formation of high-concentration territories (seed); ncRNAs can bind to diffusible proteins or ncRNAs through stoichiometric interactions (bind) and, in this manner, lead to enrichment of diffusible factors within a spatially defined region (recruit). In some cases, the recruited proteins can recruit other proteins and/or form condensates through homotypic and heterotypic interactions (not shown). b | Stoichiometric RNA–protein or DNA–protein interactions can involve single or multiple binding events, but the number of proteins recruited is limited to the number of available sites on the nucleic acid. By contrast, non-stoichiometric interactions enable binding of more proteins than available binding sites. c | Space filling model of localization of hundreds of ncRNAs (colored areas) across the nucleus in mouse embryonic stem cells. Pol II, polymerase II; XIST, X-inactive specific transcript. Image courtesy of Inna-Marie Strazhnik/Caltech.*

ncRNAs are especially well-suited to seed nuclear compartment formation because:

(1) *Transcription creates high concentrations of spatially-constrained RNA:* Once transcribed, a ncRNA can be retained at high concentration on chromatin near its site of transcription. Because transcription creates multiple copies of an RNA species, it can achieve higher local concentrations (relative to DNA) at these specific nuclear locations. In this way, a nuclear compartment can be dynamically created by controlling the expression of a specific ncRNA (**Figure 2A**).

(2) *Spatially-constrained ncRNA can bind to diffusible molecules*: Because ncRNAs contain sequence motifs and secondary structures that can bind diffusible RNA and protein molecules, these can form stoichiometric interactions that drive high local concentrations of these diffusible molecules. In specific cases, formation of these high local concentration territories can further promote concentration-dependent condensate formation through homotypic and heterotypic interactions[51,56,59] (**Figure 2A**). Interestingly, several ncRNAs contain multiple binding sites for the same protein, which can further increase valency of these interactions and local concentration[47,60–62] (**Figure 2C**).

Recently a pre-print showed that hundreds of ncRNAs can form high-concentration and stable territories near their site of transcription and therefore might similarly serve to seed the formation of various nuclear compartments of different sizes[56] (**Figure 2D)**.

In addition to ncRNAs, genomic DNA can also drive high spatial concentrations of protein complexes in the nucleus. For example, genomic DNA regions containing a high density of enhancers that each individually bind to Mediator can drive high local concentrations and condensate formation[33]. Similarly, the presence of multiple histone modifications on DNA can recruit high concentrations of chromatin "reader" proteins. Interestingly, several different chromatin reader proteins have been shown to form condensates in both normal and disease contexts (e.g., Cbx2[41,42], HP1[13,14,63], and MeCP2[64–66]). The presence of numerous protein binding sites on DNA that are present in close proximity ("spatial valency") may promote the spatial concentrations of proteins that are needed to drive condensate formation (**Figure 2C**).

These shared features – spatial anchoring in 3D space and the ability to bind and recruit diffusible molecules into high concentration territories – appear to be critical for seeding the formation of many nuclear structures.

1.3.5 INTEGRATIVE PERSPECTIVE: COMPARTMENTALIZATION REQUIRES MULTIPLE MECHANISMS *IN VIVO*

In this section, we presented various mechanisms that are critical for formation of nuclear compartments, including the ability for molecules to form stoichiometric interactions, the ability to form non-stoichiometric assemblies (condensates), and the ability for spatially anchored molecules to recruit diffusible molecules into precise nuclear territories. Importantly, these components are not mutually exclusive and assembly of complex nuclear compartments in cells likely utilize many of these – and possibly other – mechanisms simultaneously.

We note that the terminology in this field can be confusing and used differently by different groups (see **Box 2**). As an example, the term 'phase separation' is sometimes used to refer specifically to LLPS[32] and other times more broadly to describe a process by which condensates are formed[18–20]. As described above, LLPS, phase separation, and condensate formation are distinct terms representing distinct chemical and physical behaviors. Moreover, it is still quite challenging to specifically demonstrate phase separation *in vivo* and most descriptions are based on *in vitro* studies. Accordingly, the precise biophysical properties of most nuclear compartments are still unknown. To avoid ambiguity, we will use the more general term condensate except where the more specific biophysical properties of a compartment are well-characterized and critical for describing its functional role.

Importantly, condensate formation is not the only mechanism by which nuclear compartments can be formed. Alternative mechanisms include the ability for diffusible proteins to form stoichiometric molecular interactions with spatially anchored molecules such as ncRNA, DNA, and histones. For example, initiation of X chromosome inactivation (XCI) requires the recruitment of the SHARP RNA binding protein to the inactive X chromosome (Xi). This occurs through induction and retention of the Xist RNA on the Xi[67] followed by recruitment of SHARP through a stoichiometric interaction between its RRM domains and Xist[68]. Similarly, upon Herpes Simplex Virus infection, cells form a "replication compartment" in the nucleus that shows a strong enrichment of RNA Polymerase II. This enrichment is driven by preferential binding of Pol II to the viral DNA contained within this compartment[69].

1.4 NUCLEAR COMPARTMENTALIZATION FACILITATES TRANSCRIPTIONAL REGULATION

Biological processes as diverse as differentiation, response to environmental cues, and innate immunity all depend on the ability of a cell

to rapidly and specifically activate or repress transcription of specific genes[70–72]. Transcriptional regulation involves numerous cell-type specific transcription factors (TFs) that bind to DNA regulatory elements, including promoters and enhancers, through high affinity (stoichiometric) interactions with defined DNA sequences. The Mediator complex binds to enhancers, brings them near promoters, and facilitates Pol II loading during transcriptional initiation[73]. Despite important progress in deciphering the molecular mechanisms of transcriptional regulation, current models largely rely on stoichiometric molecular interactions and cannot fully explain many of the quantitative features of transcriptional regulation. Here we will discuss recent data that establishes an important role for nuclear compartmentalization in transcriptional initiation and describe how compartmentalization may enable many of the critical quantitative features of gene regulation.

1.4.1 ENHANCERS AND PROMOTERS FORM TRANSCRIPTIONAL CONDENSATES WITH RNA POL II AND MEDIATOR

The discovery of enhancers in animal genomes initially posed a challenge for classical genetics because it was not obvious how a DNA element tens or hundreds of kilobases away from a promoter could influence the activity of its target[74]. Early studies of Sonic hedgehog (*Shh*) using 3D imaging provided initial evidence that enhancers come into close physical proximity with their target genes[75]. This led to a prevailing model in which enhancer-promoter interactions form direct interactions through chromosome looping. Yet, this looping model cannot fully explain long-range enhancer regulation because an individual promoter can be simultaneously regulated by multiple enhancers[7] and, in specific cases, the spatial distance between enhancers and their promoters can increase upon transcriptional activation[76,77].

1.4.1.1 ENHANCERS AND THEIR TARGETS COEXIST IN TOPOLOGICALLY ASSOCIATING DOMAINS. Enhancer-promoter

interactions generally occur within chromosomal structures known as topologically associating domains (TADs)[8,78,79]. A TAD is a 3D structure consisting of large DNA regions (generally on the order of hundreds of kilobases of DNA) that interact more frequently with other DNA regions within the domain than with neighboring linear DNA regions (**Figure 3A**). TADs may represent the functional unit of promoter-enhancer interactions because adjusting the linear distance between *Shh* and its enhancer within a TAD has a modest effect on *Shh* expression, but disrupting the TAD boundaries substantially decreases *Shh* expression[7,80]. Many pathologies are thought to occur as a result of aberrant enhancer-promoter contacts due to loss of TAD boundaries. For example, TAD boundary disruption is thought to allow enhancers to aberrantly activate oncogenes in certain cancers[81,82]. Although some TADs are critical to facilitate specific enhancer-promoter interactions, there are also examples where they

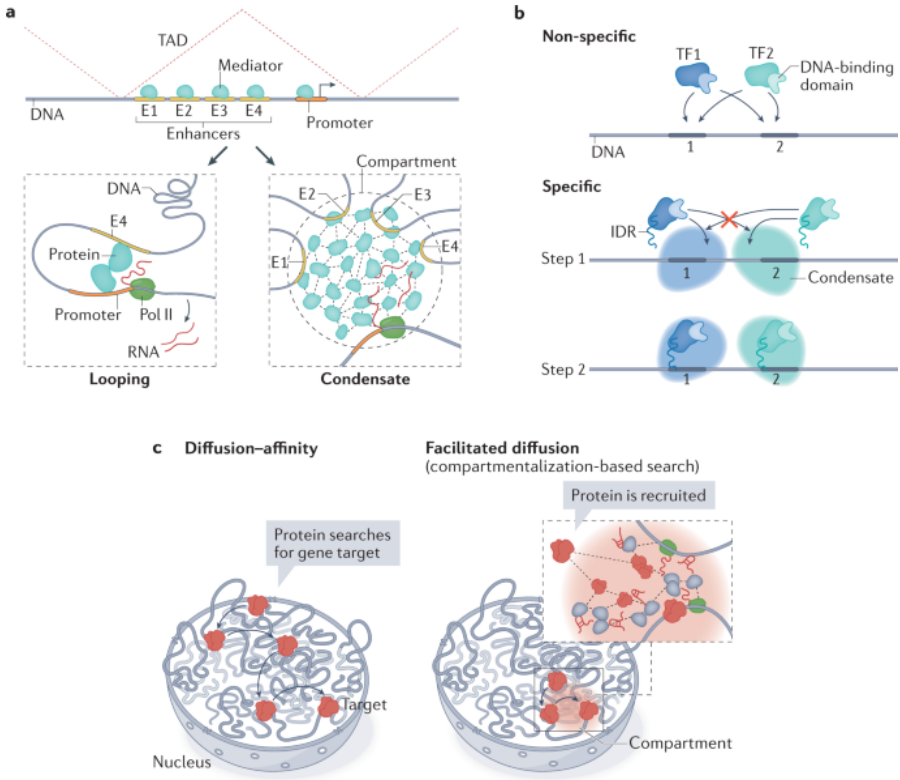


Figure 3 *| a | Example of multiple enhancers (super-enhancer) confined to a topologically associating domain (TAD; dotted red line) that contains the promoter they regulate. Comparison of a simple chromatin looping model of enhancers and promoters (left) with a condensate model (right) shows how multiple enhancers can occupy the same territory by forming cooperative interactions. b | Two distinct transcription factors (TF1 and TF2) can have the same affinity for the same DNA sequences, but occupy different genomic locations. Transcription factors lacking intrinsically disordered regions (IDRs) cannot distinguish between these high-affinity sites ('non-specific'), but transcription factors containing IDRs are targeted to specific genomic regions ('specific'). c | Comparison of diffusion–affinity models that involve stochastic movement of proteins to target sites (left) with the facilitated diffusion model, which involves a combination of 3D movement with sliding (right). Pol II, polymerase II. Image courtesy of Inna-Marie Strazhnik/Caltech. Part b adapted with permission from ref.11, AAAS.*

appear to be dispensable for gene regulation, indicating that not all TADs work in this way[83].

1.4.1.2 MEDIATOR AND POL II CONDENSATES FORM AT SUPER ENHANCERS. Many critical genes – such as pluripotency factors in embryonic stem cells, PU.1 in B cell progenitors, and *MYC* in multiple myeloma cells – contain large numbers of enhancers that control their expression[84–86]. These are referred to as "super-enhancers" and are defined as large regions of linear genomic DNA that contain a high concentration of Mediator as well as a large fraction of enhancer-associated Pol II within the cell[85].

Early imaging studies suggested that Pol II occupies clusters[87–91], but because this work relied on diffraction-limited microscopy, the precise organization of Pol II was difficult to assess. The advent of super-resolution microscopy dramatically improved the ability to directly visualize individual molecules of Pol II in the nucleus[92]. Super-resolution studies show that Pol II and Mediator co-occupy both large and small clusters in the nucleus[11]. The large Pol II/Mediator clusters (which make up <10% of total clusters) were proposed to correspond to previously described super-enhancer containing DNA loci[93]. Indeed, linking these Pol II/Mediator clusters with specific DNA loci (using immunofluorescence

and DNA FISH) demonstrated that these large clusters tend to overlap with super-enhancer containing loci[43].

1.4.1.3 TRANSCRIPTIONAL CONDENSATES CAN INCREASE THE CONCENTRATION OF REGULATORY PROTEINS ON DNA. The high concentration of Pol II/Mediator at these super-enhancer-containing genes suggests that transcriptional condensates act to increase the concentration of Pol II at highly regulated genes (**Figure 3A**). Indeed, live cell super-resolution imaging has shown that at specific loci the level of nascent transcription is directly correlated with the size of the Pol II condensate[94] (**Figure 3B**). These data suggest that super-enhancers may act to compartmentalize Mediator around target promoters and may facilitate Pol II loading. Similarly, transcription factors containing IDRs can also be recruited to genomic DNA regions containing a high density of binding sites to form a compartment that achieves higher TF concentrations than the number of binding sites present. For example, the EWS/FLI1 fusion TF can bind to repetitive DNA regions and can accumulate within clusters in the nucleus at concentrations that exceed the number of available DNA binding sites. In this way, this fusion protein can promote robust expression of genes at these sites[95].

The ability for Mediator, Pol II, and specific TFs to undergo concentration-dependent condensate formation may increase the concentration of these regulators beyond the stoichiometric concentration that can be achieved by binding to individual DNA binding sites.

1.4.2 CONDENSATES OF DNA BINDING PROTEINS MAY FACILITATE RAPID TARGET SEARCH IN THE GENOME

Cell-type specific transcription factors diffuse through the nucleus and form high affinity interactions with their cognate DNA binding sequence. While this “diffusion and affinity model” can explain many of the qualitative aspects of gene regulation (i.e., where TFs bind to DNA), measurements in bacteria demonstrate that the observed rates of

association of the Lac repressor to its target DNA site are dramatically faster than would be predicted by molecular diffusion and DNA binding affinity[96]. This discrepancy is even larger in eukaryotes, where the number of available sites on chromatin is vastly more complex[1]. For example, it would take several hundred hours for a specific TF to identify a single binding site within the nucleus using diffusion and affinity alone[2]. Yet, TFs can dynamically localize and induce transcription within minutes of stimulation in many distinct contexts. In addition, not all high affinity binding sites on DNA are occupied by a TF, indicating that there are factors beyond DNA binding affinity that are important for controlling target recognition.

To address this quantitative challenge, a biophysical model referred to as "facilitated diffusion" was proposed[1,96]. This model suggests that TFs identify their targets in two steps – (i) rapid 3D diffusion to identify genomic neighborhoods and (ii) slow 1D scanning within a neighborhood to identify high affinity targets[97,98]. Recent studies suggest that the ability for TFs – and other DNA binding proteins – to form condensates might be a critical component of this facilitated search[2].

Specifically, the IDRs contained within TFs are often critical for specific genomic DNA localization[99]. For example, two distinct TFs in yeast (Msn2 and Yap1) both have DNA binding domains with affinity for the same DNA sequences but occupy different genomic locations. This genomic specificity is encoded not by the DNA binding domains of these proteins, but by their IDRs (**Figure 3C**). Indeed, upon deletion of their IDRs, Msn2 and Yap1 can still bind to DNA, but no longer achieve specificity to their correct genomic DNA targets. Instead, both mutant TFs bind equally well to both binding sites. This example demonstrates that specificity of TF localization can be achieved through the same unstructured regions required for condensate formation.

An example where condensate formation is required to define the specificity of a DNA binding protein is in the context of dosage compensation in *Drosophila*[100]. This process entails recruitment of an

activating complex (Male Sex Lethal, *MSL*) exclusively to the promoters of genes on the male X chromosome, even though this DNA binding complex also contains high affinity DNA binding sites on autosomes. How *MSL* achieves specific localization at DNA regions on the X was a long-standing question. A recent study demonstrated that the IDR within MSL2 is required for X chromosome-specific targeting of this complex[100]. Importantly, specific localization on the X is mediated by an interaction with the roX ncRNA (which is transcribed from the X and is enriched on the X) that acts to seed a concentration-dependent condensate through homotypic interactions between the IDRs of *MSL2* on the X.

Together, these results indicate that DNA binding proteins can use their IDRs to achieve localization specificity by creating high local concentration in specific nuclear regions. In this way, compartmentalization may reduce the range of possible targets in the nucleus thereby accelerating target search (**Figure 3D**). In addition to the specific examples described above, many DNA binding proteins contain IDRs that can form condensates in the nucleus[12,45] suggesting that this may be a more general mechanism by which DNA binding proteins achieve specificity.

1.5 NUCLEAR COMPARTMENTALIZATION AND HIGHER-ORDER CHROMATIN REGULATION

Genomic DNA within the nucleus is spatially organized into active and inactive territories[4]. Post-translational modifications of histone proteins (chromatin modifications) can change the accessibility and spatial organization of large stretches of genomic DNA rather than of individual genes. These chromatin landscapes can form higher-order 3D structures that share transcriptional states. For example, large, actively transcribed regions form higher-order interactions with other active regions while heterochromatin and other repressed DNA form 3D structures that can contain DNA from multiple distinct chromosomes[10,101–103]. Here we

discuss emerging principles of compartmentalization in higher-order DNA packaging and coordinated regulation across multiple genes.

1.5.1 CONDENSATE FORMATION ENABLES AMPLIFICATION OF REPRESSIVE CHROMATIN PROTEINS ON THE INACTIVE X

One example of chromatin-mediated multi-gene regulation is X chromosome inactivation (XCI). XCI is a critical developmental process by which one of the two X chromosomes in female mammals is silenced to ensure dosage balance in X-linked gene expression between males and females[104]. XCI has emerged as a paradigm for understanding the relationship between nuclear organization, epigenetics, and gene regulation[104,105].

The Xist lncRNA initiates a cascade of events that leads to XCI[106–108]. Xist binds to the SHARP/SMRT/HDAC3 complex[108–110] to evict RNA Polymerase II from the X chromosome[67]. Although aggregate measurements across a population of cells show that Xist localizes across the entire X chromosome[67], its expression levels are not high enough for it to occupy the entire Xi within an individual cell. Specifically, there are ~200 copies of Xist in each cell[59] that would need to bind to >167 million base pairs (~1 copy of Xist/1 million base pairs of DNA). Accordingly, a direct stoichiometric interaction alone cannot explain how the Xist-SHARP complex leads to deterministic silencing of the entire X chromosome during mammalian development.

1.5.1.1 SPATIAL AMPLIFICATION OF REGULATORY PROTEINS. Recently, a super-resolution microscopy study showed that the concentration of SHARP on the Xi exceeds the concentration of Xist[59] suggesting that SHARP is recruited to the Xi in a super-stoichiometric manner. Xist can form a direct, high-affinity stoichiometric interaction with SHARP through its RRM domains and this interaction is required to recruit it to the Xi. SHARP also contains a large IDR that is dispensable

for binding to Xist and SMRT/HDAC3[68] but is essential for super-stoichiometric recruitment of SHARP to the Xi[59].

One mechanism to achieve such super-stoichiometric recruitment in an IDR-dependent manner is through the establishment of a Xist-mediated condensate that leads to high spatial concentration of the silencing complex. Specifically, once SHARP achieves sufficiently high spatial concentration on the Xi via a stoichiometric interaction with Xist, it can undergo a concentration-dependent transition to enable multiple SHARP molecules to form multivalent interactions with each other, independently of direct binding to Xist. In this way, condensate formation may act to amplify the number of repressive SHARP proteins on the Xi relative to Xist molecules and may enable transcriptional silencing independently of a direct molecular interaction with Xist **(Figure 4A**).

1.5.1.2 CONCENTRATION THRESHOLDING IN GENE EXPRESSION. This principle of local concentration of RNA and protein determining the regulatory state of a gene may also explain why transcriptional silencing is restricted to a single X chromosome during XCI. Although Xist is predominantly retained on the X chromosome from which it is expressed, it can also diffuse to autosomal genes and accumulate at a lower concentration[47,67]. Despite its occasional presence on autosomes, Xist does not appear to repress autosomal gene expression. However, when the entire X-inactivation center is moved to an autosome Xist is capable of silencing that chromosome[111]. This indicates that autosomes are not intrinsically resistant to Xist repression, but rather that the concentration of Xist required to repress transcription is not achieved on autosomes in wild-type cells. These long-observed phenotypes may be explained by a requirement for a critical concentration of Xist on a chromosomal region in order to recruit a silencing-competent SHARP/HDAC3 complexes. In this way, it can achieve deterministic silencing on the X chromosome (where it is present at its highest levels) but avoid silencing autosomal regions (where Xist never reaches this critical threshold).

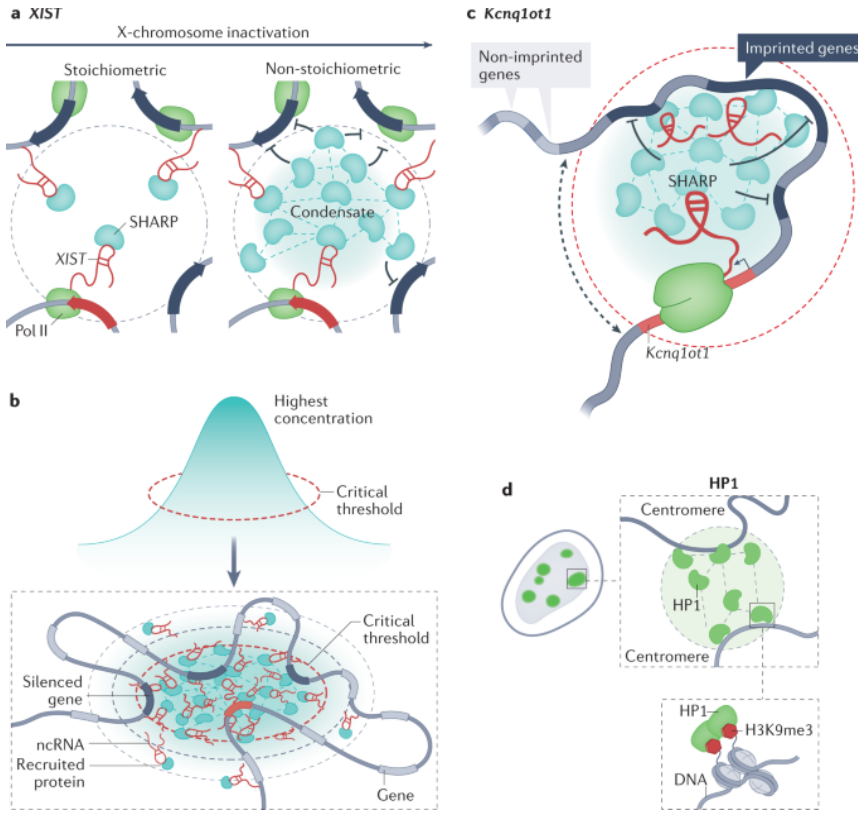


Figure 4 *a | Stoichiometric interactions occur at defined ratios of components, whereas non-stoichiometric interactions occur without fixed ratios and can exceed the binding capacity of any individual molecule. Increasing the concentration of the long non-coding RNA (ncRNA) X-inactive specific transcript (XIST) along the X chromosome increases in the concentration of the XIST-binding protein SHARP, which eventually results is silencing of most genes on the chromosome. Spatial amplification of SHARP (which exceeds XIST concentration) requires its intrinsically disordered regions (IDRs) and may be achieved through concentration-dependent condensate formation. b | Concentration thresholds may explain the 'switch-like' control of gene repression within and outside ncRNA compartments. ncRNAs can achieve their highest concentration at the centre of a spatial territory. If the ncRNA concentration at the central area reaches a critical threshold, enough proteins to mediate compartment formation and to achieve gene repression can be recruited. In this model, ncRNA molecules can still diffuse outside the silencing compartment, but not at levels high enough to recruit sufficient regulatory molecules to change gene expression. As a result, the observed effect on gene expression is binary: 'on' within the compartment and 'off' outside it. c | Schematic of the Kcnq1ot1 compartment,*

which also recruits SHARP to silence imprinted genes on the paternal allele. Dashed arrow represents a cohesin complex promoting the formation of the Kcnq1ot1 compartment. d | Schematic showing heterochromatin protein 1 (HP1)-mediated compartmentalization of centromeric heterochromatin, which is marked by trimethylated histone H3 Lys9 (H3K9me3). Pol II, polymerase II. Image courtesy of Inna-Marie Strazhnik/Caltech.

Compartmentalization may help establish physical boundaries for biological processes that require high concentrations of effectors. For example, the Kcnq1ot1 lncRNA diffuses continuously from its locus and exhibits volumetric decay with the highest concentration around its own locus and decreased concentration at regions farther away[56] (**Figure 4B**). Similar to Xist, the Kcnq1ot1 lncRNA also binds directly to SHARP[56] and represses transcription through the HDAC3 deacetylase complex[56,112] (**Figure 4B**). However, despite its diffusive properties, the repressive effect of Kcnq1ot1 is confined to a topological domain where it is present at the highest concentration and its target imprinted genes are contained[56,113,114]. Rather than displaying continuous silencing effects – with genes proportionally repressed relative to the amount of RNA observed – the effects for Kcnq1ot1 appear to be more deterministic and switch-like. This might be explained if a minimum concentration of the Kcnq1ot1-SHARP complex is required to induce concentration-dependent phase separation to silence transcription (**Figure 4C**).

1.5.2 NUCLEAR COMPARTMENTALIZATION AND EPIGENETIC MEMORY

Chromatin regulation is often characterized by long-lasting regulatory effects. For example, the inactive X chromosome remains silenced after initiation of XCI, even upon loss of the Xist RNA[111]. Similarly, chromatin regulators like HP1 that mediate heterochromatin tend to be stable across many cell divisions[115,116]. Here we discuss emerging evidence that spatial organization can play a critical role in establishing "epigenetic memory" in specific cases.

1.5.2.1 XIST ESTABLISHES A NUCLEAR CONDENSATE THAT IS REQUIRED FOR PERSISTENT AND STABLE SILENCING. After a critical development timepoint, Xist induces stable and epigenetically heritable chromosome-wide transcriptional silencing on the Xi that is maintained even in the absence of Xist[111]. Xist expression leads to the recruitment of numerous chromatin regulatory proteins, including the Polycomb Repressive Complex[108,117,118], as well as DNA methylation[119] (methylation of cytosines within CpG dinucleotides). In addition to chromatin protein recruitment, the Xi undergoes large scale structural changes and several of the proteins that are recruited to the Xi have been shown to undergo concentration-dependent condensate formation[104,120].

Recent evidence suggests that compartmentalization might play a critical role in the maintenance of XCI. Specifically, the PTBP1 RNA binding protein interacts with the E-repeat region of Xist, a tandem repeat containing more than 50 PTBP1 recognition sites[47]. PTBP1 has largely been studied because of its role as an mRNA splicing factor and was one of the first nuclear proteins shown to undergo concentration-dependent condensate formation[35]. Consistent with this observation, PTBP1 forms a condensate structure on the X chromosome that is dependent on its interaction with Xist. Intriguingly, genetic deletion of PTBP1 (or its binding sites) does not impact the initiation of transcriptional silencing on the X chromosome (in contrast to deletion of critical initiation proteins such as SHARP)[47,108]. However, when analyzing transcriptional silencing across time, the frequency at which PTBP1 mutant cells "reactivate" expression from the Xi is dramatically higher than in wild-type cells[47]. This suggests that PTBP1 might be important for maintaining silencing on the Xi. In this model, once initiation of silencing is complete and forms a condensate, this silenced state might be maintained through self-interactions between proteins within the condensate and therefore no longer require Xist expression.

1.5.2.2 CHROMATIN REGULATORS MAY SEED RE-ESTABLISHMENT OF NUCLEAR CONDENSATES. In order for this Xist-mediated condensate to play a role in the maintenance of silencing, it must be

propagated across cell divisions. While it remains unknown how this might occur, various histone modifications and DNA methylation are stably associated with the Xi through cell division[105,118,121]. These modifications may seed re-establishment of the silencing compartment following mitosis. For example, the H3K27me3 repressive histone mark is known to recruit the PRC1 complex[122–124], which has been shown to form concentration-dependent condensates. Specifically, the CBX2 protein (a component of PRC1) contains a chromodomain that binds with high affinity to modified histones, an IDR that can form condensates, and other domains required for binding and recruiting other critical PRC1/2 regulatory proteins that reinforce silencing[41]. In this way, PRC1 may be able to re-establish a nuclear compartment following cell division.

Beyond the Xi, polycomb proteins can form high concentration territories in the nucleus ("polycomb bodies"). These territories contain DNA regions that are located at distinct linear locations on a chromosome or across multiple chromosomes[125]. For example, two Hox gene clusters in Drosophila, ANT-C and BX-C, co-localize within polycomb bodies despite being separated by 10 Mb[126]. Recently, the CBX2 protein was shown to undergo condensate formation and is required for formation of these bodies[41,127]. In addition, other polycomb complex proteins are contained within these condensates. Interestingly, the IDR domain of CBX2 that is required for condensate formation is the same domain that is required for PRC1-mediated chromatin compaction, suggesting that chromatin compaction might be a product of condensate formation[41].

Other chromatin proteins can also form condensates in the nucleus following similar principles. Constitutive heterochromatin is an important feature of the eukaryotic nucleus and is associated with the heterochromatic H3K9 trimethylation histone mark, which is recognized by HP1[128]. Similar to CBX2, several of the HP1 homologs contain a chromodomain (required for recognizing H3K9me3) and IDRs and can form compartmentalized structures in the nucleus[13]. For example, HP1 forms high-concentration territories with distinct genomic DNA regions, including centromeric DNA regions from multiple chromosomes[55,56]

(**Figure 4D**). In this way, spatial organization of chromatin proteins and their regulatory targets might act to increase their concentration and re-initiate transcriptional silencing across cell division.

1.6 NUCLEAR BODIES AND POST-TRANSCRIPTIONAL RNA PROCESSING

Gene regulation involves both transcriptional modulation and various forms of post-transcriptional RNA processing such as base modification, polyadenylation, and splicing. Numerous RNA processing steps are known to occur co-transcriptionally. For example, the majority of mRNA splicing in mammals occurs co-transcriptionally, such that the nascent pre-mRNA is spliced as it is being transcribed by RNA Polymerase II[129]. Because mRNA processing is critical for cellular function as it ensures that the intact message will be translated into its functional protein product, co-transcriptional mRNA processing is thought to increase the efficiency of these processes[130,131].

Pol II has a unique, highly conserved, C-terminal domain (CTD) containing a 52 heptad repeat structure that undergoes extensive post-translational modifications during distinct stages of transcription[132,133]. These modifications act as dynamic protein binding sites for distinct mRNA processing factors. As an example, phosphorylation of serine 2 (ser2P) is associated with transcriptional elongation and has been reported to bind several proteins required for mRNA splicing and 3' cleavage and polyadenylation[134]. Deletion of the CTD of Pol II leads to a defect in the rate of co-transcriptional splicing of pre-mRNAs relative to its impact on transcription, suggesting that the CTD is important for coordination of transcription and splicing[15].

Based on these observations, the traditional view of co-transcriptional splicing is that various proteins form a complex with Pol II which brings the splicing machinery into proximity with a pre-mRNA immediately upon transcription[135–137] (**Figure 5A**). However, this direct interaction

model cannot fully explain many of the observed properties of this system. Specifically, although multiple mRNA splicing and processing factors have been reported to bind to the CTD[138], a single Pol II molecule is not big enough to accommodate a direct interaction with all of these proteins simultaneously[139]. Yet, splicing and processing factors are present at each nascent transcript. Furthermore, the relationship between transcription and splicing kinetics is non-linear whereby genes that are more highly transcribed exhibit a non-linear increase in their splicing rate[140] (**Figure 5B**). Here we will discuss emerging evidence relating spatial organization in the nucleus and kinetic coupling of co-transcriptional RNA processing.

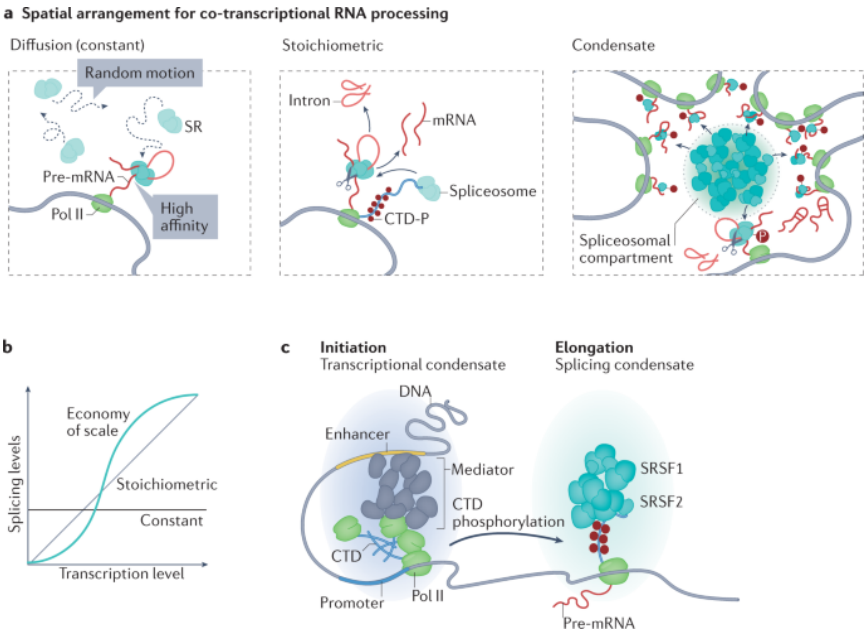


Figure 5 a *| Different models of co-transcriptional splicing. Random diffusion of splicing factors (for example, SR proteins) that have high affinity for target pre-mRNAs (left). In this model, splicing levels are independent of transcription and constant (part* ***b****). Stoichiometric model of co-transcriptional splicing, showing molecular interactions between the phosphorylated carboxy-terminal domain (CTD-P) 'tail' of RNA polymerase II (Pol II) and the spliceosome, which facilitates splicing as nascent pre-mRNAs are extruded from the polymerase (middle). In this model, splicing increases linearly with transcription rate (part* ***b****). Condensate model of co-transcriptional splicing, showing*

cooperative interactions between the CTD of Pol II, phosphorylated (P) splicing factors and a reservoir of unphosphorylated splicing factors in close proximity to nascent pre-mRNAs (right). ***b*** *| A condensate model of co-transcriptional splicing would lead to an 'economy of scale' dynamic, in which splicing increases non-linearly with increased transcription.* ***c*** *| Pol II CTD phosphorylation mediates a switch from transcriptional condensates of the Mediator complex during transcription initiation, to condensates of splicing factors during transcription elongation. SRSF1, serine/arginine-rich splicing factor 1. Image courtesy of Inna-Marie Strazhnik/Caltech.*

1.6.1 SPATIAL AND KINETIC COUPLING OF RNA POLYMERASE II TRANSCRIPTION AND MRNA SPLICING

1.6.1.1 CTD AND SPLICING FACTORS FORM NUCLEAR CONDENSATES. Pol II and various mRNA splicing and processing factors have been shown to form nuclear condensates[11,43,141]. These splicing condensates are formed through low affinity, heterotypic interactions between the IDRs contained within various splicing proteins and the CTD of Pol II (itself an IDR). In contrast to Mediator condensates that form with the hypo-phosphorylated CTD of Pol II during transcriptional initiation, condensates of splicing factors have been shown to form specifically with the hyper-phosphorylated CTD of Pol II, which is associated with transcriptional elongation[43]. As such, phosphorylation of the CTD of Pol II enables a transition from initiation to elongation and from the initiation condensate (containing enhancers and Mediator) to an elongation condensate (containing pre-mRNAs and splicing and processing factors). This elongation condensate would enable association of a large number of splicing and processing factors with the elongating RNA Pol II simultaneously (**Figure 5C**).

1.6.1.2 ACTIVELY TRANSCRIBED GENES ARE ORGANIZED ON THE PERIPHERY OF NUCLEAR SPECKLES. Recent studies showed that the genomic DNA regions of highly-transcribed Pol II genes – and their corresponding nascent pre-mRNAs – are spatially organized around nuclear speckles[10,56,142]. A nuclear speckle is a nuclear body that contains numerous splicing and processing proteins[11]. Specifically, the inner structural core is composed of serine/arginine-rich mRNA splicing factors

(i.e., SRSF1, SC35/SRSF2) and other mRNA processing factors (e.g., CPSF, CSTF, CFIm)[143,144] while the periphery consists of genomic DNA and nascent pre-mRNAs[145]. Interestingly, the distance between a gene and the nuclear speckle is inversely correlated with its transcription level[140] and genes that are within highly-transcribed neighborhoods are preferentially proximal to speckles[10,142]. Although proximity to the nuclear speckle is associated with increased gene expression[10,146] genomic DNA proximity to a nuclear speckle also includes intergenic regions and inactive genes that are contained within highly transcribed Pol II neighborhoods[10]. These observations suggest a model whereby the induction of transcription by Pol II leads to dynamic reorganization of DNA into 3D proximity of nuclear speckles.

1.6.1.3 SPATIAL ORGANIZATION MAY COORDINATE CO-TRANSCRIPTIONAL PROCESSING. While splicing does not appear to occur within nuclear speckles, the splicing factors that are contained within them have been shown to diffuse from the speckle to the nascent pre-mRNA[147]. A recent study showed that genes that show increased splicing efficiency are also closer to a nuclear speckle[140]. Compartmentalization of regulatory components around speckles increases the spatial concentration of splicing proteins near nascent pre-mRNAs and in this way may increase the rate of co-transcriptional splicing. This would lead to a non-linear relationship between Pol II concentration (and transcription rate) and splicing efficiency because the localization of a gene closer to a speckle would increase the concentration of spliceosomal components non-linearly relative to its pre-mRNA targets. Because rates of a biochemical reaction increase with increased enzyme or substrate concentrations, concentrating the relevant molecules within a specific structure would increase the kinetic rate of the specific reaction without changing the overall concentration of the molecules within the cell. This could explain why splicing efficiency behaves in an “economy of scale” manner[140], increasing non-linearly as transcription levels increase.

1.6.2 SEVERAL NUCLEAR BODIES ARE ORGANIZED AROUND TRANSCRIPTION AND PROCESSING OF SPECIFIC RNAS.

This relationship between spatial organization of genomic DNA regions, their associated nascent RNAs, and the diffusible regulators required for their co-transcriptional processing is shared across many different classes of RNA processing, suggesting that such a role might be a general feature of co-transcriptional RNA processing (**Figure 6A**). In addition to nuclear speckles and pre-mRNAs described above, the nucleus contains several nuclear bodies that are organized around transcription and processing of specialized RNA molecules[56,145,148–151].

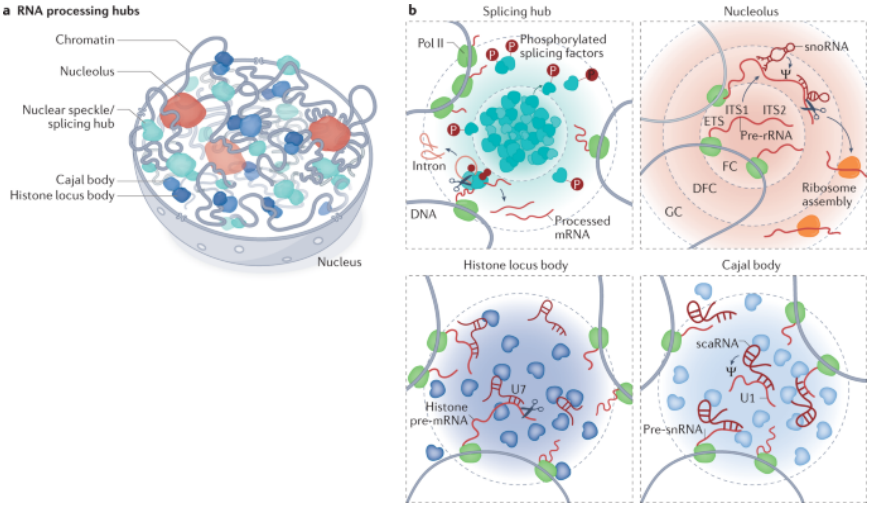


Figure 6 *a | Schematic of a cell nucleus highlighting many RNA processing hubs, including nuclear speckles (compartments of mRNA splicing factors), nucleoli (ribosome biogenesis), histone locus bodies (processing histone pre-mRNAs) and Cajal bodies (biogenesis of small nuclear RNAs (snRNAs)). b | Close-up schematics of different RNA processing hubs. A splicing hub is generally composed of two layers: an inner layer of splicing factors and an outer layer of genomic regions with active RNA polymerase II (Pol II), phosphorylated (P) splicing factors and nascent pre-mRNAs. The nucleolus is composed of three layers: inner fibrillar centre (FC), which is the site of Pol I-mediated pre-rRNA transcription; middle dense fibrillar component (DFC), where rRNA processing occurs by small nucleolar RNAs (snoRNAs) and RNase MRP (scissors); and outer granular component (GC), which is the site of ribosome assembly. Histone locus bodies are sites of*

histone pre-mRNA maturation, including by the U7 snRNA (U7). Cajal bodies are sites of snRNA maturation, including by small Cajal body RNAs (scaRNAs). Ψ, pseudouridylation. Image courtesy of Inna-Marie Strazhnik/Caltech.

1.6.2.1 NUCLEOLUS: A NUCLEAR BODY ORGANIZED AROUND TRANSCRIPTION AND PROCESSING OF RIBOSOMAL RNA. The nucleolus is the site of ribosomal biogenesis and consists of ribosomal DNA, ncRNAs including snoRNAs, and proteins involved in rRNA processing. Importantly, nascent transcription of 45S pre-rRNA is essential for formation and maintenance of the nucleolus[50,51,152,153] and direct (stoichiometric) binding of diffusible factors to the 45S pre-rRNA is required for their recruitment into this nuclear body.

The nucleolus organizes rRNA biogenesis in three liquid-like phases that correspond to the molecular components required for rRNA transcription (fibrillar center), processing (dense fibrillar component), and ribosome assembly (granular component)[39]. Several ncRNAs and proteins important for rRNA biogenesis are contained in the two inner layers, including snoRNAs (which bind to and modify pre-rRNAs) and RNase MRP (which cleaves pre-rRNAs into mature rRNAs)[154–156] (**Figure 6B**). In this way, compartmentalization may help physically separate steps of RNA biogenesis, ensuring that one step is completed before the intermediate is shuttled to the next compartment. A similar mechanism might be used to ensure completion of splicing prior to diffusion and export to the cytoplasm as well as for other forms of co-transcriptional processing[145].

1.6.2.2 HISTONE LOCUS BODIES: A NUCLEAR BODY ORGANIZED AROUND TRANSCRIPTION AND PROCESSING OF HISTONE MRNAS. Histone locus bodies (HLBs) are nuclear compartments that contain multiple histone DNA loci, U7 snRNAs and various regulatory proteins (NPAT, FLASH, NELF), and are the location of histone mRNA biogenesis[53]. Unlike other pre-mRNAs, histone genes do not contain introns and are not polyadenylated[151]. Instead, the U7 snRNP binds to the 3' end of histone mRNAs and is required for cleavage of histone pre-mRNAs before they are exported as mature transcripts to the

cytoplasm[151,157–159]. U7 binding to histone mRNAs is essential for the formation of HLBs[150]. This result suggests that histone pre-mRNA promotes the formation of HLBs and subsequently recruits other ncRNAs and protein factors within this specialized processing compartment (**Figure 6B**).

1.6.2.3 CAJAL BODIES: A NUCLEAR BODY ORGANIZED AROUND TRANSCRIPTION AND PROCESSING OF SNRNAS. Cajal bodies (CBs) are the sites of snRNA biogenesis and contain snRNA genes (U1, U2, U4, U5, and U6), scaRNAs, and proteins such as coilin and SMN[56,160]. Similar to snoRNAs in the nucleolus, scaRNAs bind directly to pre-snRNAs and mediate methylation and pseudouridylation within the CB[161,162]. Similar to RNA-processing bodies, concentrating regulatory components (scaRNAs) and targets (snRNAs) may increase the efficiency of snRNA processing. This example also highlights another potentially critical role of spatial compartmentalization in RNA processing: compartmentalization can increase the rate by which regulators identify and engage targets, which may be particularly important in cases where the regulators (e.g., scaRNAs) are expressed at low levels relative to their more abundant substrates (e.g., snRNAs) (**Figure 6B**).

1.7 GLOSSARY

Macromolecular complex – a group of molecules that assemble in fixed conformations and defined stoichiometric ratios (e.g., an enzyme and its cofactors, a ribonucleoprotein, or a holoenzyme). Complexes tend to be defined by highly specific, high-affinity interactions.

Biomolecular condensate – a concentration-dependent assembly of molecules of variable stoichiometry usually driven by multivalent and cooperative interactions that can form with or without phase separation.

Compartment – a spatial territory that is enriched for specific DNA, RNA, and/or protein molecules.

Intrinsically disordered region (IDR) – regions within a protein that do not have a single preferred structural conformation.

Phase-separation – thermodynamically driven partitioning of a homogeneous mixture into locally distinct chemical sub-mixtures (phases) with distinct density, composition, and/or other physical properties.

Liquid-Liquid Phase separation (LLPS) – a specific form of phase-separation defined by the formation of a liquid compartment within a larger liquid environment.

Phase transition – the biophysical process that leads to phase-separation. In terms of thermodynamics, it refers to a discontinuous change in the thermodynamic equilibrium state of a system in response to a change in a parameter such as temperature, pressure, or molecular concentration.

Multivalent – molecular associations that occur between multiple molecular binding sites on a single molecule and can lead to variable stoichiometric ratios.

Valency – the number of non-covalent interactions with other molecules that a single molecule or domain can support.

Homotypic interactions – interactions occurring between 2 or more copies of the same type of molecule.

Heterotypic interactions – interactions occurring between at least 2 molecules of different types.

Affinity – the strength of a non-covalent biochemical interaction, defined at the ratio of the association and dissociation rates.

Avidity – the collective strength of multiple non-covalent molecular interactions. Avidity represents the overall force conferred by multiple affinities in concert, which exceeds the sum of the strength of those interactions. It is commonly referred to as "functional affinity."

Diffusible versus constrained molecules – diffusible molecules proceed on a random walk throughout the volume that contains them. Constrained molecules proceed on a random walk preferentially within a subvolume of their overall environment, often due to high-affinity to molecules localized within that subvolume.

Diffusion and Affinity Model – the model describing how a molecule (such as a transcription factor) proceeding on a random walk through the nucleus samples many possible binding partners until it finds its high affinity cognate target site.

Seeding – the process of initiating the formation of compartment by a relatively immobile molecule or a group of molecules. This process is also referred to as nucleation.

Spatially constrained molecule – refers to molecules whose mobility in the nucleus is constrained and that display sub-diffusive properties. Such molecules can be constrained intrinsically because they are present on chromatin (e.g., DNA, histone modifications, chromatin-bound ncRNAs or proteins) or because they form high-affinity interactions with molecules that are spatially constrained, such as through integration into the nuclear envelope.

Facilitated Diffusion – in the context of compartmentalization, the process by which compartments restrict the random walks of diffusible molecules to a smaller volume. For example, a condensate with affinity to a transcription factor might constrain a factor's diffusion to a small nuclear subvolume around its target.

Stoichiometric interaction – stoichiometric biochemical interactions occur with defined ratios of components, generally with high affinity and specificity. These include multiprotein complexes or the binding of a single transcription factor to its DNA motif in the major groove.

Super-stoichiometric interaction – interactions that occur without fixed ratios of components and can exceed the binding capacity of any individual molecule. Often, super-stoichiometric assemblies involve high avidity, multivalent interactions that increase the possible number of components that can compose the assembly.

Linear and non-linear responses – linear responses are changes in reaction rate or efficiency that are directly proportional to changes in the amounts of reactants or catalysts. Non-linear responses are changes in reaction rate or efficiency that exceed those expected under a linear model given the changes in amounts of reactants or catalysts.

Topologically Associating Domains (TADs) – linear stretches of DNA that interact more frequently with one another than with neighboring regions, often due to DNA loop formation and the presence of insulators.

Super-enhancer – a genomic region containing a large number of enhancers that control an individual target genes. Often, these targets are highly-regulated genes involved in cell identity.

Mediator – a protein complex that facilitates enhancer-promoter interactions, Pol II loading, and transcriptional initiation.

Super-resolution microscopy – microscopy methods such as structured illumination (STED and SIM) or localization techniques (PALM or

STORM) that enable imaging at resolutions that exceed the diffraction limit of light.

Epigenetic memory – refers to the set of DNA and chromatin modifications that are heritable either from parents to offspring or from a mother cell to its daughters.

Splicing condensates – high local concentrations of splicing factors localized around nascent RNA transcripts. Splicing condensates are often found in proximity to nuclear speckles but are distinct from the speckles themselves.

Transcription condensates – assemblies of general transcription factors and Mediator complexes that assemble around enhancers and promoters and facilitate transcriptional activation.

Granules – small (< 1μm) condensates that generally have a simple composition compared to the more complex class of nuclear bodies.

Nuclear bodies – large (≥ 1μm) compartments that form spatially and functionally distinct territories in the nucleus. Nuclear bodies are often involved in molecular biogenesis, such as the nucleolus in forming ribosomes and Cajal bodies in processing snRNAs.

Spatial enrichment – high-local concentration of a particular molecular species within a nuclear subvolume.

Nuclear territory – a catch-all term for a three-dimensional region contained within the nucleus.

Aggregates – liquid condensates or solid precipitates that are associated with pathogenesis. Examples include amyloid plaques in Alzheimer's and C9ORF72 poly(GA) foci in ALS.

1.8 BOX 1. METHODS FOR EXPLORING COMPARTMENTALIZATION

Several methodologies have been developed to define the molecular components and functional roles of nuclear compartments. We briefly describe some of these key methods below (reviewed in detail elsewhere[175,176]).

1.8.1 IN VITRO METHODS:

- **Droplet formation assays**[177] involve increasing the concentration of a recombinantly purified protein or RNA and determining at what overall concentration (if any) the molecules undergo a phase transition.
- **Scattering methods**[178], such as small-angle X-ray scattering (SAXS) determines the maximum diameter, approximate shape, and radius of gyration of particles in solution in order to infer the level of oligomerization between molecules. Another approach is dynamic light scattering (DLS), which detects fluctuations of scattered light to calculate the distribution of particle sizes in a solution.

1.8.2 MICROSCOPY:

- ***In situ* imaging of DNA, RNA, and Protein** involves the use of fluorescent probes complementary to target RNA or DNA and antibodies against specific proteins to visualize the position of specific RNA, DNA, and proteins.
- **Multiplexed imaging.** Adaptations of FISH, like seqFISH[173,179–182], merFISH[183], and oligoPaint[184–186], involve iterative hybridization of probes to thousands of RNA and DNA molecules to enable multiplexed visualization in 3D space. Recent studies using seqFISH+ enable multiplexed DNA, RNA, and protein measurements[187].
- **Live imaging** of the distribution of fluorescently labeled proteins can be used to observe the localization and dynamics of proteins in live cells.

- **Single molecule photobleaching** is a microscopy technique to quantify the stoichiometry of subunits in macromolecular complexes. Single units of fluorophores when photo bleached display a quantized intensity decay which can be converted to a precise number of fluorophores within a complex[188].

- **Nanocages** consisting of a defined number of molecules can be used as an internal standard to quantify concentration within a spatially territory[189].

- **Fluorescence correlation spectroscopy** uses live-cell microscopy to infer molecular concentration based on fluctuations of fluorescence intensity in a given location[190].

- **Fluorescence recovery after photobleaching** measures the extent of molecular exchange within a location by extinguishing signal in a location and measuring the rate at which labeled proteins recover within that region.
- **Super-resolution microscopy** techniques like **SIM**, **PALM** and **STORM** can visualize molecules beyond the diffraction limit of light (~250 nm), which is the limitation of conventional microscopy. These approaches have enabled detection of complexes (e.g., individual Pol II clusters) at finer resolution (~20-50nm)[191,192].
- **Quantitative phase microscopy (QPM)** is a label-free method that measures the shape and concentration of individual condensates by passing light through the sample using different refractive indices which causes an optical phase delay that can be quantitatively measured.

1.8.3 GENOMIC METHODS:

- **Proximity ligation methods** (e.g., **HiC)** enable genome-wide detection of pairwise DNA regions that interact in 3D[78,193].
- **SPRITE** and **GAM**[194] are ligation-independent methods that enables genome-wide detection of multiway DNA interactions.

- **RNA-DNA SPRITE** is an adaptation that provides genome-wide maps of RNA and DNA contacts in the nucleus[56].

  TSA-Seq utilizes diffusion of free radicals to map preferential association of specific regions around nuclear bodies[142].

1.8.4 MANIPULATING COMPARTMENTS

- **OptoDroplet** is an approach that enables proteins with IDRs to undergo phase transitions upon photo-activation enabling studies of various features of phase separated structures[170].
- **CasDrop** enables the formation of liquid condensates at specific genomic loci by recruiting a modified dCas9 that dimerizes upon photoactivation with IDR-containing proteins[171].
- **CRISPR-GO** is an inducible and reversible approach that uses ligand-mediated dimerization to direct DNA regions to specific nuclear compartments[169].
- **LADL** brings together two genomic sites via light-induced heterodimerization of cryptochrome 2 and a dCas9-CIBN fusion protein[168].

1.9 BOX 2. HOW COMMON IS "PHASE SEPARATION" AS A MECHANISM IN CELLS?

Phase separation is a physical process describing how molecules can separate from their environment. This concept has important implications for thinking about the formation of membraneless structures and as such has become an area of intense focus, leading to the description of an increasingly large number of structures that have been described as "phase separated" compartments. While this physical process is useful for thinking about how molecules can behave in cells, whether many of these compartments form via phase separation or other mechanisms of cooperative multivalent association is not clear in most contexts.

A recent perspective describes some of the issues with current experiments used and the challenges with defining whether a compartment forms via phase separation within a cell[32]. The primary issue is that most experiments used to study phase separation rely on *in vitro* experiments or analysis of tagged proteins that are overexpressed in cells. Given that phase separation is a concentration-dependent process, such systems may lead to behaviors that do not accurately reflect how these molecules assemble in their endogenous contexts within a cell. The authors argue that many of the properties often ascribed to liquid-like condensates (spherical shape, fusion and fission, and rapid diffusion) are not specific to liquid-like structures. In support of this perspective, they highlight an example of a compartmentalized structure formed upon infection with the Herpes Simplex Virus that displays these three key features despite the fact that it does not form via phase separation[69]. The authors propose two additional criteria for defining LLPS within a cell – (i) that the molecules are well mixed within the compartment and (ii) demonstrating that there is a change in the diffusion rate for molecules across the boundary of the compartment. These two features are often challenging to measure with current methods, especially for small compartments (e.g., transcriptional condensates). Given these technical challenges, the precise biophysical properties of most nuclear compartments *in vivo* remain largely uncharacterized.

As more mechanistic details are established about the various biophysical properties of nuclear compartments and their mechanisms of formation, we expect that the range of terminology will expand dramatically in order to capture the true range of these processes and their importance in biological regulation. Despite the imperfect nature of our current terminology, these different considerations do not directly impact the central defining feature of a nuclear compartment, namely the enrichment of specific molecules within a spatial territory in the nucleus.

1.10 REFERENCES

1. Mirny, L. *et al.* How a protein searches for its site on DNA: The mechanism of facilitated diffusion. *J. Phys. A Math. Theor.* (2009) doi:10.1088/1751-8113/42/43/434013.
2. Jana, T., Brodsky, S. & Barkai, N. Speed-Specificity Trade-Offs in the Transcription Factors Search for Their Genomic Binding Sites. *Trends Genet.* (2020).
3. Strom, A. R. & Brangwynne, C. P. The liquid nucleome – phase transitions in the nucleus at a glance. *J. Cell Sci.* (2019) doi:10.1242/jcs.235093.
4. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).
5. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 772–772 (2016).
6. Gibcus, J. H. & Dekker, J. The Hierarchy of the 3D Genome. *Mol. Cell* **49**, 773–782 (2013).
7. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics* (2019) doi:10.1038/s41576-019-0128-0.
8. Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
9. Dundr, M. & Misteli, T. Biogenesis of nuclear bodies. *Cold Spring Harbor perspectives in biology* vol. 2 (2010).
10. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).
11. Cho, W. K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412–415 (2018).
12. Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855.e16 (2018).
13. Strom, A. R. *et al.* Phase separation drives heterochromatin

domain formation. *Nature* (2017) doi:10.1038/nature22989.
14. Larson, A. G. *et al.* Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature* (2017) doi:10.1038/nature22822.
15. Huo, X. *et al.* The Nuclear Matrix Protein SAFB Cooperates with Major Satellite RNAs to Stabilize Heterochromatin Architecture Partially through Phase Separation. *Mol. Cell* **77**, 368-383.e7 (2020).
16. Murray, D. T. *et al.* Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. *Cell* (2017) doi:10.1016/j.cell.2017.08.048.
17. Sabari, B. R., Dall'Agnese, A. & Young, R. A. Biomolecular Condensates in the Nucleus. *Trends in Biochemical Sciences* (2020) doi:10.1016/j.tibs.2020.06.007.
18. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-Liquid Phase Separation in Biology. *Annu. Rev. Cell Dev. Biol.* (2014) doi:10.1146/annurev-cellbio-100913-013325.
19. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology* (2017) doi:10.1038/nrm.2017.7.
20. Lyon, A. S., Peeples, W. B. & Rosen, M. K. A framework for understanding the functions of biomolecular condensates across scales. *Nat. Rev. Mol. Cell Biol.* **22**, 215–235 (2021).
21. Zhang, Z. *et al.* Rapid dynamics of general transcription factor TFIIB binding during preinitiation complex assembly revealed by single-molecule analysis. *Genes Dev.* (2016) doi:10.1101/gad.285395.116.
22. Struhl, K. Helix-turn-helix, zinc-finger, and leucine-zipper motifs for eukaryotic transcriptional regulatory proteins. *Trends in Biochemical Sciences* (1989) doi:10.1016/0968-0004(89)90145-X.
23. Maris, C., Dominguez, C. & Allain, F. H. T. The RNA recognition motif, a plastic RNA-binding platform to regulate

post-transcriptional gene expression. *FEBS Journal* (2005) doi:10.1111/j.1742-4658.2005.04653.x.

24. García-Mayoral, M. F. *et al.* The Structure of the C-Terminal KH Domains of KSRP Reveals a Noncanonical Motif Important for mRNA Degradation. *Structure* (2007) doi:10.1016/j.str.2007.03.006.
25. Stevens, S. W. *et al.* Biochemical and genetic analyses of the U5, U6, and U4/U6•U5 small nuclear ribonucleoproteins from Saccharomyces cerevisiae. *RNA* (2001) doi:10.1017/S1355838201019070.
26. Derewenda, Z. S. Application of protein engineering to enhance crystallizability and improve crystal properties. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2010) doi:10.1107/S090744491000644X.
27. Derewenda, Z. S. Rational protein crystallization by mutational surface engineering. *Structure* (2004) doi:10.1016/j.str.2004.03.008.
28. Deller, M. C., Kong, L. & Rupp, B. Protein stability: A crystallographer's perspective. *Acta Crystallographica Section:F Structural Biology Communications* (2016) doi:10.1107/S2053230X15024619.
29. Boeynaems, S. *et al.* Protein Phase Separation: A New Phase in Cell Biology. *Trends in Cell Biology* (2018) doi:10.1016/j.tcb.2018.02.004.
30. Brangwynne, C. P., Tompa, P. & Pappu, R. V. Polymer physics of intracellular phase transitions. *Nat. Phys.* (2015) doi:10.1038/nphys3532.
31. Kitov, P. I. & Bundle, D. R. On the Nature of the Multivalency Effect: A Thermodynamic Model. *J. Am. Chem. Soc.* **125**, 16271–16284 (2003).
32. McSwiggen, D. T., Mir, M., Darzacq, X. & Tjian, R. Evaluating phase separation in live cells: diagnosis, caveats, and functional consequences. *Genes & development* (2019) doi:10.1101/gad.331520.119.

33. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).
34. Van Der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chemical Reviews* vol. 114 6589–6631 (2014).
35. Li, P. *et al.* Phase transitions in the assembly of multivalent signalling proteins. *Nature* (2012) doi:10.1038/nature10879.
36. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* (2017) doi:10.1038/nature22386.
37. Van Treeck, B. & Parker, R. Emerging Roles for Intermolecular RNA-RNA Interactions in RNP Assemblies. *Cell* (2018) doi:10.1016/j.cell.2018.07.023.
38. Brangwynne, C. P. *et al.* Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science (80-.).* (2009) doi:10.1126/science.1172046.
39. Lafontaine, D. L. J., Riback, J. A., Bascetin, R. & Brangwynne, C. P. The nucleolus as a multiphase liquid condensate. *Nature Reviews Molecular Cell Biology* (2020) doi:10.1038/s41580-020-0272-6.
40. Seif, E. *et al.* Phase separation by the polyhomeotic sterile alpha motif compartmentalizes Polycomb Group proteins and enhances their activity. *Nat. Commun.* (2020) doi:10.1038/s41467-020-19435-z.
41. Plys, A. J. *et al.* Phase separation of polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. *Genes Dev.* (2019) doi:10.1101/gad.326488.119.
42. Tatavosian, R. *et al.* Nuclear condensates of the Polycomb protein chromobox 2 (CBX2) assemble through phase separation. *J. Biol. Chem.* (2019) doi:10.1074/jbc.RA118.006620.
43. Guo, Y. E. *et al.* Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* (2019) doi:10.1038/s41586-019-1464-0.
44. Lu, Y. *et al.* Phase separation of TAZ compartmentalizes the

transcription machinery to promote gene expression. *Nat. Cell Biol.* **22**, 453–464 (2020).
45. Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, 6400 (2018).
46. Altmeyer, M. *et al.* Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). *Nat. Commun.* **6**, 8088 (2015).
47. Pandya-Jones, A. *et al.* A protein assembly mediates Xist localization and gene silencing. *Nature* (2020) doi:10.1038/s41586-020-2703-0.
48. Padeken, J. & Heun, P. Nucleolus and nuclear periphery: Velcro for heterochromatin. *Current Opinion in Cell Biology* vol. 28 54–60 (2014).
49. Tzur, Y. B., Wilson, K. L. & Gruenbaum, Y. SUN-domain proteins: 'Velcro' that links the nucleoskeleton to the cytoskeleton. *Nat. Rev. Mol. Cell Biol.* **7**, 782–788 (2006).
50. Falahati, H., Pelham-Webb, B., Blythe, S. & Wieschaus, E. Nucleation by rRNA dictates the precision of nucleolus assembly. *Curr. Biol.* (2016) doi:10.1016/j.cub.2015.11.065.
51. Berry, J. *et al.* RNA transcription modulates phase transition-driven nuclear body assembly. *Proc. Natl. Acad. Sci. U. S. A.* (2015) doi:10.1073/pnas.1509317112.
52. Mao, Y. S., Sunwoo, H., Zhang, B. & Spector, D. L. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nat. Cell Biol.* **13**, 95–101 (2011).
53. Nizami, Z., Deryusheva, S. & Gall, J. G. The Cajal body and histone locus body. *Cold Spring Harbor perspectives in biology* vol. 2 (2010).
54. Chaumeil, J., Le Baccon, P., Wutz, A. & Heard, E. A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.* **20**, 2223–2237 (2006).
55. Maison, C. *et al.* Higher-order structure in pericentric

heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nat. Genet.* (2002) doi:10.1038/ng843.

56. Quinodoz, S. A. *et al.* RNA promotes the formation of spatial compartments in the nucleus. *bioRxiv* (2020) doi:https://doi.org/10.1101/2020.08.25.267435.
57. Bernstein, E. & Allis, C. D. RNA meets chromatin. *Genes and Development* (2005) doi:10.1101/gad.1324305.
58. Bernstein, E. *et al.* Mouse Polycomb Proteins Bind Differentially to Methylated Histone H3 and RNA and Are Enriched in Facultative Heterochromatin. *Mol. Cell. Biol.* (2006) doi:10.1128/mcb.26.7.2560-2569.2006.
59. Markaki, Y. et al. Xist-seeded nucleation sites form local concentration gradients of silencing proteins to inactivate the X-chromosome. *bioRxiv* (2020) doi:https://doi.org/10.1101/2020.11.22.393546.
60. Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nat. Genet.* **30**, 167–174 (2002).
61. Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat. Struct. Mol. Biol.* **21**, 198–206 (2014).
62. Lu, Z. *et al.* RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* (2016) doi:10.1016/j.cell.2016.04.028.
63. Wang, L. *et al.* Histone Modifications Regulate Chromatin Compartmentalization by Contributing to a Phase Separation Mechanism. *Mol. Cell* (2019) doi:10.1016/j.molcel.2019.08.019.
64. Fan, C. *et al.* Rett mutations attenuate phase separation of MeCP2. *Cell Discovery* (2020) doi:10.1038/s41421-020-0172-0.
65. Li, C. H. *et al.* MeCP2 links heterochromatin condensates and neurodevelopmental disease. *Nature* (2020) doi:10.1038/s41586-020-2574-4.
66. Wang, L. *et al.* Rett syndrome-causing mutations compromise

MeCP2-mediated liquid–liquid phase separation of chromatin. *Cell Res.* (2020) doi:10.1038/s41422-020-0288-7.

67. Engreitz, J. M. *et al.* The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science (80-.).* **341**, 1237973–1237973 (2013).
68. Dossin, F. *et al.* SPEN integrates transcriptional and epigenetic control of X-inactivation. *Nature* (2020) doi:10.1038/s41586-020-1974-9.
69. McSwiggen, D. T. *et al.* Evidence for DNA-mediated nuclear compartmentalization distinct from phase separation. *Elife* **8**, e47098 (2019).
70. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* (2000) doi:10.1091/mbc.11.12.4241.
71. Chapal, M., Mintzer, S., Brodsky, S., Carmi, M. & Barkai, N. Resolving noise-control conflict by gene duplication. *bioRxiv* (2019) doi:10.1101/634741.
72. Zobeck, K. L., Buckley, M. S., Zipfel, W. R. & Lis, J. T. Recruitment Timing and Dynamics of Transcription Factors at the Hsp70 Loci in Living Cells. *Mol. Cell* (2010) doi:10.1016/j.molcel.2010.11.022.
73. Soutourina, J. Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology* (2018) doi:10.1038/nrm.2017.115.
74. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729–740 (1983).
75. Amano, T. *et al.* Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev. Cell* **16**, 47–57 (2009).
76. Benabdallah, N. S. *et al.* Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Mol. Cell* **76**, 473-484.e7 (2019).
77. Furlong, E. E. M. & Levine, M. Developmental enhancers and

chromosome topology. *Science* vol. 361 1341–1345 (2018).
78. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
79. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Science Advances* vol. 5 eaaw1668 (2019).
80. Symmons, O. *et al.* The Shh Topological Domain Facilitates the Action of Remote Enhancers by Reducing the Effects of Genomic Distances. *Dev. Cell* (2016) doi:10.1016/j.devcel.2016.10.015.
81. Flavahan, W. A. *et al.* Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. *Nature* (2019) doi:10.1038/s41586-019-1668-3.
82. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* (2016) doi:10.1038/nature16490.
83. Williamson, I. *et al.* Developmentally regulated Shh expression is robust to TAD perturbations. *Dev.* **146**, (2019).
84. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934 (2013).
85. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
86. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
87. Van Steensel, B. *et al.* Localization of the glucocorticoid receptor in discrete clusters in the cell nucleus. *J. Cell Sci.* **108**, 3003–3011 (1995).
88. Grande, M. A., Van der Kraan, I., De Jong, L. & Van Driel, R. Nuclear distribution of transcription factors in relation to sites of transcription and RNA polymerase II. *J. Cell Sci.* **110**, 1781–1791 (1997).
89. Bregman, D. B., Du, L., Van Der Zee, S. & Warren, S. L. Transcription-dependent redistribution of the large subunit of

RNA polymerase II to discrete nuclear domains. *J. Cell Biol.* (1995) doi:10.1083/jcb.129.2.287.

90. Papantonis, A. & Cook, P. R. Transcription Factories: Genome Organization and Gene Regulation. *Chem. Rev.* **113**, 8683–8705 (2013).
91. Iborra, F. J., Pombo, A., Jackson, D. A. & Cook, P. R. Active RNA polymerases are localized within discrete transcription “factories' in human nuclei. *J. Cell Sci.* **109**, 1427 LP – 1436 (1996).
92. Cisse, I. I. *et al.* Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells. *Science (80-.).* **341**, 664 LP – 667 (2013).
93. Wang, X., Cairns, M. J. & Yan, J. Super-enhancers in transcriptional regulation and genome organization. *Nucleic Acids Research* (2019) doi:10.1093/nar/gkz1038.
94. Cho, W. K. *et al.* RNA Polymerase II cluster dynamics predict mRNA output in living cells. *Elife* (2016) doi:10.7554/eLife.13617.
95. Chong, S. *et al.* Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science (80-.).* (2018) doi:10.1126/science.aar2555.
96. Berg, O. G., Winter, R. B. & von Hippel, P. H. Diffusion-Driven Mechanisms of Protein Translocation on Nucleic Acids. 1. Models and Theory. *Biochemistry* **20**, 6929–6948 (1981).
97. Elf, J., Li, G. W. & Xie, X. S. Probing transcription factor dynamics at the single-molecule level in a living cell. *Science (80-.).* **316**, 1191–1194 (2007).
98. Li, G. W. & Xie, X. S. Central dogma at the single-molecule level in living cells. *Nature* vol. 475 308–315 (2011).
99. Brodsky, S. *et al.* Intrinsically Disordered Regions Direct Transcription Factor In Vivo Binding Specificity. *Mol. Cell* **0**, (2020).
100. Valsecchi, C. I. K. *et al.* RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature* (2021)

doi:10.1038/s41586-020-2935-z.
101. Branco, M. R. & Pombo, A. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.* **4**, 780–788 (2006).
102. Shopland, L. S., Johnson, C. V., Byron, M., McNeil, J. & Lawrence, J. B. Clustering of multiple specific genes and gene-rich R-bands around SC-35 domains: Evidence for local euchromatic neighborhoods. *J. Cell Biol.* **162**, 981–990 (2003).
103. Lomvardas, S. *et al.* Interchromosomal Interactions and Olfactory Receptor Choice. *Cell* **126**, 403–413 (2006).
104. Strehle, M. & Guttman, M. Xist drives spatial compartmentalization of DNA and protein to orchestrate initiation and maintenance of X inactivation. *Current Opinion in Cell Biology* (2020) doi:10.1016/j.ceb.2020.04.009.
105. Loda, A. & Heard, E. Xist RNA in action: Past, present, and future. *PLoS Genetics* (2019) doi:10.1371/journal.pgen.1008333.
106. Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404–416 (2015).
107. Minajigi, A. *et al.* A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science (80-.).* (2015) doi:10.1126/science.aab2276.
108. McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232–236 (2015).
109. Moindrot, B. *et al.* A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing. *Cell Rep.* (2015) doi:10.1016/j.celrep.2015.06.053.
110. Monfort, A. *et al.* Identification of Spen as a crucial factor for Xist function through forward genetic screening in haploid embryonic stem cells. *Cell Rep.* (2015) doi:10.1016/j.celrep.2015.06.067.
111. Wutz, A. & Jaenisch, R. A shift from reversible to irreversible X

inactivation is triggered during ES cell differentiation. *Mol. Cell* **5**, 695–705 (2000).
112. Pandey, R. R. *et al.* Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation. *Mol. Cell* (2008) doi:10.1016/j.molcel.2008.08.022.
113. Kanduri, C. Kcnq1ot1: A chromatin regulatory RNA. *Seminars in Cell and Developmental Biology* (2011) doi:10.1016/j.semcdb.2011.02.020.
114. Nagano, T. & Fraser, P. Emerging similarities in epigenetic gene silencing by long noncoding RNAs. *Mammalian Genome* (2009) doi:10.1007/s00335-009-9218-1.
115. Cheutin, T. *et al.* Maintenance of stable heterochromatin domains by dynamic HP1 binding. *Science (80-.).* (2003) doi:10.1126/science.1078572.
116. Maison, C. & Almouzni, G. HP1 and the dynamics of heterochromatin maintenance. *Nature Reviews Molecular Cell Biology* (2004) doi:10.1038/nrm1355.
117. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science (80-.).* (2008) doi:10.1126/science.1163045.
118. Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. Xist RNA and the mechanism of X chromosome inactivation. *Annual Review of Genetics* (2002) doi:10.1146/annurev.genet.36.042902.092433.
119. Mira-Bontenbal, H. & Gribnau, J. New Xist-Interacting Proteins in X-Chromosome Inactivation. *Current Biology* (2016) doi:10.1016/j.cub.2016.03.022.
120. Cerase, A. *et al.* Phase separation drives X-chromosome inactivation: a hypothesis. *Nature Structural and Molecular Biology* (2019) doi:10.1038/s41594-019-0223-0.
121. Wutz, A. Xist function: bridging chromatin and stem cells. *Trends in Genetics* (2007) doi:10.1016/j.tig.2007.07.004.

122. Cao, R. *et al.* Role of Histone H3 Lysine 27 Methylation in Polycomb-Group Silencing. *Science (80-.).* **298**, 1039 LP – 1043 (2002).
123. Min, J., Zhang, Y. & Xu, R. M. Structural basis for specific binding of polycomb chromodomain to histone H3 methylated at Lys 27. *Genes Dev.* **17**, 1823–1828 (2003).
124. Fischle, W. *et al.* Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by polycomb and HP1 chromodomains. *Genes Dev.* **17**, 1870–1881 (2003).
125. Illingworth, R. S. Chromatin folding and nuclear architecture: PRC1 function in 3D. *Current Opinion in Genetics and Development* (2019) doi:10.1016/j.gde.2019.06.006.
126. Pirrotta, V. & Li, H. B. A view of nuclear Polycomb bodies. *Current Opinion in Genetics and Development* (2012) doi:10.1016/j.gde.2011.11.004.
127. Eeftens, J. M., Kapoor, M. & Brangwynne, C. P. Epigenetic memory as a time integral over prior history of Polycomb phase separation. *bioRxiv* (2020).
128. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* (2001) doi:10.1038/35065132.
129. Girard, C. *et al.* Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat. Commun.* **3**, (2012).
130. Reimer, K. Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis.
131. Neugebauer, K. M. On the importance of being co-transcriptional. *J. Cell Sci.* (2002) doi:10.1242/jcs.00073.
132. Napolitano, G., Lania, L. & Majello, B. RNA Polymerase II CTD Modifications: How many tales from a single tail. *J. Cell. Physiol.* (2014) doi:10.1002/jcp.24483.
133. Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes and Development* vol. 26 2119–2137 (2012).

134. Ahn, S. H., Kim, M. & Buratowski, S. Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3′ End Processing. *Mol. Cell* (2004) doi:10.1016/S1097-2765(03)00492-1.
135. Merkhofer, E. C., Hu, P. & Johnson, T. L. Introduction to cotranscriptional RNA splicing. *Methods in Molecular Biology* (2014) doi:10.1007/978-1-62703-980-2_6.
136. Han, J., Xiong, J., Wang, D. & Fu, X. D. Pre-mRNA splicing: Where and when in the nucleus. *Trends in Cell Biology* (2011) doi:10.1016/j.tcb.2011.03.003.
137. Herzel, L., Ottoz, D. S. M., Alpert, T. & Neugebauer, K. M. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. *Nature Reviews Molecular Cell Biology* (2017) doi:10.1038/nrm.2017.63.
138. McCracken, S. *et al.* The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* (1997) doi:10.1038/385357a0.
139. Zhang, S., Aibara, S., Vos, S. M. & Agafonov, D. E. Structure of a transcribing RNA polymerase II – U1 snRNP complex. *Science (80-.).* **371**, 1–35 (2021).
140. Ding, F. & Elowitz, M. B. Constitutive splicing and economies of scale in gene expression. *Nat. Struct. Mol. Biol.* **26**, 424–432 (2019).
141. Zamudio, A. V. *et al.* Mediator Condensates Localize Signaling Factors to Key Cell Identity Genes. *Mol. Cell* **76**, 753-766.e6 (2019).
142. Chen, Y. *et al.* Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J. Cell Biol.* (2018) doi:10.1083/jcb.201807108.
143. Dopie, J., Sweredoski, M. J., Moradian, A. & Belmont, A. S. Tyramide signal amplification mass spectrometry (TSA-MS) ratio identifies nuclear speckle proteins. *J. Cell Biol.* (2020) doi:10.1083/jcb.201910207.
144. Saitoh, N. *et al.* Proteomic analysis of interchromatin granule

clusters. *Mol. Biol. Cell* (2004) doi:10.1091/mbc.E04-03-0253.
145. Lamond, A. I. & Spector, D. L. Nuclear speckles: a model for nuclear organelles. *Nat. Rev. Mol. Cell Biol.* **4**, 605–612 (2003).
146. Kim, J., Venkata, N. C., Hernandez Gonzalez, G. A., Khanna, N. & Belmont, A. S. Gene expression amplification by nuclear speckle association. *J. Cell Biol.* (2020) doi:10.1083/jcb.201904046.
147. Misteli, T. & Spector, D. L. RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. *Mol. Cell* **3**, 697–705 (1999).
148. Pederson, T. The nucleolus. *Cold Spring Harb. Perspect. Biol.* **3**, 1–15 (2011).
149. Carmo-Fonseca, M. & Rino, J. RNA seeds nuclear bodies. *Nat. Cell Biol.* (2011) doi:10.1038/ncb0211-110.
150. Shevtsov, S. P. & Dundr, M. Nucleation of nuclear bodies by RNA. *Nat. Cell Biol.* (2011) doi:10.1038/ncb2157.
151. Marzluff, W. F. & Koreski, K. P. Birth and Death of Histone mRNAs. *Trends in Genetics* (2017) doi:10.1016/j.tig.2017.07.014.
152. Scheer, U. & Benavente, R. Functional and dynamic aspects of the mammalian nucleolus. *BioEssays* (1990) doi:10.1002/bies.950120104.
153. SCHLEGEL, R. A., MILLER, L. S. & ROSE, K. M. Reduction in RNA synthesis following red cell-mediated microinjection of antibodies to RNA polymerase I. *Cell Bid. Int. Rep.* **9**, 341–350 (1985).
154. Watkins, N. J. & Bohnsack, M. T. 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* (2012) doi:10.1002/wrna.117.
155. Kiss-László, Z., Henry, Y., Bachellerie, J. P., Caizergues-Ferrer, M. & Kiss, T. Site-specific ribose methylation of preribosomal RNA: A novel function for small nucleolar RNAs. *Cell* (1996) doi:10.1016/S0092-8674(00)81308-2.

156. Ni, J., Tien, A. L. & Fournier, M. J. Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell* (1997) doi:10.1016/S0092-8674(00)80238-X.
157. Spycher, C. *et al.* 3′ end processing of mouse histone pre-mRNA: Evidence for additional base-pairing between U7 snRNA and pre-mRNA. *Nucleic Acids Res.* (1994) doi:10.1093/nar/22.20.4023.
158. Kolev, N. G. & Steitz, J. A. Symplekin and multiple other polyadenylation factors participate in 3′-end maturation of histone mRNAs. *Genes Dev.* (2005) doi:10.1101/gad.1371105.
159. Mowry, K. L. & Steitz, J. A. Identification of the human U7 snRNP as one of several factors involved in the 3′ end maturation of histone premessenger RNA's. *Science (80-.).* (1987) doi:10.1126/science.2825355.
160. Machyna, M. *et al.* The coilin interactome identifies hundreds of small noncoding RNAs that traffic through cajal bodies. *Mol. Cell* (2014) doi:10.1016/j.molcel.2014.10.004.
161. Darzacq, X. *et al.* Cajal body-specific small nuclear RNAs: A novel class of 2′-O-methylation and pseudouridylation guide RNAs. *EMBO J.* (2002) doi:10.1093/emboj/21.11.2746.
162. Kiss, A. M. *et al.* A Cajal body-specific pseudouridylation guide RNA is composed of two box H/ACA snoRNA-like domains. *Nucleic Acids Res.* (2002) doi:10.1093/nar/gkf592.
163. Erdel, F. *et al.* Mouse Heterochromatin Adopts Digital Compaction States without Showing Hallmarks of HP1-Driven Liquid-Liquid Phase Separation. *Mol. Cell* (2020) doi:10.1016/j.molcel.2020.02.005.
164. van Steensel, B. & Belmont, A. S. Lamina-associated domains: Links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169**, 780–791 (2017).
165. Reddy, K. L., Zullo, J. M., Bertolino, E. & Singh, H. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature* **452**, 243–7 (2008).
166. Kumaran, R. I. & Spector, D. L. A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional

competence. *J. Cell Biol.* **180**, 51–65 (2008).
167. Finlan, L. E. *et al.* Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS Genet.* **4**, 3 (2008).
168. Kim, J. H. *et al.* LADL: light-activated dynamic looping for endogenous gene expression control. *Nat. Methods* (2019) doi:10.1038/s41592-019-0436-5.
169. Wang, H. *et al.* CRISPR-Mediated Programmable 3D Genome Positioning and Nuclear Organization. *Cell* (2018) doi:10.1016/j.cell.2018.09.013.
170. Shin, Y. *et al.* Spatiotemporal Control of Intracellular Phase Transitions Using Light-Activated optoDroplets. *Cell* (2017) doi:10.1016/j.cell.2016.11.054.
171. Shin, Y. *et al.* Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome. *Cell* (2018) doi:10.1016/j.cell.2018.10.057.
172. Takei, Y. *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021).
173. Eng, C. H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* (2019) doi:10.1038/s41586-019-1049-y.
174. Collombet, S. *et al.* RNA polymerase II depletion from the inactive X chromosome territory is not mediated by physical compartmentalization. *bioRxiv* 2021.03.26.437188 (2021) doi:10.1101/2021.03.26.437188.
175. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology* (2016) doi:10.1038/nrm.2016.104.
176. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* (2019) doi:10.1038/s41576-019-0195-2.
177. Ganser, L. R. & Myong, S. Methods to Study Phase-Separated Condensates and the Underlying Molecular Interactions. *Trends Biochem. Sci.* **45**, 1004–1005 (2020).
178. Robinson, C. V., Sali, A. & Baumeister, W. The molecular

sociology of the cell. *Nature* **450**, 973–982 (2007).
179. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* (2016) doi:10.1016/j.neuron.2016.10.001.
180. Shah, S. *et al.* Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell* (2018) doi:10.1016/j.cell.2018.05.035.
181. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods* (2014) doi:10.1038/nmeth.2892.
182. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* (2017) doi:10.1038/nature20777.
183. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci. U. S. A.* (2019) doi:10.1073/pnas.1912459116.
184. Beliveau, B. J. *et al.* Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc. Natl. Acad. Sci. U. S. A.* (2012) doi:10.1073/pnas.1213818110.
185. Beliveau, B. J. *et al.* Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nat. Commun.* (2015) doi:10.1038/ncomms8147.
186. Beliveau, B. J., Apostolopoulos, N. & Wu, C. T. Visualizing genomes with oligopaint FISH probes. *Curr. Protoc. Mol. Biol.* (2013) doi:10.1002/0471142727.mb1423s105.
187. Takei, Y. *et al.* Global architecture of the nucleus in single cells by DNA seqFISH+ and multiplexed immunofluorescence. *bioRxiv* (2020).
188. Dey, S. & Maiti, S. Single-molecule photobleaching: Instrumentation and applications. *J. Biosci.* **43**, 447–454 (2018).
189. Hsia, Y. *et al.* Design of a hyperstable 60-subunit protein

icosahedron. *Nature* (2016) doi:10.1038/nature18010.
190. Elson, E. L. Introduction to fluorescence correlation Spectroscopy—Brief and simple. *Methods* **140**–**141**, 3–9 (2018).
191. Huang, B., Babcock, H. & Zhuang, X. Breaking the diffraction barrier: Super-resolution imaging of cells. *Cell* (2010) doi:10.1016/j.cell.2010.12.002.
192. Godin, A. G., Lounis, B. & Cognet, L. Super-resolution microscopy approaches for live cell imaging. *Biophysical Journal* (2014) doi:10.1016/j.bpj.2014.08.028.
193. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* (2009) doi:10.1126/science.1181369.
194. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).

Chapter 2

RNA PROMOTES THE FORMATION OF SPATIAL COMPARTMENTS IN THE NUCLEUS

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

*equal contribution

A modified version of this chapter was published as: Quinodoz, S. A. *et al.* RNA promotes the formation of spatial compartments in the nucleus. *Cell* **184**, 5775-5790.e30 (2021).

"Everything is related to everything else, but near things are more related than distant things." – Waldo Tobler, The First Law of Geography

2.1 ABSTRACT

RNA, DNA, and protein molecules are highly organized within three-dimensional (3D) structures in the nucleus. Although RNA has been proposed to play a role in nuclear organization, exploring this has been challenging because existing methods cannot measure higher-order RNA and DNA contacts within 3D structures. To address this, we developed RNA & DNA SPRITE (RD-SPRITE) to comprehensively map the spatial organization of RNA and DNA. These maps reveal higher-order RNA-chromatin structures associated with three major classes of nuclear function: RNA processing, heterochromatin assembly, and gene regulation. These data demonstrate that hundreds of ncRNAs form high-concentration territories throughout the nucleus, that specific RNAs are required to recruit various regulators into these territories, and that these RNAs can shape long-range DNA contacts, heterochromatin assembly, and gene expression. These results demonstrate a mechanism where RNAs form high-concentration territories, bind to diffusible regulators, and guide them into compartments to regulate essential nuclear functions.

2.2 INTRODUCTION

The nucleus is spatially organized in three-dimensional (3D) structures that are important for various functions including transcription and RNA processing [1–3]. To date, genome-wide studies of nuclear organization have focused primarily on the role of DNA [1,4], yet nuclear structures are known to contain DNA, RNA, and protein molecules that are involved in shared functional and regulatory processes. These include classical compartments like the nucleolus [5] (which contains transcribed ribosomal RNAs and their processing molecules) and nuclear speckles [6] (which contain nascent pre-mRNAs and mRNA splicing components), as well as more recently described transcriptional condensates (which contain Mediator and RNA Pol II) [7,8]. Because the complete molecular architecture of the nucleus has not been globally explored, the extent to which such compartments exist and contribute to nuclear function remains unknown. Even for the specific nuclear compartments that have been characterized, the mechanism by which intrinsically diffusible RNA and protein molecules become spatially organized remains unclear.

Nuclear RNA has long been proposed to play a central role in shaping nuclear structure [9,10]. Over the past decade it has become clear that mammalian genomes encode thousands of nuclear-enriched ncRNAs [11], several of which play critical regulatory roles [12]. These include ncRNAs involved in splicing of pre-mRNAs (snRNAs) [13,14], cleavage and modification of pre-ribosomal RNAs (snoRNAs, Rnase MRP) [15,16], 3'-end cleavage and processing of the non-polyadenylated histone pre-mRNAs (U7 snRNA) [17], and transcriptional regulation (e.g., Xist [18] and 7SK [19]). Many of these ncRNAs localize within specific compartments in the nucleus [3]. For example, snoRNAs and the 45S pre-ribosomal RNA localize within the nucleolus [5], the Xist lncRNA localizes on the inactive X chromosome (Barr body) [20], and snRNAs and Malat1 localize within nuclear speckles [21].

In each of these examples, RNA, DNA, and protein components simultaneously interact within precise structures. While the localization of

specific ncRNAs have been well studied, the localization patterns of most nuclear ncRNAs remain unknown 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 are involved in nuclear organization, (ii) which nuclear compartments are dependent on RNA, and (iii) what mechanisms RNAs utilize to organize nuclear structures.

Microscopy is currently the only way to relate RNA and DNA molecules in 3D space, yet it is limited to examining a small number of components and 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 [22–26]. While these can provide genome-wide pairwise maps of RNA-DNA interactions, they do not provide information about the 3D organization of these molecules. Moreover, we recently showed that proximity-ligation methods can fail to identify pairwise contacts between molecules if they are not close enough in space to be directly ligated [27]. Consistent with this, existing methods fail to identify known RNA-DNA contacts within nuclear bodies including nucleoli, histone locus bodies, and Cajal bodies [24–26].

We recently developed SPRITE, which utilizes split-and-pool barcoding to generate comprehensive and multi-way 3D maps of the nucleus across a wide range of distances [27]. We showed that SPRITE accurately maps the spatial organization of DNA arranged around two nuclear bodies – nucleoli and nuclear speckles. However, our original version could not detect the majority of RNAs, including low abundance ncRNAs known to organize within several well-defined nuclear structures. Here, we introduce a dramatically improved method, RNA & DNA SPRITE (RD-SPRITE), which enables simultaneous, 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 and hundreds of ncRNAs that form high concentration territories throughout the nucleus. Focusing on specific

examples, we show that many of these RNAs recruit diffusible ncRNA and protein regulators and can shape long-range DNA contacts, heterochromatin assembly, and gene expression within these territories. Together, our results highlight a role for RNA in the formation of compartments involved in essential nuclear functions including RNA processing, heterochromatin assembly, and gene regulation.

2.3 RESULTS

2.3.1 RD-SPRITE GENERATES ACCURATE MAPS OF HIGHER-ORDER RNA AND DNA CONTACTS

To explore the role of RNA in shaping nuclear structure, we improved the efficiency of the RNA-tagging steps of our SPRITE method [27] to enable detection of all classes of RNA (see **Methods**). We refer to this new approach as RNA & DNA SPRITE (RD-SPRITE). It 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 fragmented into smaller complexes, (iii) molecules within each complex are tagged with an RNA or DNA-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 complex, (v) DNA and RNA are sequenced, and (vi) all reads sharing identical barcodes are merged into a SPRITE cluster (**Figure 1A, S1A-B**). Because RD-SPRITE does not rely on proximity ligation, it can detect multiple RNA and DNA molecules that associate simultaneously.

We performed RD-SPRITE in an F1 hybrid female mouse ES cell line engineered to induce Xist from a single allele. We sequenced libraries on

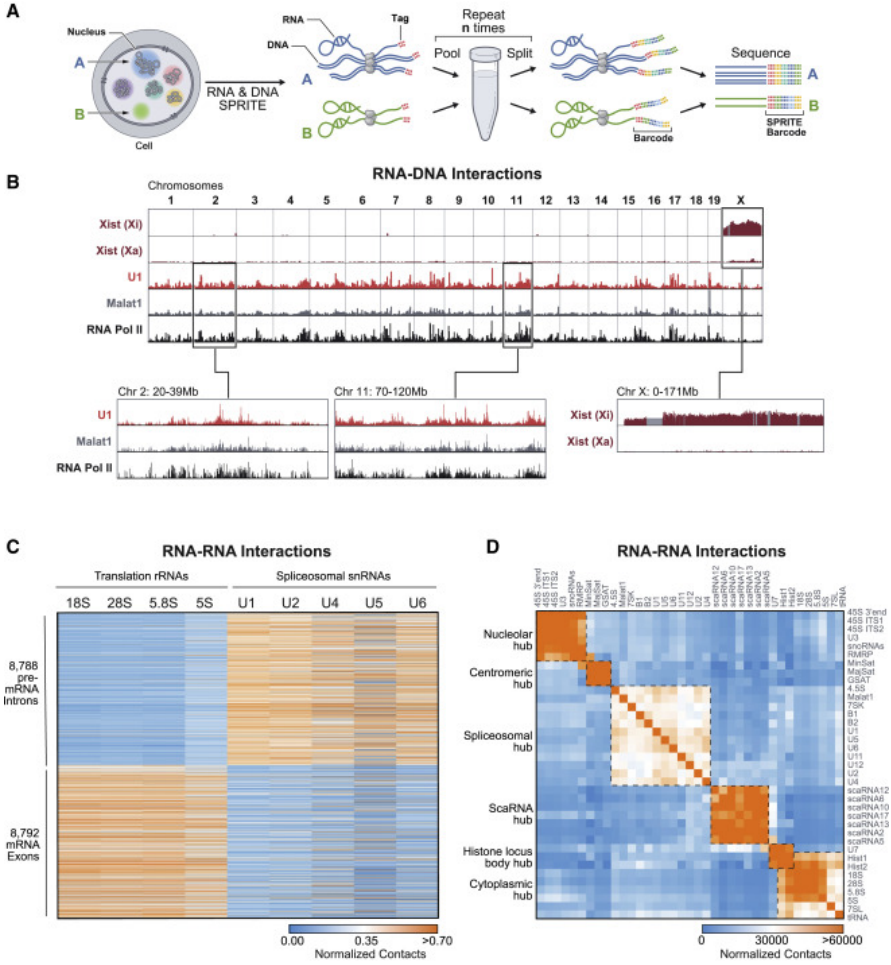


Figure 1. RD-SPRITE generates maps of higher-order RNA and DNA contacts. *(A) Schematic of RD-SPRITE: crosslinked cells are fragmented, DNA and RNA are barcoded through multiple rounds of split-and-pool barcoding, and SPRITE clusters are defined as a group of molecules sharing a barcode.(B) Xist unweighted contacts on the inactive (Xi) or active X chromosome (Xa), U1 and Malat1 weighted contacts, and RNA Pol II (ENCODE) across the genome. Gray demarcates masked regions. (C) Heatmap showing unweighted RNA-RNA contacts between translation-associated RNAs or splicing RNAs (columns) and introns or exons of mRNAs (rows). (D) Heatmap of unweighted RNA-RNA contact frequencies for several classes of RNA. Boxes denote hubs.*

a NovaSeq S4 run to generate ~8 billion reads corresponding to ~720 million SPRITE clusters (**Figure S1C, Table S2-3**). To ensure that RD-SPRITE accurately measures *bona fide* RNA interactions, we focused on RNA-DNA contacts for several ncRNAs that were previously mapped to chromatin and reflect a range of known *cis* and *trans* localization patterns. We observed strong enrichment of: (i) Xist over the inactive X (Xi), but not the active X chromosome (Xa) (**Figure 1B, S1D**) [20]; (ii) Malat1 and U1 over actively transcribed Pol II genes (**Figure 1B**) [28,29]; and (iii) telomerase RNA component (Terc) over telomere-proximal regions of all chromosomes (**Figure S1E**) [30,31].

Next, we focused on known RNA-RNA contacts in different cellular locations. We observed a large number of contacts between translation-associated RNAs in the cytoplasm, including all RNA components of the ribosome and ~8000 individual mRNAs (exons), but not with pre-mRNAs (introns) (**Figure 1C**). Conversely, we observed many contacts between snRNA components of the spliceosome and individual pre-mRNAs (introns) in the nucleus (**Figure 1C**).

Together, these results demonstrate that RD-SPRITE accurately measures RNA-DNA and RNA-RNA contacts in the nucleus and cytoplasm. While we focus primarily on contacts within the nucleus, RD-SPRITE can also be utilized to study RNA compartments beyond the nucleus [32].

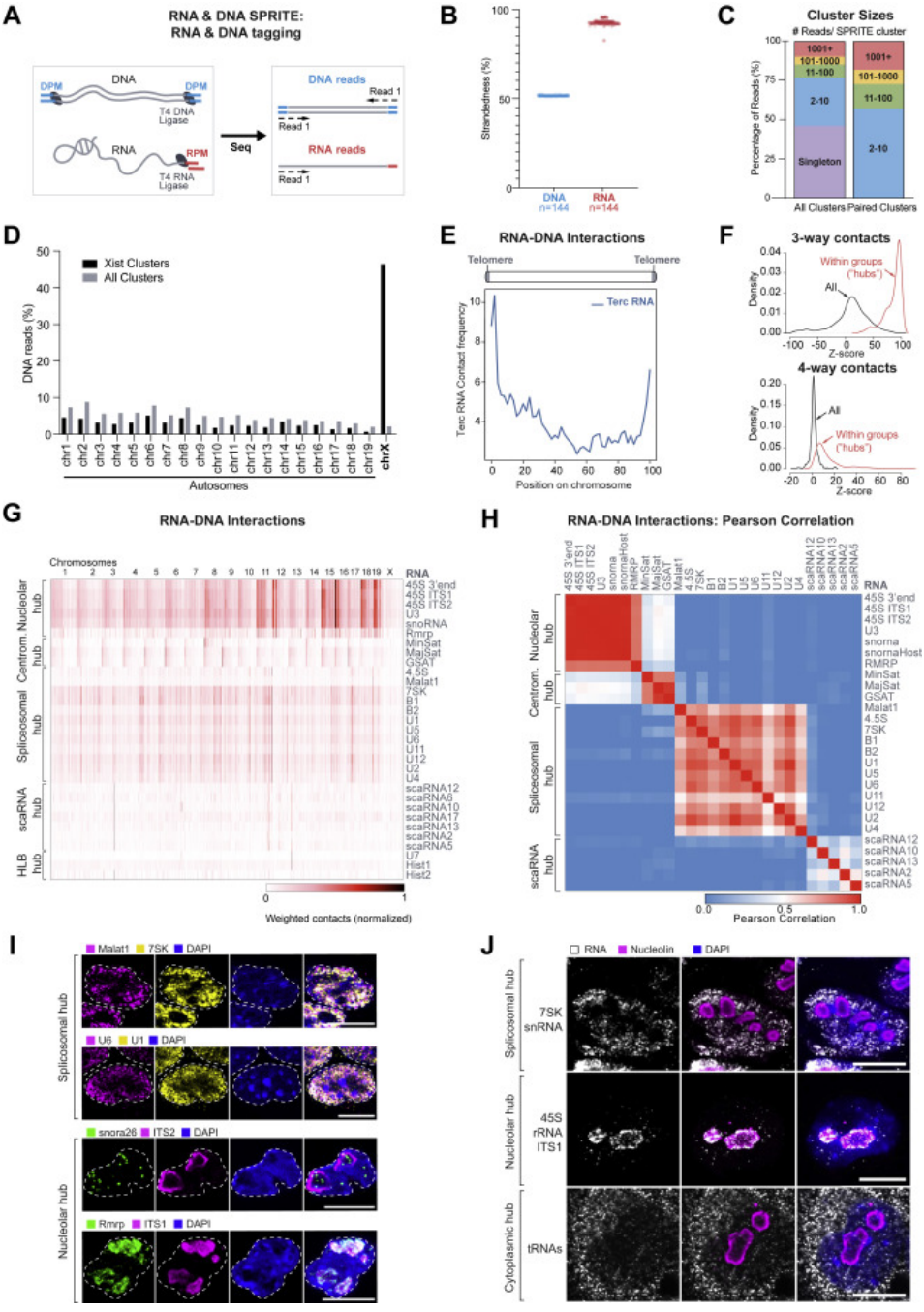

A
RNA & DNA SPRITE:
RNA & DNA tagging
DNA
DPM
DPM
T4 DNA Ligase
RNA
RPM
T4 RNA Ligase
Seq
DNA reads
Read 1
Read 1
RNA reads
Read 1
B
Strandedness (%)
DNA n=144
RNA n=144
C
Cluster Sizes
Reads/ SPRITE cluster
Percentage of Reads (%)
1001+
101-1000
11-100
2-10
Singleton
All Clusters
Paired Clusters
D
Xist Clusters
All Clusters
DNA reads (%)
Autosomes
chrX
E
RNA-DNA Interactions
Telomere
Telomere
Terc RNA
Terc RNA Contact Frequency
Position on chromosome
F
3-way contacts
Within groups ("hubs")
All
Density
Z-score
4-way contacts
G
RNA-DNA Interactions
Chromosomes
RNA
Nucleolar hub
Centrom. hub
Spliceosomal hub
scaRNA hub
HLB hub
Weighted contacts (normalized)
H
RNA-DNA Interactions: Pearson Correlation
Pearson Correlation
I
Malat1 7SK DAPI
U6 U1 DAPI
snora26 ITS2 DAPI
Rmrp ITS1 DAPI
Spliceosomal hub
Nucleolar hub
J
RNA Nucleolin DAPI
7SK snRNA
45S rRNA ITS1
tRNAs
Spliceosomal hub
Nucleolar hub
Cytoplasmic hub

Figure S1. 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, DPM and RPM, 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. 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 contacts 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 STAR 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 or "hub" (within group, red), defined by sets of pairwise interacting RNAs (see 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). Panels show individual RNAs (left), DAPI (right-middle); and overlays (right). Scalebar is 10μm. (J) RNA FISH (left) of specific, hub-associated ncRNA along with nucleolin immunofluorescence (middle) and DAPI (right). tRNAs are visualized using pooled RNA FISH probes (see STAR Methods). Scalebar is 10μm.*

2.3.2 MULTIPLE NCRNAS CO-LOCALIZE WITHIN SPATIAL COMPARTMENTS IN THE NUCLEUS

To explore which RNAs localize within spatial compartments, we first mapped pairwise RNA-RNA and RNA-DNA contacts and identified several groups of RNAs that 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 localize to similar genomic DNA regions (**Figure S1G-H**). Using a combination of RNA FISH and immunofluorescence (IF), we

confirmed that RNAs within a group co-localize (**Figure S1I**) while RNAs in distinct groups localize to different regions of the cell (**Figure S1J**).

We next explored whether groups of pairwise interacting RNAs simultaneously associate within higher-order structures. To do this, we compared the frequency of contacts between 3 or more distinct RNAs to the expected frequency if these RNAs were randomly distributed. We observed many significant multi-way contacts between RNAs within each group (**Table S1**). Overall, we observed a significantly higher number of multi-way contacts among RNAs within a group than between RNAs from distinct groups (~50-fold for 3-way contacts, **Figure S1F**). Because these groups of RNAs are found in higher-order structures, we refer to them as "hubs" and explore them below.

2.3.3 NCRNAS 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 RNAs within these hubs (RNA-RNA interactions), their location relative to genomic DNA (RNA-DNA interactions), and the 3D organization of these DNA loci (DNA-DNA interactions).

(i) ***ncRNAs involved in ribosomal RNA processing organize around transcribed ribosomal RNA genes.*** We identified a hub that includes the 45S pre-ribosomal RNA, RNase MRP, and dozens of snoRNAs involved in rRNA biogenesis (**Figure 1D, S2A**). rRNA is transcribed as a single 45S precursor RNA, is cleaved by RNAse MRP, and is modified by various snoRNAs to generate the mature 18S, 5.8S, and 28S rRNAs [33]. We found that these ncRNAs form multi-way contacts with each other ($p<0.01$, z-score=31, **Table S1**) and localize at genomic locations proximal to ribosomal DNA repeats that encode the 45S pre-rRNA and other genomic regions that organize around the nucleolus [27] (**Figure 2A, S2B**). We explored the DNA-DNA interactions that occur within SPRITE

clusters containing multiple nucleolar hub RNAs and observed that these RNAs and genomic DNA regions are organized together in 3D space (**Figure 2B**, **S2C**). Our results demonstrate that the nascent 45S pre-rRNA, along with the diffusible snoRNAs and RNase MRP, are spatially enriched near the DNA loci from which rRNA is transcribed.

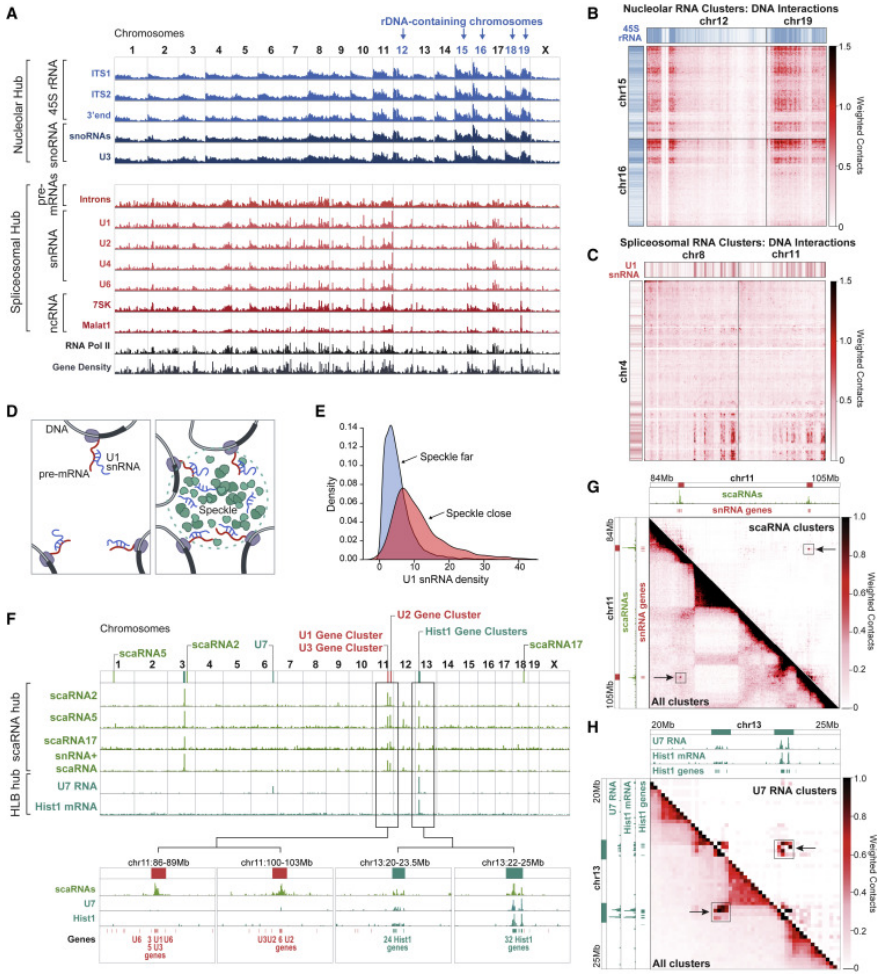


Figure 2. Non-coding RNAs involved in RNA processing organize within hubs. *(A) Weighted RNA-DNA contacts (1-Mb resolution) for several RNAs within the nucleolar and spliceosomal hubs are plotted alongside Pol II occupancy (ENCODE) and gene density.*

Chromosomes with rDNA are shown in blue. (B) Weighted DNA-DNA contacts in SPRITE clusters containing nucleolar hub RNAs are shown between chromosomes 12 + 19 and 15 + 16. Blue/white color bar represents high and low 45S rRNA RNA-DNA contacts. (C) Weighted DNA-DNA contacts in SPRITE clusters containing spliceosomal hub RNAs are shown between chromosomes 4 and 8 + 11. Red/white color bar represents U1 snRNA RNA-DNA contacts. (D) Illustration of two possible snRNA localization models: (left) localization occurs primarily through association with nascent pre-mRNAs, and (right) localization depends on 3D position of an individual gene within the nucleus. (E) U1 snRNA density over genomic DNA regions with comparable expression levels that are close (red) or far (blue) from nuclear speckles. (F) Weighted RNA-DNA contacts for clusters containing various scaRNAs or both scaRNAs and snRNAs (green) or U7 and histone pre-mRNAs (teal). (G) Weighted DNA-DNA contacts across a genomic region containing snRNA genes for all (bottom) or scaRNA-containing (top) SPRITE clusters. scaRNA RNA-DNA contacts are shown along the top and side axes. Enriched loci are highlighted by a black box and arrow. (H) Weighted DNA-DNA contacts in a genomic region containing histone genes for all (bottom) or U7-containing (top) SPRITE clusters. U7 and histone pre-mRNA RNA-DNA contacts are shown along the top and side axis. Enriched loci are marked with a black box and arrow.

(ii) ncRNAs involved in mRNA splicing are spatially concentrated around genes containing a high density of Pol II. We identified a hub that contains nascent pre-mRNAs, major and minor spliceosomal ncRNAs, and other ncRNAs associated with transcriptional regulation and mRNA splicing (**Figure 1D, Table S1**). Nascent pre-mRNAs are known to be directly bound and cleaved by spliceosomal RNAs to generate mature mRNA transcripts [34], yet it is unclear how spliceosomal RNAs are organized in the nucleus relative to target pre-mRNAs and genomic DNA [35,36]. 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 the DNA occupancy of splicing RNAs to be proportional to mRNA transcription levels, regardless of the 3D position of an individual gene in the nucleus. However, we find that splicing RNAs do not show a uniform occupancy over all genes but are more highly enriched over DNA regions containing a high-density of actively transcribed Pol II genes (r = 0.86-0.90, **Figure 2A, S2B,D**). When we explored the higher-order DNA contacts of these RNAs, we found that these genomic DNA regions form preferential inter-chromosomal contacts and are comparable to regions organized around

nuclear speckles [27] (**Figure 2C, S2E**). 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 2D**), even when focusing on genes with comparable levels of transcription (**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 organized around snRNA gene clusters. We identified a hub containing several small Cajal body-associated RNAs (scaRNAs) and snRNAs (**Figure 1D, Table S1, Figure S2F**). 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 [37]. scaRNAs directly hybridize to snRNAs to guide these modifications [38]. We found that scaRNAs are highly enriched at discrete genomic regions containing multiple snRNA genes in close linear space (**Figure 2F**). Although we cannot directly distinguish between the spatial localization of nascent snRNAs and mature snRNAs, we found that SPRITE clusters containing snRNAs and scaRNAs are highly enriched at genomic DNA regions containing snRNA genes **(Figure 2F)**, indicating that nascent snRNAs are enriched near their transcriptional loci. Despite being separated by large genomic distances, these DNA regions form long-range contacts (**Figure 2G**) and scaRNAs, snRNAs, and their associated DNA loci simultaneously interact within higher-order SPRITE clusters (**Figure S2G**). 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 [39] (**Figure S2J**). However, Cajal bodies are traditionally defined by the presence of Coilin foci in the nucleus [40–42] and based on this definition, our mES cells do not contain visible Cajal bodies (**Figure S2L**). 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.

(iv) The histone processing U7 snRNA is enriched around histone gene loci. We identified a hub containing U7 and various histone mRNAs (**Figure 1D**). Unlike most pre-mRNAs, histone pre-mRNAs are not polyadenylated; their 3'ends are bound and cleaved by the U7 snRNP complex to produce mature histone mRNAs [43,44]. This process is thought to occur within nuclear structures called Histone Locus Bodies (HLBs) [45], demarcated by NPAT protein (**Figure S2H**). We observed that U7 localizes at genomic DNA regions containing histone mRNA genes, specifically at two histone gene clusters on chromosome 13 (**Figure 2F**). To determine whether U7, histone genes, and histone pre-mRNAs spatially co-occur, we focused on DNA-DNA contacts from U7-containing clusters and observed long-range DNA contacts between the two histone gene clusters on chromosome 13 (**Figure 2H**). Consistent with previous observations that HLBs and Cajal bodies are often adjacent to each other in the nucleus [45], 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 2F, S2G, S2I-L**).

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 forms of RNA processing.

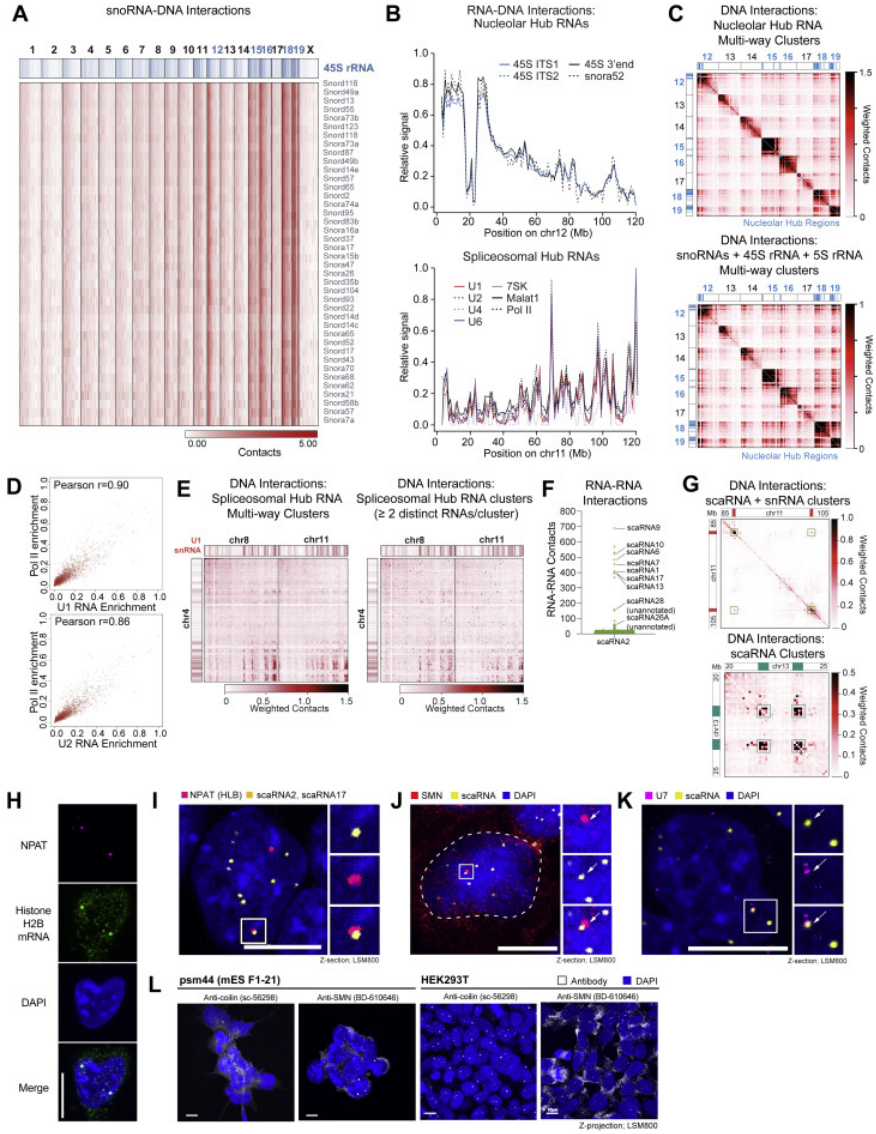


Figure S2. Various RNA processing bodies are organized around the transcriptional loci of their targets. *(A) Genome-wide localization of each individual snoRNA, as determined by unweighted RNA-DNA contact frequency. Blue track shows 45S pre-rRNA localization on DNA. Chromosomes containing ribosomal DNA (rDNA) genes are denoted in blue. (B) RNA-DNA contact frequencies on (top) chromosome 12 for various RNAs within the*

nucleolar hub and on (bottom) 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 (top) and snoRNAs, 45S, and 5S (bottom) simultaneously. (D) Genome-wide 1Mb enrichment of spliceosomal hub RNA-DNA interactions (U1 and U2 snRNA) compared to enrichment of Pol II ChIP-seq signal (ENCODE). 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 snRNA contacts on DNA are shown as a heatmap (red-white scale) along the top and side axes. (F) RNA-RNA contact frequency between scaRNA2 and all RNAs. Top hits include annotated scaRNAs and two previously unannotated scaRNAs, which we identified (see STAR Methods). (G) Weighted DNA-DNA contacts within (top) SPRITE clusters containing both scaRNAs and snRNAs are shown across a region on chromosome 11 which contains snRNA gene clusters (red boxes) and (bottom) SPRITE clusters containing scaRNAs across a region on chromosome 13 which contains histone gene clusters (teal boxes). (H) IF of NPAT (magenta), RNA FISH of Histone H2B mRNA (green), nuclear stain with DAPI (blue) and overlaid images in mES cells. Scalebar is 10μm. (I) 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. (J) 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). It is possible that these snRNA processing bodies might represent nuclear gems (Matera and Frey, 1998), which contain SMN protein, or "residual bodies," which are Coilin negative (Nizami et al., 2010; Tucker et al., 2001). We observe SMN foci in our mES cells and that some, but not all, scaRNAs colocalize with SMN protein in the nucleus. Scalebar is 10μm. (K) RNA FISH image of a 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. (L) Immunofluorescence imaging of classical Cajal Body (Coilin) and nuclear gem (SMN) markers in mouse ES cells (left) and HEK293T cells (right). Cajal bodies are traditionally defined by the presence of Coilin foci in the nucleus (Machyna et al., 2015; Nizami et al., 2010; Ogg and Lamond, 2002) and based on this definition, our mES cells do not contain visible Cajal bodies with multiple 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.

2.3.4 RNA PROCESSING COMPARTMENTS ARE DEPENDENT ON 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 [46]), 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 this, we treated cells with actinomycin D (ActD), a drug that inhibits RNA Pol I and Pol II transcription [47], for 4 hours and performed RD-SPRITE (**Figure 3A, S3A**). We confirmed that ActD treatment led to robust inhibition of various nascent RNAs (e.g., 45S, histone mRNAs), but did not impact the steady-state RNA levels of their associated diffusible ncRNAs (snoRNAs, U7, scaRNAs) (**Figure 3B, S3B-C**). Next, we explored the spatial organization of DNA and RNA. Strikingly, while we did not observe structural changes of most DNA structural features (e.g., chromosome territories, A/B compartments, **Figure S3I**), we observed large-scale disruption of DNA and RNA organization within the nuclear structures associated with ribosome, snRNA, and histone biogenesis.

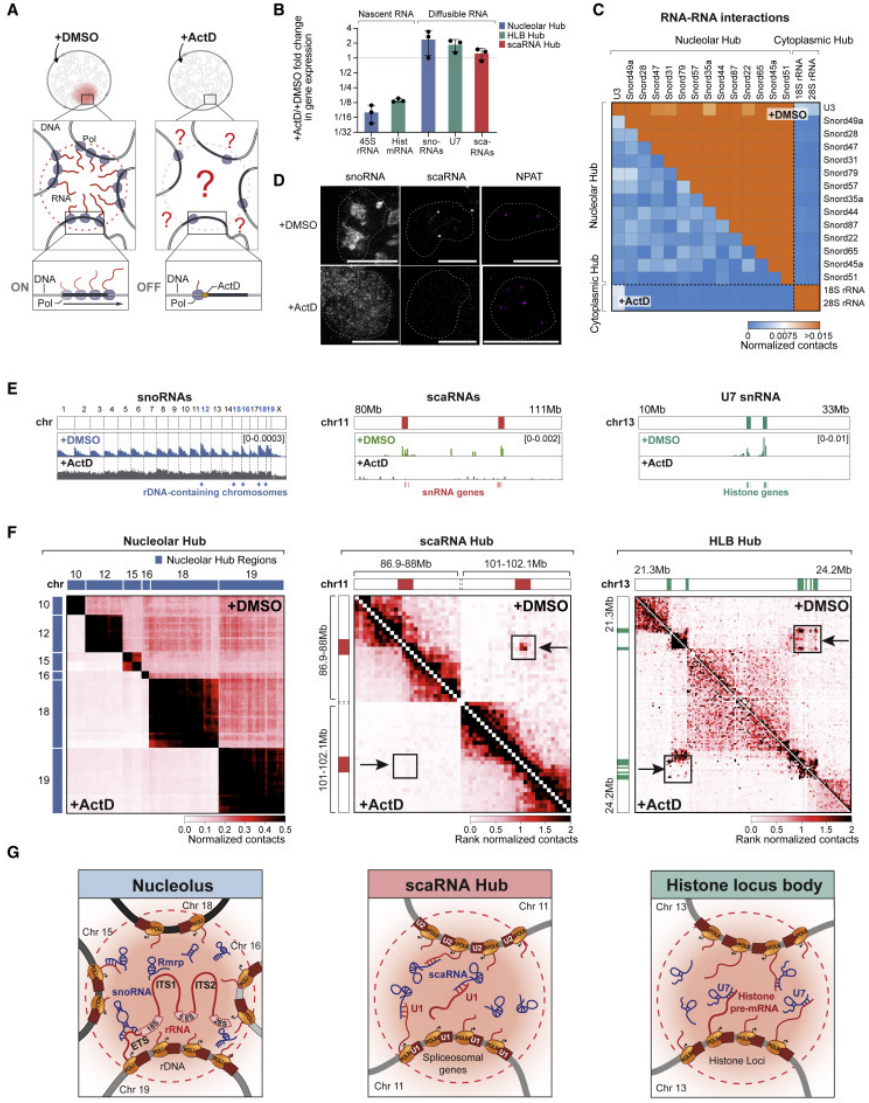


Figure 3. Inhibition of nascent RNA transcription disrupts RNA processing hubs. *(A) Schematic of transcriptional inhibition of Pol I and Pol II in cells treated with actinomycin D (+ActD) or control (+DMSO). (B) Gene expression changes of RNAs of interest following ActD treatment. Error bars represent standard deviation of three replicate experiments. (C) Unweighted RNA-RNA contact frequency of snoRNAs and rRNAs following ActD (bottom) or DMSO (top) treatment. (D) Imaging of snoRNA, scaRNA, or*

NPAT protein upon ActD or DMSO treatment. Scale bar represents 10 μm. (E) Weighted RNA-DNA contacts upon DMSO (top) or ActD (bottom) treatment for aggregated snoRNAs (left, cluster size 1,001–10,000), scaRNAs (middle), and U7 (right). (F) DNA-DNA contact matrices upon ActD (bottom) or DMSO (top) treatment. (Left) Nucleolar-hub-associated genomic regions (previously described in Quinodoz et al., 2018). (Middle) Two regions on chromosome 11 containing snRNA clusters. (Right) Region on chromosome 13 containing histone gene clusters. (Middle right) Rank normalized contacts are defined by rescaling contact frequency based on their rank-order to enable comparison between samples. (G) Model of how nascent transcription of RNA organizes diffusible ncRNAs and genomic DNA to form each hub.

Focusing on the nucleolar hub, we observed a strong depletion of RNA-RNA contacts between the various snoRNAs (**Figure 3C**) and global disruption of snoRNA localization at nucleolar DNA sites (**Figure 3D-E, S3D**) such that snoRNA and RMRP localization became diffusive throughout the nucleus (**Figure 3D, S3E,H**). We also observed a dramatic reduction in inter-chromosomal contacts between genomic DNA regions contained within the nucleolar hub (**Figure 3F, S3G**). These results indicate that transcription of 45S pre-rRNA (which is known to interact with snoRNAs and RNase MRP [48,49] acts to concentrate these diffusible ncRNAs and organize DNA loci into the nucleolar compartment (**Figure 3G**).

Similarly, ActD treatment led to a loss of focal localization of scaRNAs at snRNA genes (**Figure 3E, S3D**), a change from focal to diffusive localization throughout the nucleus (**Figure 3D**), and a striking reduction in the long-range DNA-DNA contacts between snRNA genes (**Figure 3F, S3G**). In addition, we observed a loss of focal localization of U7 at the histone genes (**Figure 3E, S3D**), loss of long-range DNA-DNA interactions between the histone loci (**Figure 3F**), and an increase in the number of nuclear foci containing HLB-associated proteins (NPAT) within each cell (**Figure 3D, S3F**). These results indicate that nascent transcription of snRNAs and histone pre-mRNAs is required to drive organization of these nuclear compartments (**Figure 3G**).

Although we did not observe major changes in DNA-DNA or RNA-DNA contacts within the splicing hub, this may be because ActD only led to a

modest reduction (<2-fold) in nascent pre-mRNA (introns) levels (**Figure S3A**). Consistent with this, we previously observed significant changes in snRNA localization at active DNA sites following treatment with flavopiridol (FVP) [29], a transcriptional inhibitor that leads to robust reduction of nascent pre-mRNA levels.

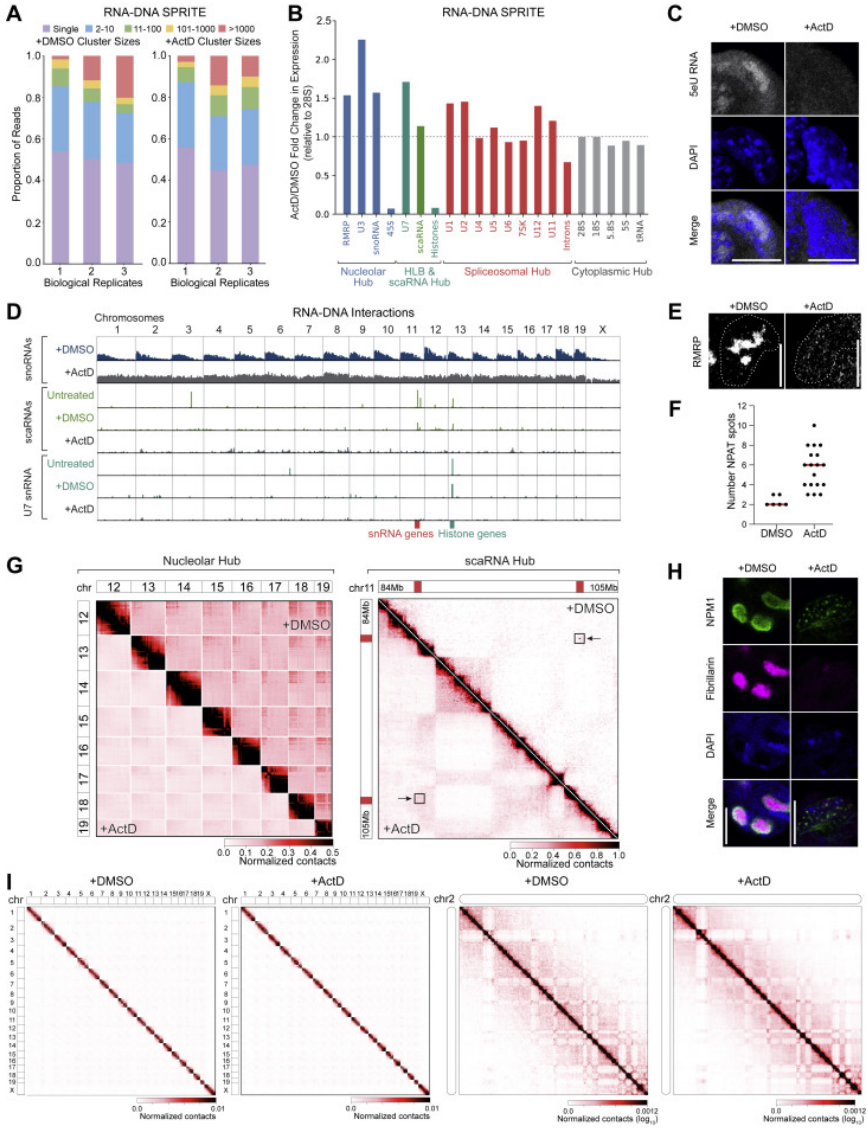


Figure S3. Transcriptional inhibition with actinomycin D leads to structural changes in the nucleolar hub, scaRNA hub, and HLB hubs. *(A) Cluster size distribution in RD-SPRITE for DMSO-treated (left) and ActD-treated (right) samples. Independent results from three biological replicates are shown. (B) 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 STAR Methods). (C) Microscopy image of nascent RNA in DMSO-treated cells or ActD-treated cells. Nascent transcription was visualized by incubating cells with 5EU (see STAR 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, gray) or control treatment (+DMSO, blue). Contacts for top-expressing snoRNAs in SPRITE clusters of size 1001-10000 reads were aggregated (see STAR Methods) (Middle) DNA localization for scaRNAs following ActD transcriptional inhibition (+ActD, gray) or control treatment (+DMSO, green). (Bottom) DNA localization of U7 snRNA following ActD transcriptional inhibition (+ActD, gray) 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. Scalebar is 10 μm. (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 STAR 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 STAR Methods). (H) IF stain for NPM1 (green), 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.

2.3.5 SATELLITE-DERIVED NCRNAS ORGANIZE HP1 LOCALIZATION AT INTER-CHROMOSOMAL HUBS

In addition to RNA processing, we identified a hub containing ncRNAs transcribed from minor and major satellite DNA regions within centromeric and pericentromeric regions, respectively (**Figure 1D**). We

found that these ncRNAs localize primarily over centromere-proximal regions (**Figure 4A-B, S4B**) and organize into higher-order structures containing these ncRNAs and multiple centromere-proximal regions from different chromosomes (**Figure 4C, S4A**). To confirm this, we performed DNA FISH on the major and minor satellite DNA and observed higher-order structures where multiple centromeres interact simultaneously (**Figure 4D**), indicating that satellite-derived ncRNAs demarcate nuclear compartments where centromeric regions from multiple chromosomes associate with each other.

Centromeric and pericentromeric DNA (chromocenters) are enriched for various heterochromatin enzymes and chromatin modifications, including the HP1 protein and H3K9me3 modifications [50]. Previous studies have shown that global disruption of RNA by RNase A leads to disruption of HP1 localization at chromocenters [50]. However, RNAse A is not specific and can impact several structures in the nucleus, including nucleoli [51]. Because major and minor satellite-derived ncRNAs localize exclusively within centromere-proximal structures, we hypothesized that these ncRNAs might be important for HP1 localization. To test this, we used an antisense oligonucleotide (ASO) to degrade either the major or minor satellite RNAs (**Figure S4C-D**) and observed depletion of HP1 proteins over these centromere-proximal structures (**Figure 4E-F, S4E**) without impacting overall HP1 protein levels (**Figure S4F**). Because disruption of the major satellite RNAs also led to reduced minor satellite RNA levels (**Figure S4C-D**), we cannot exclude that altered HP1 localization is solely due to depletion of minor satellite RNA.

Our results demonstrate that satellite-derived ncRNAs are enriched close to their transcriptional loci and recruit HP1 into centromere-proximal nuclear compartments (**Figure 4G**). Consistent with this, previous studies have shown that disruption of the major satellite-derived RNA prior to the formation of chromocenters during preimplantation development leads to loss of chromocenter formation, lack of heterochromatin formation, and embryonic arrest [52].

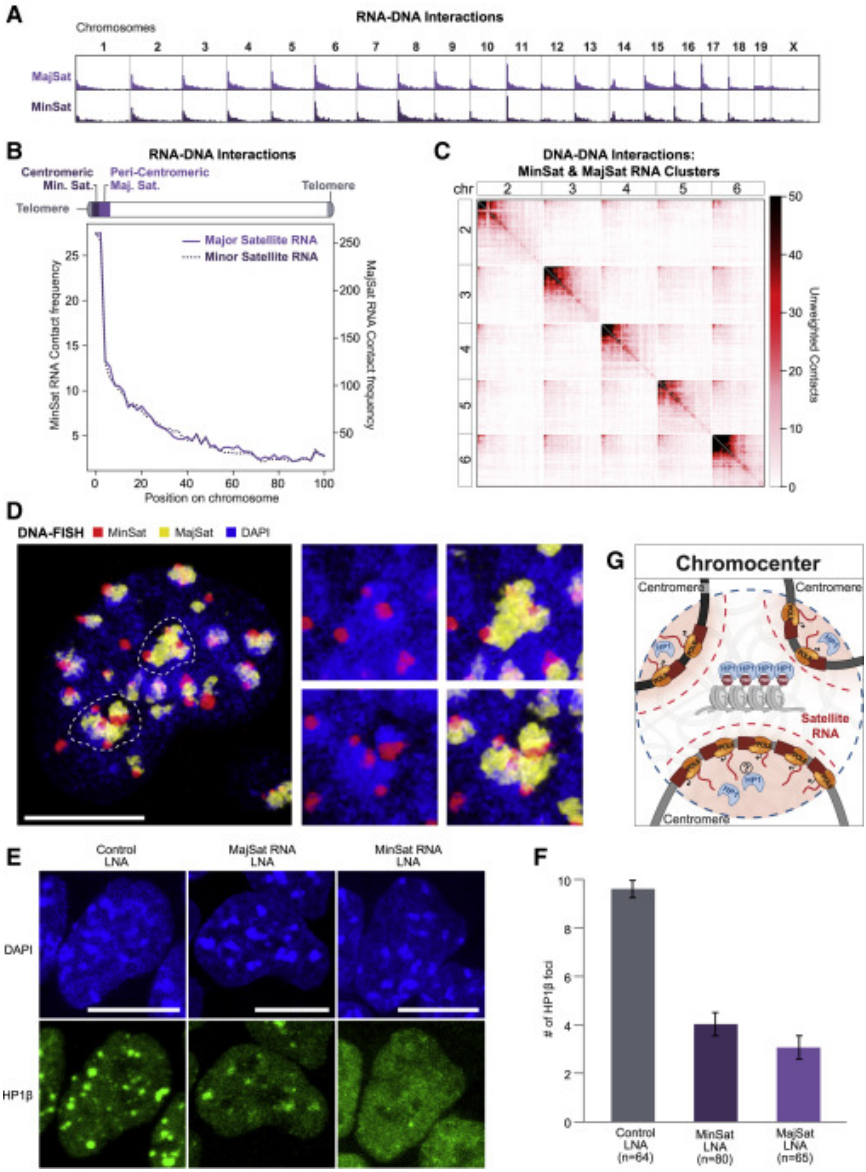


Figure 4. Satellite-derived ncRNAs organize HP1 at inter-chromosomal hubs. *(A and B) Unweighted RNA-DNA contact frequencies of major and minor satellite-derived ncRNAs (A) across the genome or (B) aggregated across all chromosomes. (C) Unweighted DNA-DNA contacts for chromosomes 2 to 6 within clusters containing a satellite-derived RNA.*

(D) DNA FISH of major (yellow) and minor (red) satellite DNA in the nucleus (DAPI, blue). Dashed lines demarcate the two DAPI-dense structures shown as zoom-ins on the right. Scale bar represents 10 μm. (E) HP1β IF following locked nucleic acid (LNA)-mediated knockdown of major (MajSat) and minor (MinSat) satellite-derived RNAs. Scale bar represents 10 μm. (F) Quantification of the mean number of HP1β foci per cell following LNA knockdown. n = number of cells analyzed; error bars represent standard error. (G) Schematic of chromocenter hub. Satellite RNAs are spatially concentrated (red gradient) near centromeric DNA. Individual centromeres assemble into a heterochromatic chromocenter structure highly enriched with HP1 protein.

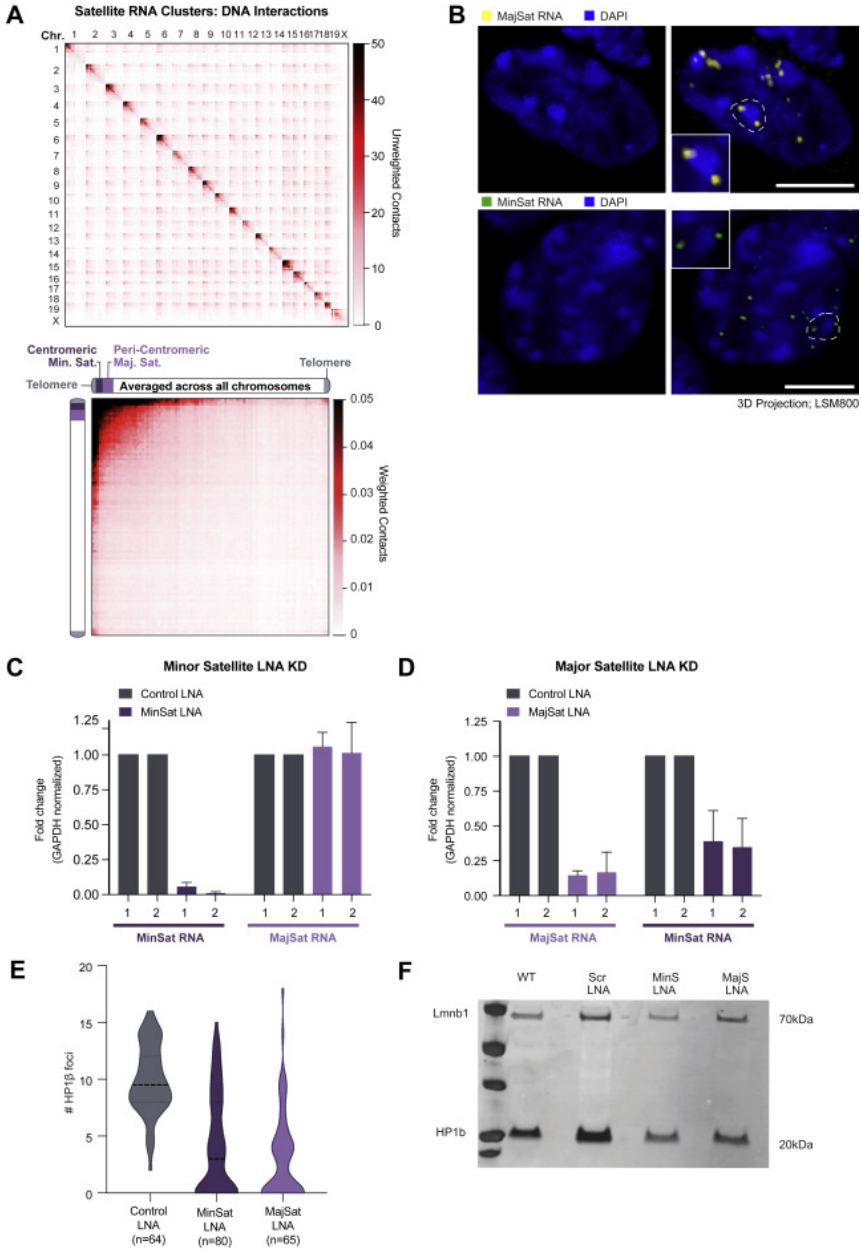

A
Satellite RNA Clusters: DNA Interactions
Chr. 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Unweighted Contacts
Centromeric Min. Sat.
Peri-Centromeric Maj. Sat.
Telomere
Telomere
Averaged across all chromosomes
Weighted Contacts
B
MajSat RNA
DAPI
MinSat RNA
DAPI
3D Projection; LSM800
C
Minor Satellite LNA KD
Control LNA
MinSat LNA
Fold change (GAPDH normalized)
MinSat RNA
MajSat RNA
D
Major Satellite LNA KD
Control LNA
MajSat LNA
Fold change (GAPDH normalized)
MajSat RNA
MinSat RNA
E
HP1β foci
Control LNA (n=64)
MinSat LNA (n=80)
MajSat LNA (n=65)
F
WT
Scr LNA
MinS LNA
MajS LNA
Lmnb1
HP1b
70kDa
20kDa

Figure S4. Satellite-derived ncRNAs mediate higher-order heterochromatin organization at centromeric clusters. *(A) (Top) Unweighted, genome-wide DNA-DNA contact matrices constructed from SPRITE clusters containing minor or major satellite RNAs. (Bottom) Weighted, inter-chromosomal DNA-DNA contact matrices averaged over all chromosomes from SPRITE clusters containing minor and major satellite RNA. DNA-DNA contacts occurring between regions on all pairs of chromosomes (1 through X) were computed, averaged, and plotted as an aggregate heatmap (see STAR Methods). (B) RNA FISH images of either MajSat RNA (top, yellow) or MinSat RNA (bottom, green). DAPI (blue) only images are shown on the left; merged images are on the right. Dashed lines and corresponding inset boxes zoom in on a single DAPI-dense chromocenter structure. Scalebar is 10μm. (C) Quantification of major and minor satellite RNA gene expression changes following LNA knockdown for minor satellite RNA (2 primer sets) compared to control LNA. Error bars represent standard deviation across 3 biological replicates. (D) Quantification of major and minor satellite RNA gene expression changes following LNA knockdown for major satellite RNA (2 primer sets) compared to control LNA. Error bars represent standard deviation across 3 biological replicates. (E) Quantification of the number of HP1β foci per cell shown in Figure 5E depicted as a violin plot. Control: n = 64 cells, MinSat LNA: n = 80 cells, MajSat LNA: n = 65 cells. (F) Western blot for Lmnb1 protein and HP1β protein in untreated (WT), scramble LNA (scr LNA), Minor Satellite-targeting LNA (MinS LNA) or Major Satellite-targeting LNA (MajS LNA)-treated cells.*

2.3.6 HUNDREDS OF NON-CODING RNAS LOCALIZE IN SPATIAL PROXIMITY TO THEIR TRANSCRIPTIONAL LOCI

Thousands of nuclear-enriched ncRNAs are expressed in mammalian cells, but only a handful have been mapped on chromatin. We mapped ~650 lncRNAs in ES cells and observed a striking difference in chromatin localization between these and mature mRNAs (**Figure 5A, S5A-B,** see **Methods**). Specifically, we found that the vast majority (93%) of the lncRNAs are strongly enriched within 3D proximity of their transcriptional loci (**Figure 5B-D, S5C**). This is consistent with previous microscopy measurements that showed that most lncRNAs measured form enriched foci in the nucleus [53]. In contrast, we find that mature mRNAs are depleted near their transcriptional loci and at all other genomic locations (chromatin enrichment score <0), consistent with their localization in the cytoplasm (**Figure 5A, S5B,D-E**). We observed a similar lack of chromatin enrichment for a subset of lncRNAs, including

Norad which functions in the cytoplasm (**Figure 5A-B)** [54]. Additionally, not all lncRNAs with high chromatin enrichment are restricted to the 3D territory around their locus. For example, Malat1 is strongly enriched on chromatin but localizes broadly across all chromosomes (**Figure 5A-B, S5C**).

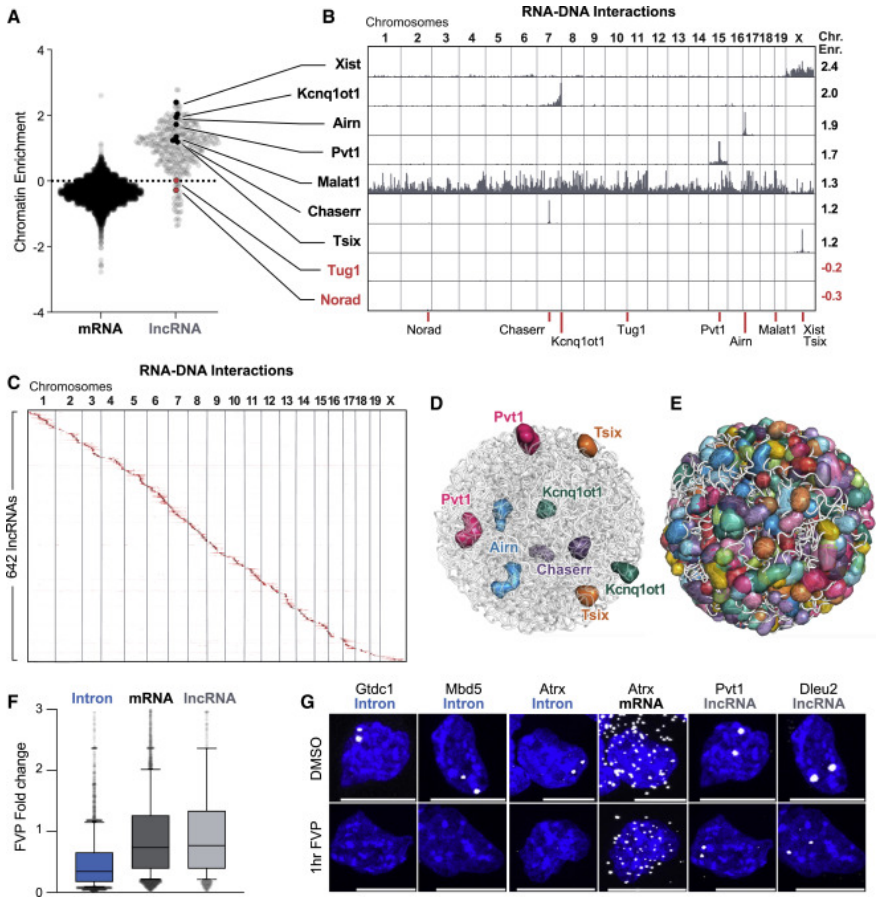


Figure 5. Most lncRNAs localize at genomic targets in 3D proximity to their transcriptional loci. *(A) Chromatin enrichment score for mRNAs and lncRNAs. Values >0 and <0 represent RNAs enriched and depleted on chromatin, respectively. (B) Unweighted RNA-DNA localization maps for selected chromatin-enriched (black) and chromatin-depleted (red) lncRNAs. Chromatin enrichment scores (Chr. Enr.) are listed (right). Red lines (bottom) indicate transcriptional locus for each RNA. (C) Unweighted RNA-DNA*

localization map of 642 lncRNAs ordered by genomic position of their transcriptional loci. (D and E) 3D space filling nuclear structure model of the selected lncRNAs (D) or 543 lncRNAs (E) that display at least 50-fold enrichment in the nucleus. Each sphere corresponds to a 1-Mb region or larger where an individual lncRNA is enriched. (F) Change in RNA levels between untreated and flavopiridol (FVP)-treated mouse ESCs (Jonkers et al., 2014) for introns, mRNAs, and lncRNAs. Plot: line represents median, box extends from 25th to 75th percentiles, and whiskers from 10th to 90th percentiles. (G) RNA FISH for selected introns, mRNA exons, and lncRNAs following FVP (bottom) or DMSO (top) treatment for 1 h. Scale bar represents 10 μm.

Localization of lncRNAs in proximity to their transcriptional loci could represent either unstable RNA products transiently associated with their transcriptional loci prior to degradation (consistent with nascent pre-mRNA localization [55]) or stable association of mature RNAs after transcription (**Figure S5A**). To test whether they represent transient RNA products, we measured the expression of lncRNAs after FVP treatment. We explored a previously published RNA sequencing experiment performed after 50 minutes of treatment with FVP in mES cells [56]. Consistent with previous reports [57], we found that virtually all lncRNAs were dramatically more stable than nascent pre-mRNAs and comparable in stability to mature mRNAs (**Figure 5F**). To confirm this, we performed RNA FISH for 4 lncRNAs, 6 nascent pre-mRNAs (introns), and 1 mature mRNA (exons) in untreated cells and upon FVP treatment. We found that all of these lncRNAs form stable nuclear foci that are retained upon transcriptional inhibition (**Figure 5G, S5F**). In contrast, all nascent pre-mRNA foci are lost upon transcriptional inhibition, even though we observe no impact on their mature mRNA products (**Figure 5G**).

Together, these results demonstrate that many hundreds of lncRNAs form high concentration spatial territories throughout the nucleus (**Figure 5E**).

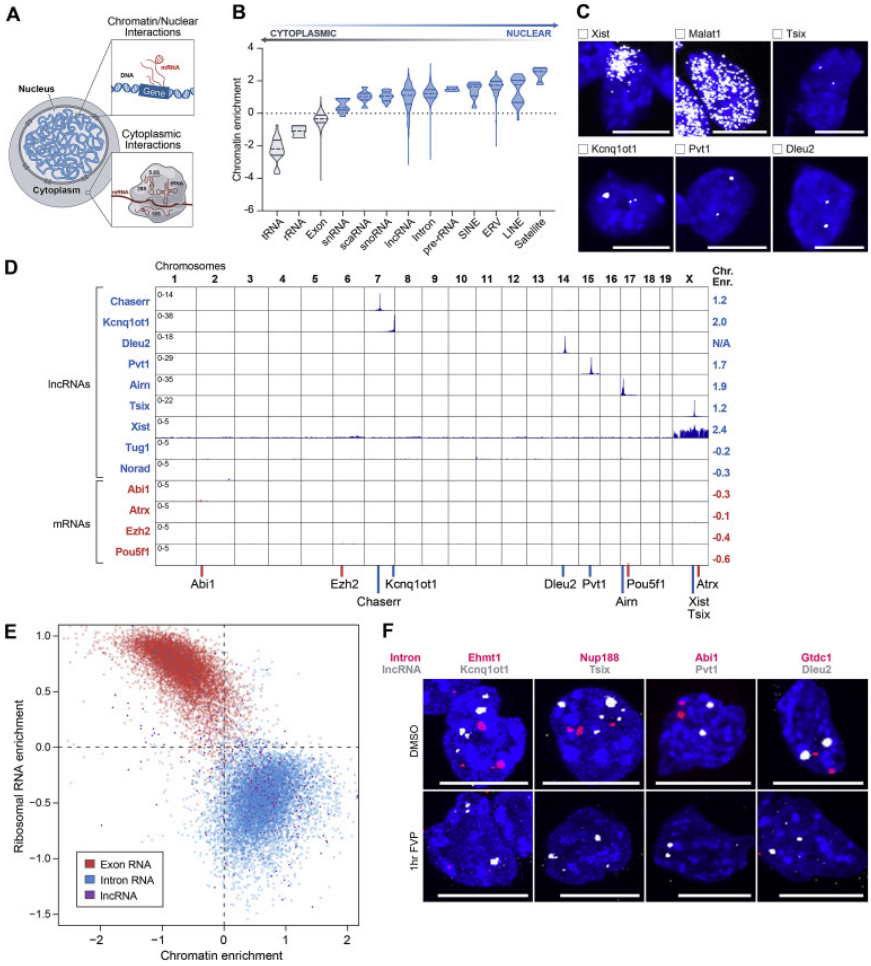


Figure S5. Many lncRNAs localize within 3D proximity to their transcriptional loci in the nucleus. *(A) Schematic illustration of our chromatin enrichment score which computes the frequency of an RNA interaction with chromatin (top inset) compared to the frequency of interactions without chromatin, such as interactions with rRNA, tRNA, and mRNA in the cytoplasm (bottom inset). (B) Chromatin enrichment score for multiple classes of RNAs. tRNAs, rRNAs, and exons are predominantly depleted on chromatin (enrichment score < 0) versus other classes of RNAs, including introns, scaRNAs, lncRNAs, are enriched on chromatin (enrichment score > 0). (C) RNA FISH localization patterns of multiple lncRNAs (Xist, Malat1, Tsix, Kcnq1ot1, Pvt1, and Dleu2 lncRNAs) in the nucleus (DAPI). Scalebar is 10μm. (D) Genome-wide normalized RNA-DNA interactions for several lncRNAs (blue) and mRNAs (red). Each RNA locus is demarcated at the bottom. (E)*

Chromatin enrichment scores (x-axis) versus ribosomal RNA enrichment scores (y-axis) for exons (red), introns (blue), and lncRNAs (purple). (F) RNA FISH for 4 mRNA introns and 4 lncRNAs treated for 1 hour with DMSO (top) or FVP (bottom). As a control, we co-stained lncRNAs (white) and introns (red) within the same cell. Scalebar is 10μm.

2.3.7 NON-CODING RNAS GUIDE REGULATORY PROTEINS TO NUCLEAR TERRITORIES TO REGULATE GENE EXPRESSION

Because hundreds of lncRNAs are enriched in territories throughout the nucleus, we explored whether RNAs might impact protein localization within these territories. Recently, we and others showed that SHARP (also called Spen) directly binds Xist [58,59] and recruits the HDAC3 histone deacetylase complex to the X chromosome to silence transcription [58,60] **Figure 6A**, **S6A**). To explore the nuclear localization of SHARP more globally, we performed super-resolution microscopy and found two types of localization: low-level diffusive localization throughout the nucleus and compartmentalized localization within dozens of well-defined foci (~50-100 foci/nucleus; **Figure 6B, Video S1**). To determine whether the SHARP foci are dependent on RNA, we deleted the RNA binding domains from SHARP (ΔRRM) and visualized its localization (**Figure 6A**). We observed diffuse localization of the mutant protein and loss of all compartmentalized SHARP foci (**Figure 6B, Video S2**) even though there was no change in overall SHARP protein levels (**Figure S6B**). These results demonstrate that RNA is required for SHARP localization to dozens of spatial territories throughout the nucleus.

To explore whether these ncRNA-mediated territories might act to regulate gene expression, we purified SHARP and mapped its interactions with specific RNAs. We identified strong binding to several RNAs, including a ~600 nucleotide region at the 5' end of Kcnq1ot1 (**Figure 6C**), a lncRNA that leads to parental imprinting of several genes within the Cdkn1c locus and is associated with the pediatric Beckwith-Wiedemann overgrowth syndrome [61]. We found that Kcnq1ot1 localizes within the topologically associating domain (TAD) that contains all of the known imprinted genes (Kcnq1, Cdkn1c, Slc22a18, Phlda2; [61,62], but excludes other genes that are linearly close in the genome (e.g. Cars, Nap1l4;

Figure 6D). We hypothesized that Kcqn1ot1 acts to guide SHARP to this territory. To test this, we induced Kcnq1ot1 expression and measured the concentration of SHARP over the two distinct alleles: the allele expressing the Kcnq1ot1 RNA (+Kcnq1ot1) and the allele lacking it (-Kcnq1ot1). We observed an enriched focus of SHARP only over the +Kcqn1ot1 allele (**Figures 6E and S6C, Video S3**). This demonstrates that Kcnq1ot1 localization acts to recruit SHARP to a precise territory.

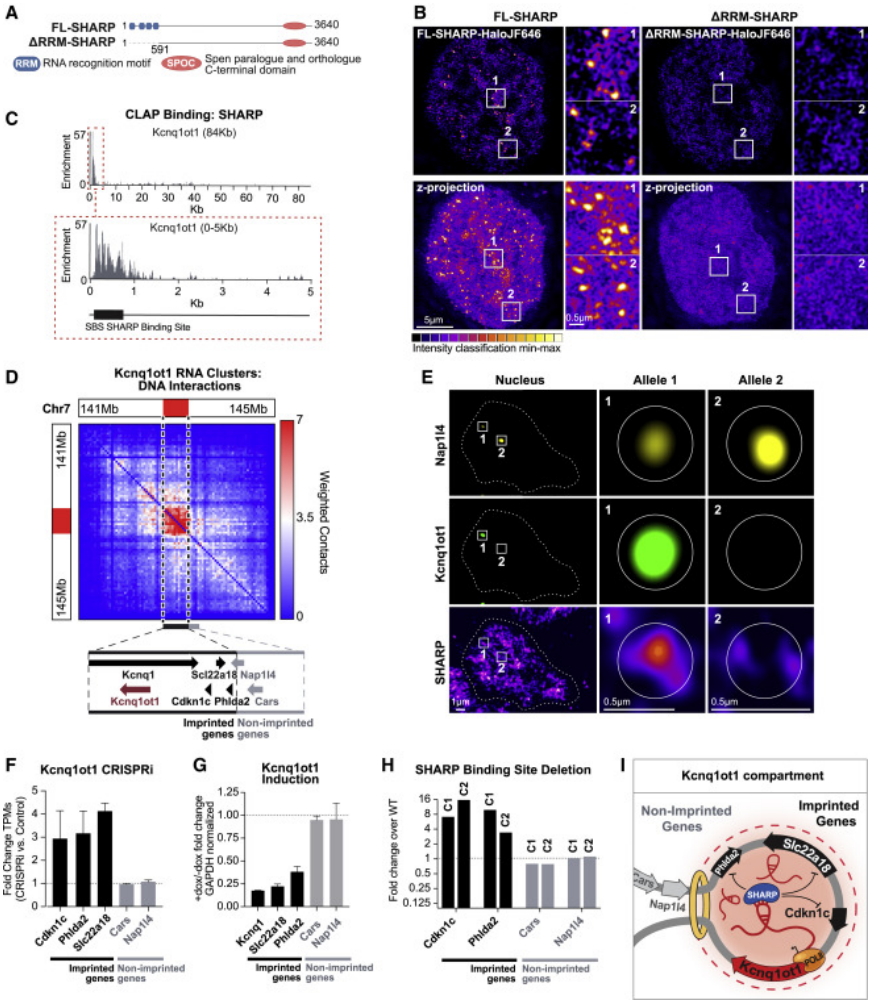


Figure 6. SHARP is enriched within dozens of RNA-mediated compartments in the nucleus and can regulate gene expression within specific compartments. *(A) Full-length (FL) SHARP (also referred to as Spen) contains four RNA recognition motif (RRM, blue) domains and one Spen paralog and ortholog C-terminal (SPOC, orange) domain. SHARP lacking its RNA-binding motifs (ΔRRM) was generated by deleting the first 591 amino acids. (B) 3D-structured illumination microscopy (3D-SIM) intensity of Halo-tagged FL-SHARP (left) and ΔRRM-SHARP (right). Shown are 125-nm optical sections (top) and z-projections (bottom). FL-SHARP localizes in foci throughout the nucleus (zoom in panels 1 and 2), while ΔRRM-SHARP localization is more diffuse. Bar: 5 μm; insets: 0.5 μm. (C) SHARP-binding profile to Kcnq1ot1 including its SHARP-binding site (SBS; black box)*

mapped using covalent linkage and affinity purification (CLAP). (D) Weighted DNA-DNA contacts within clusters containing Kcnq1ot1 RNA. Dashed line indicates the location of the Kcnq1ot1-enriched territory. (Zoom box) Genomic locations of the Kcnq1ot1 gene (burgundy), the imprinted Kcnq1, Slc22a18, Cdkn1c, and Phlda2 (black), and non-imprinted Nap1l4 and Cars (gray) genes. (E) RNA FISH combined with IF of Nap1l4 RNA, Kcnq1ot1 RNA, and SHARP. Maximum intensity z-projections (left) are shown alongside individual z-section slices of the actively transcribed Kcnq1ot1 allele (center) and the inactive Kcnq1ot1 allele (right). Scale bars represent 1 μm (left) and 0.5 μm (center, right). (F) Changes in gene expression upon CRISPR interference (CRISPRi) of Kcnq1ot1. Error bars represent standard deviation between two biological replicates. (G) Changes in gene expression with or without induction of Kcnq1ot1 (+dox/–dox). Error bars represent standard deviation between two biological replicates. (H) Comparison of gene expression between two clonal lines lacking the Kcnq1ot1 SBS to wild-type cells. (I) Model of how Kcnq1ot1 seeds the formation of an RNA-mediated compartment in spatial proximity to its transcriptional locus. After transcription, Kcnq1ot1 binds and recruits the SHARP protein into this compartment to silence imprinted target genes.

To explore the functional contribution of this Kcnq1ot1-mediated SHARP territory, we downregulated Kcnq1ot1 using CRISPRi and observed specific upregulation of genes within the Kcnq1ot1-localized territory (**Figure 6F**). Conversely, induction of Kcnq1ot1 expression led to silencing of these target genes (**Figure 6G**). In both cases, there was no impact on the genes outside of this Kcnq1ot1-localized domain (**Figure 6F-G, S6H**). To determine if SHARP binding to Kcnq1ot1 RNA is essential for Kcnq1ot1-mediated transcriptional silencing, we deleted the SHARP binding site on Kcnq1ot1 (ΔSBS) and observed upregulation of its known target genes in two independent clones (**Figure 6H, S6D-E**). Because SHARP is known to recruit HDAC3 [58], we tested whether induction of Kcnq1ot1 leads to a reduction of histone acetylation over this territory. We performed ChIP-seq against H3K27ac and observed depletion specifically over the imprinted cluster upon Kcnq1ot1 induction (**Figure S6F**). Moreover, we tested whether histone deacetylase activity is required for Kcnq1ot1-mediated silencing by treating cells with a small molecule that inhibits HDAC activity (TSA) and observed specific loss of Kcnq1ot1-mediated silencing of its target genes (**Figure S6G**). Together, these results demonstrate that Kcnq1ot1 localizes at a high concentration within the TAD containing its transcriptional locus, binds directly to

SHARP, and recruits SHARP and its associated HDAC3 complex to silence transcription of genes within this nuclear territory (**Figure 6I**).

We also identified several other lncRNAs that localize within specific nuclear territories around their transcriptional loci containing their functional targets. For example: **(i)** Airn localizes within a TAD containing its reported imprinted target genes [63] but excludes other neighboring genes (**Figure S6I**); **(ii)** Pvt1 localizes to a TAD containing Myc and multiple enhancers of Myc (**Figure S6J**) and has been shown to repress Myc expression [64]; **(iii)** Chaserr localizes within the TAD containing Chd2 (**Figure S6K**) and has been shown to repress Chd2 expression [65,66].

These results demonstrate that the localization pattern of a lncRNA in 3D space can act to guide recruitment of regulatory proteins to specific nuclear territories and highlights an essential role for these lncRNA-enriched nuclear territories in gene regulation.

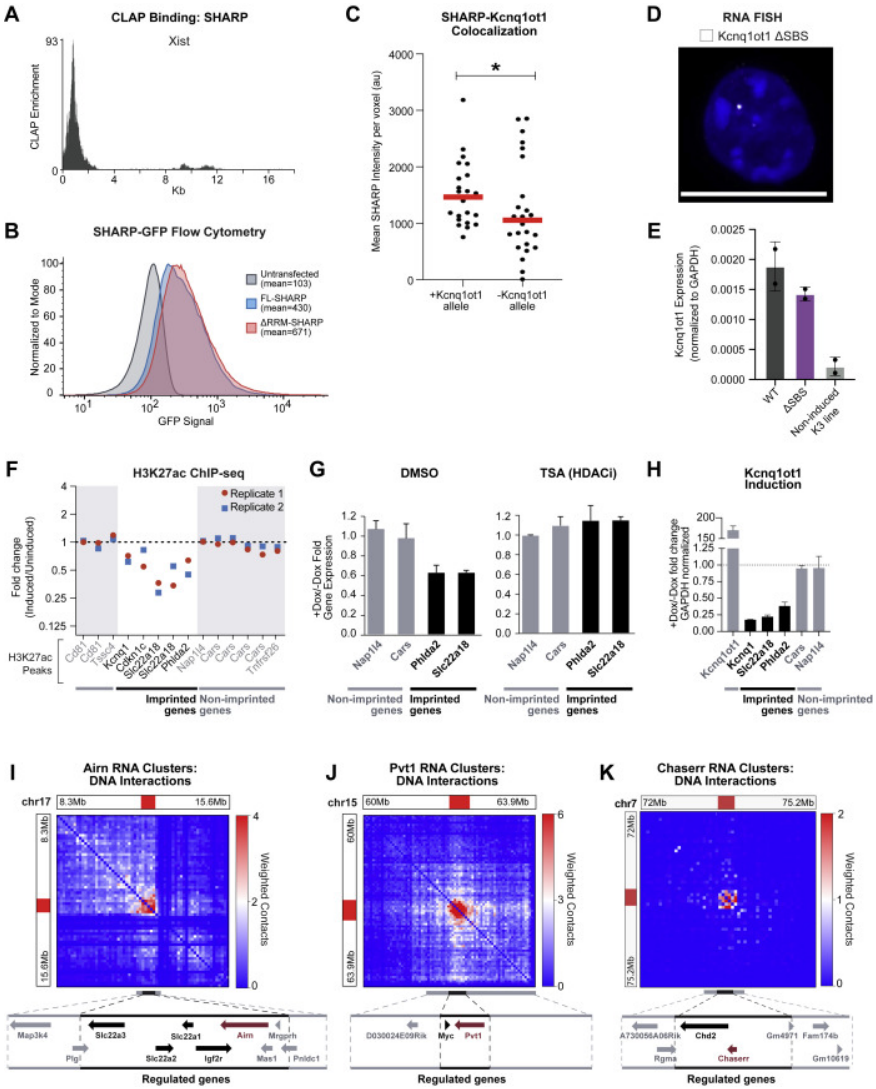


Figure S6. lncRNAs regulate target gene expression precisely with high-concentration territories in the nucleus. *(A) CLAP binding profile of SHARP protein to the Xist lncRNA. SHARP particularly binds at the 0-2kb region of Xist. (B) Detection of GFP-tagged FL-SHARP (blue) or ΔRRM-SHARP (red) protein expression by flow cytometry. (C) Quantification of SHARP localization on the Kcnq1ot1-expressing allele (left) versus the non-expressing allele (right) for images in Figure 6E. Red bar indicates mean intensity. ∗ indicates a p value of less than < 0.05 by Kolmogorov-Smirnov statistical test. Allele 1: n = 22 cells, Allele 2: n = 24 cells. (D) RNA FISH of Kcnq1ot1 in cell lines genetically*

engineered to delete the internal SHARP-Binding Site (ΔSBS) in Kcnq1ot1, Scalebar is 10μm. (E) Relative Kcnq1ot1 RNA expression in induced cells with the dox-inducible Kcnq1ot1 promoter (Kcnq1ot1 WT), induced cells lacking the SHARP-binding site (Kcnq1ot1 ΔSBS) or non-induced cells (non-induced K3 cells). Bars depict the mean of three primer sets. Error bars represent standard deviation across two biological replicates. (F) Relative promoter H3K27 acetylation (H3K27ac) in Kcnq1ot1-expression induced versus non-induced cells. Fold-change in enrichment is computed at all H3K27ac peaks for imprinted genes (black) and non-imprinted genes (gray). ChIP-seq results from two biological replicates are shown in red and blue, respectively. (G) Mean gene expression differences of Kcnq1ot1-regulated and Kcnq1ot1-non-regulated genes between induced (+Dox) and non-induced (-Dox) samples treated with DMSO (left) or the HDAC inhibitor, Trichostatin A (TSA) (right) (see STAR Methods). Error bars represent standard deviation. (H) Gene expression fold-change upon dox-induction of Kcnq1ot1 for Kcnq1ot1-regulated and Kcnq1ot1-non-regulated genes. Regulated genes (black) show robust repression while unregulated genes not within the imprinted TAD (gray) show no change. Error bars represent standard deviation. (I) Weighted DNA-DNA interaction matrix for Airn RNA-containing SPRITE clusters showing Airn lncRNA localization on DNA in a region confined to the genes Airn is known to regulate (Rom et al., 2019). (J) Weighted DNA-DNA interaction matrix for Pvt1 RNA-containing SPRITE clusters showing Pvt1 lncRNA localization on DNA in a region occupied by Pvt1 and Myc genes. (K) Weighted DNA-DNA interaction matrix for Chaserr RNA-containing SPRITE clusters. Chaserr RNA is confined to a TAD containing the Chaserr gene and its known regulatory target, Chd2.

2.4 DISCUSSION

Our results demonstrate that ncRNAs can act as seeds to drive spatial localization of otherwise diffusive ncRNA and protein molecules. 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 cases, we observed a common theme where (i) specific RNAs localize at high concentrations in proximity to their transcriptional loci and (ii) diffusible ncRNA and protein molecules that bind to them are enriched within these structures. 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 thus act to dynamically change the distribution of diffusible molecules such that they become enriched within these territories ("recruit," **Figure 7**). By recruiting diffusible regulatory factors to multiple 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.

More generally, we showed that hundreds of nuclear ncRNAs are preferentially localized within precise territories in the nucleus, suggesting that RNA may represent a widespread class of molecules that act as seeds to drive spatial organization of diffusible molecules. This mechanism utilizes a unique role for RNA in the nucleus (relative to DNA or proteins): the process of transcription produces many copies of an RNA, which accumulate at high concentrations in proximity to their transcriptional locus. In contrast, proteins are translated in the cytoplasm and therefore lack positional information in the nucleus, and DNA is present at a single copy and therefore cannot achieve high local concentrations.

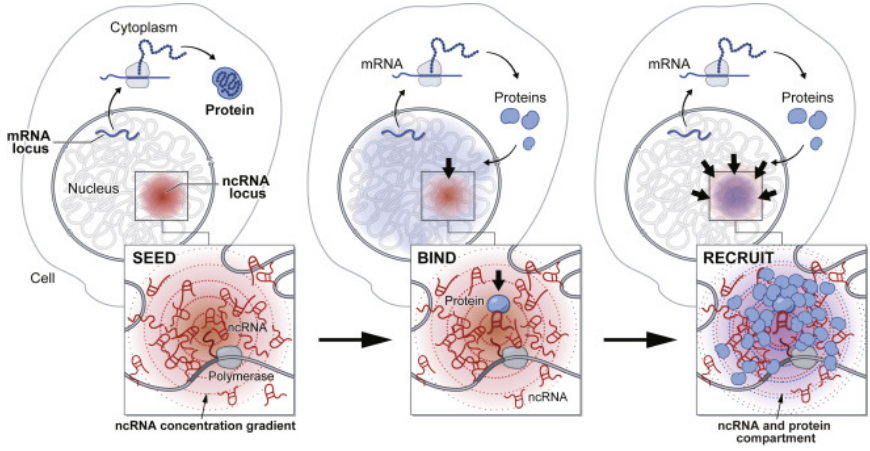


Figure 7. A model for the mechanism by which ncRNAs drive the formation of nuclear compartments. Once *transcribed, mRNAs are exported to the cytoplasm while ncRNAs are*

retained in the nucleus. ncRNA transcription creates a transcript concentration gradient, highest near its transcriptional locus (SEED, left panel). Because ncRNAs can bind with high affinity to diffusible RNAs and proteins immediately upon transcription (BIND, middle panel), they can concentrate other RNAs and proteins in a spatial compartment (RECRUIT, right panel). In this way, ncRNAs can drive the organization of nuclear compartments.

Central to this mechanism is the fact that ncRNAs can form high affinity interactions immediately following transcription and thus can recruit proteins and RNAs. In contrast, mRNAs require translation and therefore generally do not form stable interactions with regulatory molecules in the nucleus. Our results suggest that any RNA that functions independently of its translated product could act in this way. For example, we find that histone pre-mRNAs can seed organization of nuclear compartments even though their processed RNAs are also translated into protein products. Other nascent pre-mRNAs may also have protein-independent functions and form high-affinity interactions within the nucleus that are important for spatial organization. This seeding role for RNA might also contribute to the formation of other recently described nuclear compartments such as transcriptional condensates, which inherently produce high levels of RNAs, including enhancer-associated RNAs and pre-mRNAs. 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 locally constrained while others diffuse throughout 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 are required for a wide range of essential nuclear functions, including RNA processing, heterochromatin organization, and gene regulation (**Figure S7**). 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-

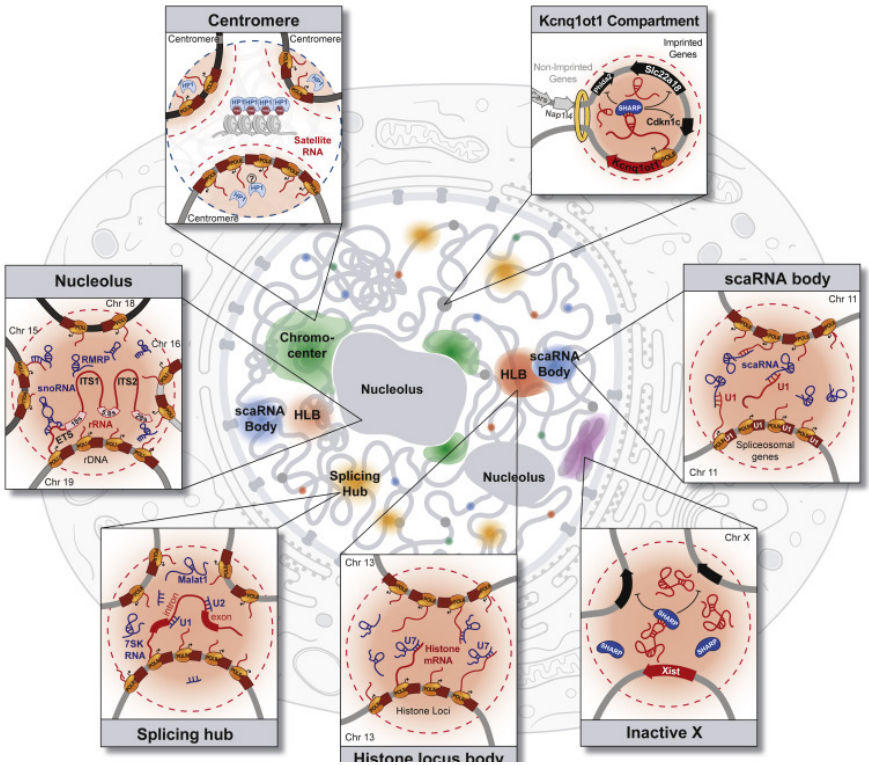


Figure S7. A widespread role for ncRNAs in shaping compartments throughout the nucleus that are associated with various nuclear functions. *A model schematic of the different nuclear compartments within the nucleus and the molecular components contained within them. In each of these cases, an RNA seeds 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.*

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.

2.5 LIMITATIONS OF THE STUDY

We note several technical limitations of the RD-SPRITE method. It requires crosslinking, which may lead to biases in the types of interactions that are detected. Because this approach takes a snapshot in time, it cannot measure dynamic events. 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 may act in this way, this mechanism may not hold true for every RNA. Future work is needed to explore the functional and mechanistic roles of individual ncRNAs.

2.6 METHODS

2.6.1 RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Mitchell Guttman (mguttman@caltech.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

- SPRITE datasets generated during this study have been deposited on GEO and are publicly available as of the date of publication at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151515. Accession numbers are listed in the key resources table.
- The original code for the SPRITE analysis pipeline used in this study is available on Github at https://github.com/GuttmanLab/sprite2.0-pipeline and https://github.com/GuttmanLab/sprite-pipeline. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

2.6.2 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 F1-21 mouse 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) pSM33 male ES cells (gift from K. Plath). These cells express Xist from the endogenous locus under control of a tetracycline-inducible promoter. (v) TX1072, a female mouse embryonic stem cell line (gift from E. Heard [69]). These cells express Xist from the endogenous locus under control of a tetracycline-inducible promoter. (vi) HEK293T, a

female human embryonic kidney cell line (ATCC Cat# CRL-3216, RRID:CVCL_0063).

Cell culture conditions. All mouse ES cell lines 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, and cells were grown. Fresh medium was replaced every 24-48 hours depending on culture density, and 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.

TX1072 mouse ES cells were grown on gelatin-coated flasks in serum-containing ES cell medium (high glucose DMEM (Sigma), 15% FBS (Gibco), 2 mM L-glutamine (Gibco), 1 mM sodium pyruvate (Gibco), 0.1 mM MEM non-essential amino acids (Gibco), 0.1 mM β-mercaptoethanol, 1000 U/mL leukemia inhibitory factor (LIF, Chemicon), and 2i (3 µM Gsk3 inhibitor CT-99021, 1 µM MEK inhibitor PD0325901). Cell culture media was changed daily.

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.

2.6.3 METHOD DETAILS

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 [70]. 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.

3D-SIM SHARP-Halo cell culture conditions. pSM33 cells were seeded in 4-well imaging chambers (ibidi) equipped with a high precision glass bottom and plasmids were transfected with lipofectamine 3000 24 hours prior to imaging according to the manufacturer's instructions. Addition of doxycycline 8hrs prior to imaging was performed to transiently induce full-length (FL) SHARP and ΔRRM-SHARP SHARP (also known as Spen) expression from the Sp22 clone as previously described [71]. The ΔRRM clone (SHARPΔ1-591) was generated using PIPE mutagenesis using the Sp22 Full Length entry clone as template. It was recombined with appropriate destination vectors using Gateway LR recombination. 1μM JF646 Halo ligand was introduced to the media for 30 min, washed-off twice with PBS and exchanged with fresh media which were incubated

for another 15 min. Live-cell 3D-SIM imaging was performed at 37C and 5% CO_2 in media without phenol red.

Doxycycline Inducible Kcnq1ot1 cell line development. The endogenous promoter of Kcnq1ot1 was CRISPR-targeted (nicking gRNA pairs ACAGATGCTGAATAATGACT and CACGTCACCAAGGTCTTGGT or GCAGCCACGACACTGTTGAT and GTCACCAAGGTCTTGGTAGG) to insert a TRE and minimal CMV promoter into the same cell line with integrated Tet transactivator (M2rtTA) used to generate Dox-inducible Xist (see above). Clones were screened for ablation of endogenous Kcnq1ot1 expression and upregulation of expression upon administration of doxycycline (**Supplemental Figure 6E,H**).

CRISPRi: dCas9-4XSID cell line generation. A catalytically dead Cas9 (dCas9) fused to 4 copies of the SID repressive domain (4XSID) expressed from an Ef1a promoter was integrated into a single copy locus in the genome (mm10 - chr6:86,565,487-86,565,506; gRNA sequence AATCTTAGTACTACTGCTGC) using CRISPR targeting (cells hereby referred to as dCas9-4XSID).

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 12-16 hours prior to RNA harvesting for qRT-PCR or induced for 24 hours prior to cell crosslinking with 1% formaldehyde for ChIP-seq.

Trichostatin (TSA) treatment. For HDAC inhibitor experiments, cells were treated with either DMSO (control) or 5μM TSA (Sigma T8552-1MG) in fresh 2i media or 2i media containing 2μg/ml doxycycline for induction of Kcnq1ot1 expression.

Flavopiridol (FVP) Treatment. FVP transcriptional inhibition was performed by culturing cells in FVP (Sigma F3055-1MG) or DMSO at 1 μM final concentration in 2i media for 1 hour.

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.

2.6.4 RNA & DNA-SPRITE

RD-SPRITE is an adaptation of our initial SPRITE protocol [27] with significant improvements to the RNA molecular biology steps that enable generation of higher complexity RNA libraries.

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 [27]. (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.

The approach was performed as follows:

Crosslinking, lysis, sonication, and chromatin digestion. *pSM44* mES 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. 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 supernatant was removed and 0.5M Tris pH 7.5 was added for 1 hour at 4°C to quench coupling. Beads were subsequently washed post coupling three times with 1mL of Modified RLT buffer and three times with 1mL of SPRITE Wash buffer.

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. DPM Ligation was performed using 11 μL of 4.5 μM DPM adaptor in a 250 μL reaction using Instant Sticky End Mastermix (NEB) at 20°C for 30 minutes with shaking. 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 using High Concentration T4 RNA Ligase I [72]. Briefly, beads were resuspended in a solution consisting of 30 µL 100% DMSO, 154 µL H_2O, and 20 µL of 20 µM RPM adaptor, heated at 65°C for 2 minutes to denature secondary structure of RNA and the RPM adaptor, then immediately put on ice. An RNA ligation master mix was added on top of this mixture consisting of: 40 µL 10x NEB T4 RNA Ligase Buffer, 4 µL 100mM ATP (NEB), 120 µL 50% PEG 8000 (NEB), 20 µL Ultra Pure H_2O, 6 µL Ribolock RNAse Inhibitor, 7 µL NEB T4 RNA Ligase, High Concentration (M0437M) for 24°C for with shaking 1 hour 15 minutes. 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 [27]). 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 were performed for 45min to 1 hour at 20°C. Specifically, each well contained the following: 2.4 µL well-specific 0.45 µM SPRITE tag (IDT), 6.4 µL custom SPRITE ligation master mix, 5.6 µL SPRITE wash buffer (described above), and 5.6 µL Ultra-Pure H_2O. For all SPRITE ligations, we make a custom SPRITE ligation master mix (3.125x concentrated) combining 1600 µL of 2x Instant Sticky End Mastermix (NEB; M0370),

600 µL of 1,2-Propanediol (Sigma-Aldrich; 398039), and 1000 µL of 5x NEBNext Quick Ligation Reaction Buffer (NEB; B6058S). 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.**

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/CCGATCTGGTGGTCT TT
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.

2.6.5 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[73] 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 [74] 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 [75]. Filtered reads were subsequently realigned to our custom collection of repeat sequences using Bowtie v2.3.5 [76], 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 [77]. RepeatMasker [78] defined regions with milliDev ≤ 140 along with blacklisted v2 regions were filtered out using Bedtools v2.29.0 [79].

SPRITE cluster file generation. RNA and DNA reads were merged, and a cluster file was generated for all downstream analysis. MultiQC v1.6 [80] 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.

2.6.6 MICROSCOPY IMAGING

3D-Structured Illumination Microscopy (3D-SIM): 3D-SIM super-resolution imaging was performed on a DeltaVision OMX-SR system (Cytiva, Marlborough, MA, USA) equipped with a 60x/1.42 NA Plan Apo oil immersion objective (Olympus, Tokyo, Japan), sCMOS cameras (PCO, Kelheim, Germany) and 642 nm diode laser. Image stacks were acquired with z-steps of 125 nm and with 15 raw images per plane. The raw data were computationally reconstructed with the soft-WoRx 7.0.0 software package (Cytiva, Marlborough, MA, USA) using a wiener filter set to 0.002 and channel-specifically measured optical transfer functions (OTFs) using an immersion oil with a 1.518 refractive index (RI). 32-bit raw datasets were imported to ImageJ and converted to 16-bit stacks.

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)[81,82].

RNA-FISH for FVP experiments. To compare the relative stability of lncRNAs and pre-mRNAs, we obtained intron FISH probes for targets of comparable gene length to lncRNAs. This was done to ensure that any differences in RNA stability upon FVP treatment are not due to differences in the time it takes to transcribe each RNA. Specifically, we obtained probes for pre-mRNAs that are 57.87kb (Nup188), 73.7kb (Mbd5), 99.8kb (Abi1), 129.7kb (Ehmt1),131.8kb (Atrx), and 297.2kb (Gtdc1) in length. For lncRNAs, we obtained probes for RNAs of lengths 53.4kb (Tsix), 79.5kb (Dleu2), 93.1kb (Kcnq1ot1), and 340kb (Pvt1).

RNA-FISH combined with immunofluorescence of SHARP at Kcnq1ot1 loci. Dox inducible Kcnq1ot1 mESC were cultured in dox for 24 hours and fixed for RNA-FISH against Kcnq1ot1 and Nap1l4 combined with immunofluorescence for SHARP. Images were acquired on a Zeiss LSM980 confocal microscope with 63x oil immersion objective lens using the Airyscan 2.0 detector. The number of z-slices and size of the image were determined based on Zeiss recommendations for optimal Airyscan 2.0 acquisition. All images were deconvoluted using ZEN Blue Software with the same settings and were analyzed using Imaris software. To visualize the locations of the two alleles, we used the spot detection module to identify 3D surfaces corresponding to either Nap1l4 or Kcnq1ot1 signals. Spots positive for Nap1l4 RNA but not Kcnq1ot1 are referred to as Kcnq1ot1- and spots positive for Kcnq1ot1 are referred to as Kcnq1ot1+ alleles in this manuscript. The same thresholds and size filters were used across all images and the determined 3D objects were of the same volumes. For quality control, we confirmed that the majority of cells only contained a single Kcnq1ot1 volume and filtered the few individual cells containing zero or two volumes. This ensured that we focus only on cells with monoallelic expression of Kcnq1ot1. We quantified fluorescence intensity in these 3D objects by taking the sum of intensity within those volumes across all channels and plotted the resulting values.

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 [83] 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.

2.6.7 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 [84]. 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 [27]. 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 [27].

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 [27]. 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.

2.6.8 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 [85]. 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 [85]) 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.

2.6.9 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.

2.6.10 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 *pSM44* mES cells. Briefly, treated cells were crosslinked, lysed, and sonicated as described for RNA-DNA SPRITE above. 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 as described above. For DPM adaptor ligation, a unique set of DPM adaptors (Plate 6) 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 (Odd, Even, Odd, Terminal tags). After split-and-pool barcoding, samples were aliquoted into 5% aliquots and reverse crosslinked overnight at 65˚C as described above. DNA was isolated using Zymo DNA Clean and Concentrator column and PCR amplified for library generation as described above.

RNA & DNA SPRITE. RD-SPRITE was performed on ActD or DMSO treated *pSM44* mES 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 the SPRITE pipeline. 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 using the SPRITE 2.0 pipeline 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 [27], 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 [85].

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 [27].

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.

2.6.11 SATELLITE-DERIVED NCRNA KNOCKDOWNS AND HP1 MEASUREMENTS

LNA transfections. LNA antisense oligonucleotides designed against Major Satellite and Minor Satellite were transfected using Lipofectamine™ RNAiMAX Transfection Reagent according to manufacturer protocol (Thermo Fisher Scientific #13778030). We designed LNAs targeting the forward and reverse strand of the satellite-derived RNAs. These probes, targeting distinct regions of the transcript, were mixed to a final concentration of 10 μM each and 5 μL of the mix was transfected to each well of a 24-well plate containing cells. As a

control, non-targeting LNA were transfected at the same concentrations. After 48h or 72h in culture, cells were used for further procedures. KD for both LNA were confirmed by RT-qPCRs (**Supplemental Figure 4C-D**). We note that the LNA-depletion of MinSat RNA does not impact expression of the MajSat RNA, but MajSat RNA depletion does moderately reduce MinSat RNA (**Supplemental Figure 4C-D**).

LNA sequences. LNAs were designed by Qiagen. The following sequences were used. Minor Satellite (forward): ACTCACTCATCTAATA, Minor Satellite (reverse): TGGCAAGACAACTGAA, Major Satellite (forward): AGGTCCTTCAGTGTGC, Major Satellite (reverse): ACATTCGTTGGAAACG. Control: Negative control A Antisense LNA GapmeR (#339515).

Reverse transcription and quantitative PCR (RT-qPCR). Total RNA was extracted from mES cells with Silane beads (Sigma) according to manufacturer conditions and treated with Turbo DNase (Life Technologies) for 15min at 37C to remove genomic DNA. RT reactions were performed according to Superscript II protocol (Thermo Fisher Scientific #18064022) with random 9mer. qPCRs were performed in technical replicates using a Roche Lightcycler and a representative of three biological replicates is shown. Plots were generated using GraphPad software. ddCt values were calculated by normalizing Ct values to GAPDH and to samples transfected with control LNA to compare gene expression differences between samples.

qPCR primers used for analysis.

GAPDH:CATGGCCTTCCGTGTTCCTA
GCCTGCTTCACCACCTTCTT
MinS_1: GAACATATTAGATGAGTGAGTTAC
GTTCTACAAATCCCGTTTCCAAC
MinS_2: GATGGAAAATGATAAAAACC
CATCTAATATGTTCTACAGTGTGG

MajS_1: GACGACTTGAAAAATGACGAAATC CATATTCCAGGTCCTTCAGTGTGC

MajS_2: GCACACTGAAGGACCTGGAATATG GATTTCGTCATTTTTCAAGTCGTC

Image analysis of HP1 foci. Image visualization and analysis was performed with Icy software and ImageJ software with a minimum of 10 cells observed per condition. For HP1 foci quantification, we computed a binary mask based on relative intensity threshold (>100 for HP1ß staining replicate 1, >120 for HP1ß replicate 2) in which the relative signal intensity was set from 10 to 200.

Western Blot for HP1 levels. To access the levels of HP1ß after LNA-mediated knockdown, we performed a western blot for HP1ß. Cells were transfected as previously described and then 4 wells out of a 24 well plate pooled and flash frozen. The cells were lysed completely by resuspending frozen cell pellets in 100 µL of ice-cold lysis buffer (50 mM HEPES, pH 7.4, 100 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% Sodium Deoxycholate) supplemented with 1X Protease Inhibitor Cocktail (Roche), 20 U Turbo DNase (Ambion), and 1X Manganese/Calcium Mix (0.5 mM $CaCl_2$, 2.5 mM $MnCl_2$). Samples were incubated on ice for 10 minutes to allow lysis to proceed. The lysates were then incubated at 37°C for 10 minutes at 700 rpm shaking on a Thermomixer (Eppendorf). Following, lysates were run through a Qiashredder column (Qiagen) and cleared by centrifugation at 15,000 x g for 2 minutes. The supernatant was transferred to new tubes, mixed with LDS loading buffer and reducing buffer, heated to 95C for 3 minutes and then cooled on ice for 2 minutes. The samples were then run on a 4-12% SDS gel in MES-SDS buffer. Gel transfer to a nitrocellulose membrane was done using the P2 setting of the iBlot transfer system (Thermofisher). The nitrocellulose membrane was washed 3 times with 1x PBS and blocked for 30 minutes in LI-COR blocking buffer. The blocked membrane was incubated with primary antibodies - HP1ß (mouse, 1:1000) and LaminB1 (rabbit; 1:1000) - overnight at 4°C on a shaker. Unbound primary antibody was then removed by washing 3 times with 1x PBS + 0.1% Tween. The membrane was then incubated with secondary

antibodies (LI-COR, 1:10,000) for 45 minutes at room temperature and washed 2 more times with 1xPBS. The membranes were developed using the LI-COR Imaging System.

2.6.12 MAPPING LNCRNA LOCALIZATION

Defining lncRNAs. We used Gencode release 95 (GRCm38.p6, https://ftp.ensembl.org/pub/release-95/gtf/mus_musculus/Mus_musculus.GRCm38.95.gtf.gz) to define all lncRNAs in this study. Specifically, we included all annotations with the "lincRNA" or "antisense" biotypes to define all lncRNAs. For example, lncRNAs such as Tsix, Airn, and Kcnq1ot1 are annotated as "antisense" rather than "lincRNA." We included all lncRNAs that contained coverage in our mouse ES data by filtering the list to those that were contained in at least 10 SPRITE clusters. This yielded a list of 642 lncRNAs.

Calculation of chromatin enrichment scores. To determine the extent to which RNA transcripts are in contact with chromatin, we calculated a chromatin enrichment score for each RNA transcript. The chromatin enrichment score is computed as the ratio of the number of SPRITE clusters containing a given RNA that also contains DNA ("chromatin bound") relative to all SPRITE clusters containing the RNA transcript. We normalize these counts by the SPRITE cluster size in which it was observed (described above). We determined an "expected" DNA to RNA contact ratio by calculating mean DNA to RNA contact ratio across all RNA transcripts. Chromatin enrichment scores were calculated as the natural log of the observed DNA to RNA contact ratio divided by the expected ratio. Positive chromatin enrichment scores indicate RNA transcripts with higher ratios of DNA to RNA contacts than the mean. We performed a similar analysis to calculate enrichment scores for different sets of RNA transcripts. For example, we compute a ribosomal RNA enrichment score based on the ratio of ribosomal RNA contacts to all RNA contacts for a given RNA transcript.

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 (**Figure S5A-B**). 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.

lncRNA RNA-DNA genome wide heatmap. We plotted these 642 lncRNAs across the genome at 10Mb resolution. For each lncRNA, we computed the number of SPRITE clusters that co-occur within each 10Mb bin. We then normalized this count by the average contacts across all genomic bins. We refer to this ratio as an enrichment score. This enrichment score is intrinsically normalized for the different expression levels of different lncRNAs. We plotted all bins that have an enrichment value greater than 5-fold. We zoomed in on selected examples and plotted them across the entire genome at 1Mb resolution. In these examples, we plotted the enrichment scores across all values as a continuous bedgraph in IGV.

Calculation of lncRNAs enriched around their transcriptional loci. Using these values, we defined a lncRNA as enriched in proximity to its transcriptional locus if it was >20-fold enriched within the 10Mb bin containing its transcriptional loci. At this cutoff, lncRNAs that have very broad distribution patterns across the genome such as Malat1 are excluded, while the vast majority of lncRNAs (596 lncRNAs, 92.8%) are highly enriched around their transcriptional loci.

Visualizing proportion of lncRNAs or mRNAs on chromatin. To visually compare the fraction of different RNAs that are retained on chromatin across the genome, we computed a weighted score accounting for the counts within a given genomic bin relative to the total fraction of SPRITE clusters contained off chromatin. Specifically, we identified all SPRITE clusters containing a given RNA and computed the number that also contained a DNA read (on chromatin count) and the number that do not contain DNA (off chromatin count). We computed a score for each genomic bin defined as the number of SPRITE clusters containing an RNA and genomic bin by dividing this count by the total number of SPRITE clusters containing the same RNA that did not have a paired DNA read (off-DNA count). We multiplied this number by 100 to linearly scale values. This score accounts for different abundance levels of different RNAs allowing us to compare them directly to each other and accounts for the proportion of the RNA that is present on chromatin versus off-chromatin.

Generating nuclear structure models of lncRNA localization. To visualize the localization of lncRNAs in 3D, we generated 3D models of the genome based on SPRITE DNA-DNA contacts. We modeled each chromosome as a linear polymer composed of N monomers, where N is the number of 1Mb bins on the chromosome. Each chromosome polymer is initialized as a random walk, and then a Brownian dynamics simulation is performed on all chromosomes using an energy function composed of the following forces: 1) a harmonic bond force between adjacent monomers, 2) a spherical confinement force, 3) a repulsive force to prevent monomers from overlapping, 4) an attractive force based on SPRITE contact frequencies to ensure that preferential contacts determined by SPRITE are accurately reflected by the models. Simulations were performed using the open-source molecular simulation software OpenMM. The outputs of simulations were visualized using Pymol 2 (pymol.org/2). Chromosomes were visualized as cartoon tubes and lncRNAs were visualized by drawing a surface over the genomic

regions where lncRNA enrichment was greater than 50-fold over background.

FVP treatment and analysis. GRO-seq data from Jonkers *et al.* [56] were obtained from NCBI GEO (accession GSE48895) and aligned to mm10 using HISAT2. Raw read counts were determined for each gene using deepTools module multiBamSummary for untreated and 50 min FVP conditions. Raw read counts were converted to transcripts per million (TPM) values using a custom Python script, and fold change in TPM was calculated for each gene by dividing 50 min FVP TPM values by untreated TPM values. Cumulative distribution plots were generated using R and box-and-whisker plots were generated using PRISM.

2.6.13 KCNQ1OT1 PROTEIN BINDING, PERTURBATIONS, AND GENE EXPRESSION MEASUREMENTS

Kcnq1ot1 CRISPR interference. dCas9-4XSID cells were transfected using multiplexed gRNA vector constructs, containing an episomal polyoma origin of replication, puromycin resistance driven by a PGK promoter, and four tandem U6-gRNA cassettes, allowing for simultaneous expression of four sgRNAs. Negative control gRNA sequences recognizing the *Saccharomyces cerevisiae* Upstream Activation Sequence (UAS) and the Tetracycline Response Element (TRE) were multiplexed together (referred to as sgTUUT; gRNAs are as follows: TCTCTATCACTGATAGGGAG, GAGGACAGTACTCCGCTCGG, GCGGAGTACTGTCCTCCGAG, and TCTCTATCACTGATAGGGAG). Four gRNA sequences targeting the Kcnq1ot1 promoter were multiplexed together (referred to as sgKcnq1ot1; gRNAs are as follows: GCCTAGCCGTTGTCGCTAGG, GCCCTGTACTGCATTGAGGT, GCCTGCACAGTAGGATTCCA, and GGAGGATGGGTCGAGTGGCT).

dCas9-4XSID cells were transfected with either sgTUUT or sgKcnq1ot1 and selected for three days with 1μg/mL of puromycin in standard 2i culture conditions. Cells were subsequently passaged and maintained in 0.5μg/mL puromycin for an additional 7 days prior to RNA harvesting. Data presented are from two separate transfections and biological replicates.

SHARP binding to Kcnq1ot1 RNA using Covalent linkage and Affinity Purification (CLAP). We transfected an expression vector containing full-length SHARP with an N-terminal Halo-FLAG (HF) fusion protein into mouse ES cells containing a doxycycline inducible Xist gene. Cells were washed once with PBS and then crosslinked on ice using 0.25 J cm−2 (UV2.5k) of UV at 254 nm in a Spectrolinker UV Crosslinker. Cells were then scraped from culture dishes, washed once with PBS, pelleted by centrifugation at 1,500g for 4 min, and flash-frozen in liquid nitrogen for storage at -80°C. We lysed batches of 5 million cells by completely resuspending frozen cell pellets in 1 mL of ice cold iCLIP lysis buffer (50 mM Hepes, pH 7.4, 100 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% Sodium Deoxycholate) supplemented with 1X Protease Inhibitor Cocktail (Promega), 200 U of Murine RNase Inhibitor (New England Biolabs), 20 U Turbo DNase (Ambion), and 1X Manganese/Calcium Mix (0.5mM $CaCl_2$, 2.5 mM $MnCl_2$). Samples were incubated on ice for 10 minutes to allow lysis to proceed. The lysates were then incubated at 37°C for 10 minutes at 1150 rpm shaking on a Thermomixer (Eppendorf). Lysates were cleared by centrifugation at 15,000g for 2 minutes. The supernatant was collected and kept on ice until bound to the HaloLink Resin.

We used 200 μL of 25% HaloLink Resin (50 μL of HaloLink Resin total) per 5 million cells. Resin was washed three times with 2 mL of 1X TBS (50 mM Tris pH 7.5, 150 mM NaCl) and incubated in 1X Blocking Buffer (50 mM HEPES, pH 7.5, 10 μg/mL Random 9-mer, 100 μg/mL BSA) for 20 minutes at room temperature with continuous rotation. After the incubation, resin was washed three times with 1X TBS. The cleared lysate was mixed with 50 μL of HaloLink Resin and incubated at 4 °C for 3-16

hrs with continuous rotation. The captured protein bound to resin was washed three times with iCLIP lysis buffer at room temperature and then washed three times at 90°C for 2 minutes while shaking at 1200 rpm with each of the following buffers: 1X ProK/NLS buffer (50 mM HEPES, pH 7.5, 2% NLS, 10 mM EDTA, 0.1% NP-40, 10 mM DTT), High Salt Buffer (50 mM HEPES, pH 7.5, 10 mM EDTA, 0.1% NP-40, 1M NaCl), 8M Urea Buffer (50 mM HEPES, pH 7.5, 10 mM EDTA, 0.1% NP-40, 8 M Urea), and Tween buffer (50 mM HEPES, pH 7.5, 0.1% Tween 20, 10 mM EDTA). Finally, we adjusted the buffer by washing with Elution Buffer (50 mM HEPES, pH 7.5, 0.5 mM EDTA, 0.1% NP-40) three times at 30°C. The resin was resuspended in 83 µL of Elution Buffer and split into a 75 µL (ProK elution) and 8 µL (TEV elution) reaction. 25 µL of 4X ProK/NLS Buffer and 10 µL of ProK were added to the ProK elution tube and the sample was incubated at 50°C for 30 minutes while shaking at 1200 rpm. 2.3 µL of ProTEV Plus Protease (Promega) was added to the TEV Elution and the sample was incubated at 30°C for 30 minutes while shaking at 1200 rpm.

For each experiment, we ensured that we successfully purified the Halo-tagged protein. To do this, the TEV elution sample was mixed with 1X LDS Sample Buffer (Invitrogen) and 1X Reducing Agent (Invitrogen) and heated for 6 minutes at 70°C. The sample was run on a 3-8% Tris Acetate Gel (Invitrogen) for 1 hour at 150 V. The gel was transferred to a nitrocellulose membrane using an iBlot Transfer Device (Invitrogen). The nitrocellulose membrane was blocked with Odyssey Blocking Buffer (LI-COR) for 30 minutes. We incubated the membrane in Anti-FLAG mouse monoclonal Antibody (Sigma-Aldrich Cat# F3165, RRID:AB_259529) and V5 rabbit polyclonal antibody (Santa Cruz Biotechnology Cat# sc-83849-R, RRID:AB_2019669) at a 1:2500 dilution for 2 hours at room temperature to detect the protein. We visualized the protein by incubating the membrane in 1:17,500 dilution of both IRDye 800CW Goat anti-Rabbit IgG (LI-COR Biosciences Cat# 925-32210, RRID:AB_2687825) and IRDYE 680DR Goat anti-Mouse IgG (LI-COR Biosciences Cat# 925-

68070, RRID:AB_2651128) for 1 hour at room temperature followed by imaging on a LI-COR Odyssey.

RNA was purified from the Proteinase K elution sample and an RNA-Seq library was constructed as previously described. Briefly, after Proteinase K elution, the RNA was dephosphorylated (Fast AP) and cyclic phosphates removed (T4 PNK) and then cleaned up on Silane beads as previously described [72]. The RNA was then ligated to an RNA adapter containing a RT primer binding site. The ligated RNA was reverse transcribed (RT) into cDNA, the RNA was degraded using NaOH, and a second adapter was ligated to the single stranded cDNA. The DNA was amplified, and Illumina sequencing adaptors were added by PCR using primers that are complementary to the 3' and 5' adapters. The molarity of PCR amplified libraries was measured by Agilent Tapestation High Sensitivity DNA and all samples were pooled at equal molarity. The pool was then purified and size selected on a 2% agarose gel and cut between 150-700 nts. The final libraries were measured by Agilent Bioanalyzer and Qubit High Sensitivity DNA to determine the loading density of the final pooled sample. Pooled samples were paired-end sequenced on an Illumina HiSeq 2500 with read length 35 x 35nts.

Sequencing reads were trimmed to remove adaptor sequences and any bases containing a quality scores <10 using Trimmomatic[86]. We filtered out all read-pairs where either read was trimmed to <25 nucleotides. We excluded PCR duplicates using the FastUniq tool [87]. The remaining reads were then aligned to Ribosomal RNAs (rRNAs) using the Tagdust program[88] with a database of 18S, 28S, 45S, 5S, 5.8S rRNA sequences. TagDust was chosen because it allowed more permissive alignments to rRNA reads that contained mismatches and indels due to RT errors induced by rRNA post-transcriptional modifications. The remaining reads were then aligned to the mouse genome using STAR aligner [89]. Only reads that mapped uniquely in the genome were kept for further analysis.

Stability of SHARP protein lacking RNA recognition motifs (ΔRRM). We generated mouse embryonic stem cells (TX1072; gift from E. Heard

[69]) that express either full length SHARP or a truncated version of SHARP lacking the four RRM domains (SHARPΔ1-591) using stable random integration with Piggy-Bac. Both these SHARP variants were tagged with eGFP. To assess the stability of the ΔRRM-SHARP protein, we measured single cell eGFP expression using flow cytometry. Cells expressing full length (FL) or ΔRRM-SHARP were trypsinized to single cell suspension, as described previously, and resuspended in 1xPBS. Fluorescence was detected using the MACSQuan VYB cell analyzer. We gated on the single cell population and plotted the distribution GFP fluorescence levels for each sample. At least 10,000 cells were analyzed for each condition.

Genetic deletion of SHARP Binding Site (ΔSBS) in Kcnq1ot1. F1 2-1 line were CRISPR-targeted with gRNAs targeting the SHARP-Binding Site (SBS) (SHARP Binding Site Coordinates: mm10 - chr7:143,295,789-143,296,455; gRNA sequences were ATGCACCATCATAGACCACG and TCATAGCCTCCCCCTCCTCG). Following selection using 1 μg/mL of puromycin in standard 2i culture conditions, transfected cells were allowed to recover in standard 2i media prior to sub-cloning. Clone were subsequently screened using genomic DNA PCR, using primers flanking the deletion region (CAGCATCTGTCCAATCAACAG and GCAAAATACGAGAACTGAGCC). In contrast to the wild type 1048bp band, successfully targeted alleles produced a 305bp band. Sub-clones homozygous for the targeted allele were subject to RT-qPCR and GAPDH-normalized gene expression was further normalized to the F1 parent line.

HDAC inhibitor treatment. The inducible Kcnq1ot1 cell line was treated with either DMSO (control) or 5μM TSA in fresh 2i media or 2μg/mL doxycycline in standard 2i. RNA was extracted, reverse transcribed, and qPCR was performed. Ct values were normalized to GAPDH to compare gene expression differences between induced and non-induced samples within the same pharmacologic condition (i.e., GAPDH-normalized "Induced DMSO" to GAPDH-normalized "Non-Induced DMSO

Vehicle") to generate fold gene expression ratios. RT-qPCR data presented is summarized from two separate replicate experiments.

ChIP-seq of H3K27Ac upon induction of Kcnq1ot1. The inducible Kcnq1ot1 cell line was treated with either DMSO (control; -dox) or 2μg/mL doxycycline (+dox) in standard 2i for 24 hours to induce expression in two biological replicates. 10 million cells equivalents were then harvested and crosslinked in suspension at room temperature with 1% Formaldehyde for 10 minutes. After crosslinking, the formaldehyde crosslinker was quenched for 5 minutes with addition of 2.5M Glycine for final concentration of 0.5M. Cells were spun down, washed three times with 1x PBS + 0.5% RNAse Free BSA (AmericanBio #AB01243-00050) and final cell pellets flash frozen at -80C for storage.

For cell lysis with nuclear enrichment, the cell pellets were resuspended in 1 ml of Gagnon Hypotonic lysis buffer (10mM Tris pH 7.5, 10mM NaCl, 3mM $MgCl_2$, 0.3% NP-40 (v/v), 10% glycerol (v/v)) + 1:50 PIC, incubated on ice for 10 minutes, vortexed, and pelleted by centrifugation for 3min at 1250g. The isolated nuclei were resuspended in 600 μL of Mammalian Lysis Buffer (50mM HEPES, 150mM NaCl, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS) + 1X PIC and transferred to 15mL conical tubes (Diagenode adaptors - C30010009). Chromatin was fragmented using a Bioruptor waterbath sonicator for 27 cycles at max intensity for 30 seconds followed by 30 seconds of rest. To remove debris, the lysate was centrifuged at 13000RPM for 10 minutes at 4C and cleared by incubating at room temperature for 1 hour with 100 μL of Protein G beads in 500μL of 1X RIPA (10mM Tris-HCl pH 7.5, 1mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% Sodium deoxycholate, 100nM NaCl) + 1:50 PIC. The resulting supernatant was diluted in 1800 μL of Hanks' Balanced Salt Solution (Thermo Scientific 88284) + 2400 μL of 2X RIPA + 1:50 PIC. A 1% aliquot (48 μL) was taken to serve as input.

H3K27Ac antibody-Protein G bead complexes were prepared a day in advance. 5 μg of H3K27Ac Antibody (Active Motif, 39134) was incubated with 100 μL of Dynabeads Protein G (ThermoFisher Scientific

10003D) in 500 µL of 1X RIPA + 1:50 PIC for 4 hours with rotation at 4C. The beads were washed twice with 1X RIPA + 1:50 PIC and stored at 4C until use.

Prepared chromatin (~4.8ml of mixture) was coupled to the prepared Antibody-Bead complexes (200 µL in 1X RIPA) overnight (12-15hrs) at 4C while rotating end-to-end on a hula mixer. Coupled beads were then washed 1X with Low Salt Immune Complex Buffer (50mM Tris-HCl pH 8.1, 150mM NaCl, 0.1% SDS, 1% Triton X-100, 2mM EDTA), 1X with High Salt Immune Complex Wash Buffer (50mM Tris-HCl pH 8.1, 50mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100), 1X with LiCl Immune Complex Wash Buffer (10mM Tris-HCl pH 8.1, 0.25 M LiCl, 1mM EDTA, 1% Igepal-CA630, 1% deoxycholic acid) and 1X with TE Buffer (10mM Tris-Hcl pH 8, 10mM EDTA). DNA molecules were eluted from the beads by reverse crosslinking overnight (~12-13 hours) at 65°C in NLS Elution Buffer (20mM Tris-HCl pH 7.5, 10mM EDTA, 2% Sodium-Lauroylsarcosine, 50mM NaCl) supplemented with 10 µL Proteinase K (NEB). The eluted DNA was purified using the Zymo DNA Clean Up and Concentrator Kit.

Sequencing was performed on an Illumina HiSeq 2500, 100 base pair paired end flowcell. Sequencing reads were trimmed using Trimmomatic [86] to remove adaptor sequences and any bases containing a quality scores <10. Reads were then aligned to the mouse GRCm38.p6 genome using STAR aligner [89] and only reads that mapped uniquely were kept for further analysis. RepeatMasker [78] defined regions with milliDev ≤ 140 along with blacklisted v2 regions were filtered out using Bedtools v2.29.0 [79]. Using the aligned and filtered read set, H3K27 acetylation peaks were called using MACS2 with default settings [90].

H3K27 ChIP-seq Analysis. For each gene of interest, windows over the promoter region were defined using the H3K27ac peaks in the -dox control sample. For some genes, multiple H3K27ac peaks were detected, and each peak window was analyzed separately. The number of reads falling within the promoter-overlaying window was counted and normalized to the total

reads in the experiment. Then, the change in promoter acetylation following Kcnq1ot1 induction was calculated for each gene by taking the ratio of normalized reads in the +dox condition to the -dox condition. Analysis was performed and reported separately for two replicates.

2.6.14 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. Precision measures such as mean, standard deviation, confidence intervals are described in the corresponding figure legends.

2.7 REFERENCES

1. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology* **16**, 245–257 (2015).
2. Strom, A. R. & Brangwynne, C. P. The liquid nucleome - phase transitions in the nucleus at a glance. *Journal of cell science* Preprint at https://doi.org/10.1242/jcs.235093 (2019).
3. Dundr, M. & Misteli, T. Biogenesis of nuclear bodies. *Cold Spring Harbor perspectives in biology* vol. 2 Preprint at https://doi.org/10.1101/cshperspect.a000711 (2010).
4. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219–226 (2017).
5. Pederson, T. The nucleolus. *Cold Spring Harbor Perspectives in Biology* **3**, 1–15 (2011).
6. Spector, D. L. & Lamond, A. I. Nuclear speckles. *Cold Spring Harbor Perspectives in Biology* **3**, 1–12 (2011).
7. Cho, W. K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* (2018) doi:10.1126/science.aar4199.
8. Guo, Y. E. *et al.* Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* Preprint at https://doi.org/10.1038/s41586-019-1464-0 (2019).
9. Nickerson, J. A., Krochmalnic, G., Wan, K. M. & Penman, S. Chromatin architecture and nuclear RNA. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 177–81 (1989).
10. Rinn, J. L. & Guttman, M. RNA and dynamic nuclear organization. *Science* **345**, 1240–1241 (2014).
11. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* (2019) doi:10.1093/nar/gky955.
12. Rinn, J. L. & Chang, H. Y. Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry* (2012) doi:10.1146/annurev-biochem-051410-092902.

13. Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annual Review of Biochemistry* (2003) doi:10.1146/annurev.biochem.72.121801.161720.
14. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* Preprint at https://doi.org/10.1038/nature08909 (2010).
15. Kiss-László, Z., Henry, Y., Bachellerie, J. P., Caizergues-Ferrer, M. & Kiss, T. Site-specific ribose methylation of preribosomal RNA: A novel function for small nucleolar RNAs. *Cell* (1996) doi:10.1016/S0092-8674(00)81308-2.
16. Watkins, N. J. & Bohnsack, M. T. 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* Preprint at https://doi.org/10.1002/wrna.117 (2012).
17. Kolev, N. G. & Steitz, J. A. Symplekin and multiple other polyadenylation factors participate in 3′-end maturation of histone mRNAs. *Genes and Development* (2005) doi:10.1101/gad.1371105.
18. Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A. & Panning, B. Xist RNA and the Mechanism of X Chromosome Inactivation . *Annual Review of Genetics* (2002) doi:10.1146/annurev.genet.36.042902.092433.
19. Egloff, S., Studniarek, C. & Kiss, T. 7SK small nuclear RNA, a multifunctional transcriptional regulatory RNA with gene-specific features. *Transcription* Preprint at https://doi.org/10.1080/21541264.2017.1344346 (2018).
20. Engreitz, J. M. *et al.* The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science* **341**, 1237973–1237973 (2013).
21. Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell* (2010) doi:10.1016/j.molcel.2010.08.011.

22. Bell, J. C. *et al.* Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. *eLife* (2018) doi:10.7554/eLife.27024.
23. Li, X. *et al.* GRID-seq reveals the global RNA-chromatin interactome. *Nature Biotechnology* (2017) doi:10.1038/nbt.3968.
24. Yan, Z. *et al.* Genome-wide colocalization of RNA–DNA interactions and fusion RNA pairs. *Proceedings of the National Academy of Sciences of the United States of America* (2019) doi:10.1073/pnas.1819788116.
25. Sridhar, B. *et al.* Systematic Mapping of RNA-Chromatin Interactions In Vivo. *Current Biology* (2017) doi:10.1016/j.cub.2017.01.011.
26. Bonetti, A. *et al.* RADICL-seq identifies general and cell type-specific principles of genome-wide RNA-chromatin interactions. *bioRxiv* (2019) doi:10.1101/681924.
27. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).
28. West, J. A. *et al.* The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites. *Molecular Cell* (2014) doi:10.1016/j.molcel.2014.07.012.
29. Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159**, 188–199 (2014).
30. Schoeftner, S. & Blasco, M. A. Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nature Cell Biology* (2008) doi:10.1038/ncb1685.
31. Mumbach, M. R. *et al.* HiChIRP reveals RNA-associated chromosome conformation. *Nature Methods* (2019) doi:10.1038/s41592-019-0407-x.
32. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: Organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology* Preprint at https://doi.org/10.1038/nrm.2017.7 (2017).

33. Baßler, J. & Hurt, E. Eukaryotic Ribosome Assembly. *Annual Review of Biochemistry* (2019) doi:10.1146/annurev-biochem-013118-110817.
34. Lee, Y. & Rio, D. C. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry* (2015) doi:10.1146/annurev-biochem-060614-034316.
35. Herzel, L., Ottoz, D. S. M., Alpert, T. & Neugebauer, K. M. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. *Nature Reviews Molecular Cell Biology* Preprint at https://doi.org/10.1038/nrm.2017.63 (2017).
36. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics* vol. 15 163–175 Preprint at https://doi.org/10.1038/nrg3662 (2014).
37. Tycowski, K. T., You, Z. H., Graham, P. J. & Steitz, J. A. Modification of U6 spliceosomal RNA is guided by other small RNAs. *Molecular Cell* (1998) doi:10.1016/S1097-2765(00)80161-6.
38. Darzacq, X. *et al.* Cajal body-specific small nuclear RNAs: A novel class of 2′-O-methylation and pseudouridylation guide RNAs. *EMBO Journal* (2002) doi:10.1093/emboj/21.11.2746.
39. Machyna, M., Heyn, P. & Neugebauer, K. M. Cajal bodies: Where form meets function. *Wiley Interdisciplinary Reviews: RNA* Preprint at https://doi.org/10.1002/wrna.1139 (2013).
40. Ogg, S. C. & Lamond, A. I. Cajal bodies and coilin - Moving towards function. *Journal of Cell Biology* vol. 159 17–21 Preprint at https://doi.org/10.1083/jcb.200206111 (2002).
41. Nizami, Z., Deryusheva, S. & Gall, J. G. The Cajal body and histone locus body. *Cold Spring Harbor perspectives in biology* vol. 2 Preprint at https://doi.org/10.1101/cshperspect.a000653 (2010).
42. Machyna, M., Neugebauer, K. M. & Staněk, D. Coilin: The first 25 years. *RNA Biology* (2015) doi:10.1080/15476286.2015.1034923.

43. Marzluff, W. F., Wagner, E. J. & Duronio, R. J. Metabolism and regulation of canonical histone mRNAs: Life without a poly(A) tail. *Nature Reviews Genetics* Preprint at https://doi.org/10.1038/nrg2438 (2008).
44. Marzluff, W. F. & Koreski, K. P. Birth and Death of Histone mRNAs. *Trends in Genetics* Preprint at https://doi.org/10.1016/j.tig.2017.07.014 (2017).
45. Nizami, Z., Deryusheva, S. & Gall, J. G. The Cajal body and histone locus body. *Cold Spring Harbor perspectives in biology* vol. 2 Preprint at https://doi.org/10.1101/cshperspect.a000653 (2010).
46. Jády, B. E. & Kiss, T. A small nucleolar guide RNA functions both in 2′-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. *EMBO Journal* (2001) doi:10.1093/emboj/20.3.541.
47. Bensaude, O. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription* **2**, 103–108 (2011).
48. Goldfarb, K. C. & Cech, T. R. Targeted CRISPR disruption reveals a role for RNase MRP RNA in human preribosomal RNA processing. *Genes and Development* (2017) doi:10.1101/gad.286963.116.
49. Cech, T. R. & Steitz, J. A. The noncoding RNA revolution - Trashing old rules to forge new ones. *Cell* Preprint at https://doi.org/10.1016/j.cell.2014.03.008 (2014).
50. Maison, C. *et al.* Higher-order structure in pericentric heterochromatin involves a distinct pattern of histone modification and an RNA component. *Nature Genetics* (2002) doi:10.1038/ng843.
51. Barutcu, A. R., Blencowe, B. J. & Rinn, J. L. Differential contribution of steady-state RNA and active transcription in chromatin organization . *EMBO reports* (2019) doi:10.15252/embr.201948068.

52. Casanova, M. *et al.* Heterochromatin Reorganization during Early Mouse Development Requires a Single-Stranded Noncoding Transcript. *Cell Reports* (2013) doi:10.1016/j.celrep.2013.08.015.
53. Cabili, M. N. *et al.* Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biology* **16**, 20 (2015).
54. Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* (2016) doi:10.1016/j.cell.2015.12.017.
55. Levesque, M. J. & Raj, A. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nature methods* **10**, 246–8 (2013).
56. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **2014**, (2014).
57. Clark, M. B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Research* (2012) doi:10.1101/gr.131037.111.
58. McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232–236 (2015).
59. Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161**, 404–416 (2015).
60. Żylicz, J. J. *et al.* The Implication of Early Chromatin Changes in X Chromosome Inactivation. *Cell* (2019) doi:10.1016/j.cell.2018.11.041.
61. Kanduri, C. Kcnq1ot1: A chromatin regulatory RNA. *Seminars in Cell and Developmental Biology* Preprint at https://doi.org/10.1016/j.semcdb.2011.02.020 (2011).
62. Nagano, T. & Fraser, P. Emerging similarities in epigenetic gene silencing by long noncoding RNAs. *Mammalian Genome* Preprint at https://doi.org/10.1007/s00335-009-9218-1 (2009).

63. Braidotti, G. *et al.* The Air noncoding RNA: An imprinted cis-silencing transcript. in *Cold Spring Harbor Symposia on Quantitative Biology* (2004). doi:10.1101/sqb.2004.69.55.
64. Olivero, C. E. *et al.* p53 Activates the Long Noncoding RNA Pvt1b to Inhibit Myc and Suppress Tumorigenesis. *Molecular Cell* (2020) doi:10.1016/j.molcel.2019.12.014.
65. Rom, A. *et al.* Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nature Communications* (2019) doi:10.1038/s41467-019-13075-8.
66. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* (2016) doi:10.1038/nature20149.
67. Matera, A. G. & Frey, M. R. Coiled bodies and gems: Janus or gemini? *American Journal of Human Genetics* (1998) doi:10.1086/301992.
68. Tucker, K. E. *et al.* Residual Cajal bodies in coilin knockout mice fail to recruit Sm snRNPs and SMN, the spinal muscular atrophy gene product. *Journal of Cell Biology* (2001) doi:10.1083/jcb.200104083.
69. Schulz, E. G. *et al.* The Two Active X Chromosomes in Female ESCs Block Exit from the Pluripotent State by Modulating the ESC Signaling Network. *Cell Stem Cell* **14**, 203–216 (2014).
70. Engreitz, J. M. *et al.* The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science* **341**, 1237973–1237973 (2013).
71. Markaki, Y. *et al.* Xist-seeded nucleation sites form local concentration gradients of silencing proteins to inactivate the X-chromosome. *bioRxiv* (2020) doi:https://doi.org/10.1101/2020.11.22.393546.
72. Shishkin, A. A. *et al.* Simultaneous generation of many RNA-seq libraries in a single reaction. *Nature Methods* **12**, 323–325 (2015).
73. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* (2011) doi:10.14806/ej.17.1.200.

74. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods* (2015) doi:10.1038/nmeth.3317.
75. Jin, Y., Tam, O. H., Paniagua, E. & Hammell, M. TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btv422.
76. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
77. Krueger, F. & Andrews, S. R. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Research* (2016) doi:10.12688/f1000research.9037.1.
78. Smit, A., Hubley, R. & Grenn, P. RepeatMasker Open-4.0. *RepeatMasker Open-4.0.7.* (2015).
79. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* (2010) doi:10.1093/bioinformatics/btq033.
80. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btw354.
81. Chan, P. P. & Lowe, T. M. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research* (2009) doi:10.1093/nar/gkn787.
82. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Research* (2016) doi:10.1093/nar/gkv1309.
83. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biology* (2005) doi:10.1371/journal.pbio.0030157.
84. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* Preprint at https://doi.org/10.1038/nbt.1754 (2011).

85. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods* (2012) doi:10.1038/nmeth.2148.
86. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu170.
87. Xu, H. *et al.* FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. *PLoS ONE* (2012) doi:10.1371/journal.pone.0052249.
88. Lassmann, T., Hayashizaki, Y. & Daub, C. O. TagDust - A program to eliminate artifacts from next generation sequencing data. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp527.
89. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) doi:10.1093/bioinformatics/bts635.
90. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).

Chapter 3

A METABOLIC LABELLING METHOD TO ENHANCE PURIFICATION OF NASCENT PRE-MRNAS

Prashant Bhat and Mitchell Guttman

"We're born, we live a little while, we die." E.B. White in Charlotte's Web

3.1 ABSTRACT

One of the challenges in genome-wide assays for RNA biology is the availability of methods to purify and enrich for low abundance RNAs – including nascent pre-mRNAs and mature lncRNAs – which may only comprise 1-5% of the total population of RNA in the cell. Genome-wide steady-state measurements of spliced mRNAs provide a mixture of splicing events that occur co-transcriptionally and post-transcriptionally, making it difficult to measure the effect of specific ncRNAs (i.e. MALAT1) on co-transcriptional splicing of nascent transcripts. Thus, a genome-wide method is needed to (1) enrich for nascent pre-mRNAs and (2) measure their splicing rates at the co-transcriptional level. Commonly used enrichment strategies, including manufacturer protocols only yield 2-3 fold enrichment of target, low abundance RNAs. Here, I will describe a genome-wide method to capture specific classes of low abundance RNAs that achieves over 50-fold enrichment of target RNAs.

3.2 PROTOCOL

This protocol is optimized for a starting amount of 5 – 10 ug of 10-minute 5EU labeled RNA. Cells grown to confluency in one well of a 6 well plate should yield enough material for this protocol.

3.2.1 PROTOCOL OVERVIEW

Step 1: Biotinylate 5eu RNA: click reactions with biotin picolyl azide
Step 2: Streptavidin Dot Blot Assay
Step 3: Sequential Capture Protocol
Step 4: Dephosphorylation/End repair
Step 5: First RNA ligation
Step 6: Reverse Transcription
Step 7: cDNA ligation
Step 8: PCR amplification of cDNA
Step 9: Library Pooling and Gel Cut (HiSeq 200 cycle, Paired End Kit)
Step 10: Data processing, alignment to mm10 reference genome, FastQC

STEP 1: BIOTINYLATE 5EU RNA: CLICK REACTIONS WITH BIOTIN PICOLYL AZIDE

1. Dissolve biotin picolyl azide in anhydrous DMSO to working concentration of 10 mM (I dissolve the stock to 100 mM and then further dilute to 10 mM for a working concentration) and aliquot azide into tubes of 50 ul each. Minimize freeze-thaw cycles to no more than 5 cycles.
2. Dissolve THPTA to 400 mM stock concentration in nuclease free H2O. 0.2 μm filter and aliquot into 5 μL single use tubes. THPTA can be freeze-thawed a maximum of 3 times after which it must be discarded.
3. To prepare 1 mL of 400 mM sodium ascorbate, measure out 79 mg of sodium ascorbate and transfer to a 1.7 ml microfuge tube. Wait until all reaction components have been mixed before dissolving sodium ascorbate as it has a very short half-life in water and oxidizes rapidly.

a. Note: Shelf life of sodium ascorbate is ONE YEAR. Mark the date when the new bottle is opened and discard after one year has elapsed or if the powder starts turning yellow. Minimize exposure to air when measuring, and seal bottle cap with parafilm once finished.

4. Mix sample, water, HEPES, biotin picolyl azide, and RNAse inhibitor in that order.
5. Premix $CuSO_4$ and THPTA in a separate tube to generate Cu(I).
6. Working quickly, add 1 mL of nuclease-free H2O to 79 mg of sodium ascorbate. Invert up and down six to eight times to dissolve.
7. Add 5.25 ul of pre-mixed $CuSO_4$ and THPTA to sample. Vortex sample briefly, touch spin and move quickly to step 8.
8. Add 1.5 ul of sodium ascorbate to initiate reaction, vortex briefly, touch spin.
9. Ensure caps are completely fastened (or use safe-lock tubes) from air, which will oxidize sodium ascorbate and result in lower click reaction efficiency.
10. Incubate reactions for 1 hour on thermomixer at 25C, 1000 rpm.

Stock concentration	**Component**	**Final concentration**	**Sample**	**7x master mix**
25 pmol/ul	5eU IVT RNA standard or alkyne-modified DNA oligonucleotide		32.75	

variable	5eu RNA or EdU DNA		variable	
10 mM	Biotin picolyl azide	1 mM	5	35
25 mM	$CuSO_4$	2 mM	4	28
400 mM	THPTA	10 mM	1.25	8.75
400 mM	Sodium ascorbate	12 mM	1.5	
1000 mM	HEPES, pH = 7.4	100 mM	5	35
	Ribolock RNase inhibitor		0.5	3.5
	H2O		variable	
		Final volume	50	

11. After reactions are completed, clean and concentrate RNA using >17 nt Zymo Clean and Concentrate protocol on a **Zymo IC column.**
12. Elute in 30 µL nuclease-free H2O.

********Save 1 ul of this as pre-capture input to be used in Step 4: End repair!*********

STEP 2: STREPTAVIDIN DOT BLOT ASSAY

1. Take a PINK box for western, spray and wash 3X with 70% Ethanol – dry thoroughly.
2. Cut appropriately sized nitrocellulose membrane to fit standards and samples and label the grid.
3. Load equal volume of standards and samples (Here, I dotted 2 ul of standards and samples) by directly pipetting volumes onto the membrane.
4. Wait 5 min to let dry before proceeding to Step 5.
5. Carefully remove blue paper and only place white nitrocellulose membrane in UV crosslinker.
6. Place the membrane in the UV Crosslinker and click “optimal crosslink” and press start.
7. Once completed, place back in Licor box, and incubate in 10 ml 1X PBS (RNase Free) for 2 minutes.
8. Incubate in 15 mL 1X Licor blocking buffer (Dilute 1:1 into 1X PBS RNase Free), rocking at RT for 30 minutes.
    - Combine 7.5 ml 2x Licor blocking buffer + 7.5 ml of 1X PBS
9. Remove blocking buffer and incubate in 15mL of a 1:2000 dilution (7.5 ul) of Streptavidin IRdye 800 Conjugate in licor blocking buffer + 0.1% Tween + 0.1% Streptavidin, **(spin down stock of streptavidin IRdye conjugate prior to making dilution and avoid any precipitate)**, rock at RT for 30 minutes.
10. Remove Streptavidin and wash blot 3X with 15 mL 1X PBS (RNASe Free) + 0.1% Tween, 5-minute incubations each time.
11. Finally, rinse the blot 3X in 15mL of 1X PBS (RNAse Free)
12. Image on Licor (set focal offset to 0 μm).

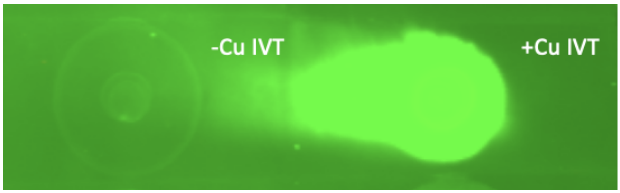


Figure 1: Example image of negative and positive control for successful click reactions.

STEP 3: SEQUENTIAL CAPTURE PROTOCOL

Prepare Streptavidin Beads (can be done in advance for each capture)

1. Aliquot 20ul of Dynabeads MyOne Streptavidin C1 into a 1.7mL tube per capture.
    a. For example: 50 ul x 6 samples x 3 captures = **900 ul SA beads**
2. Spin down, magnet separate, and remove supernatant storage solution.
3. Wash 3x with 1000uL of M2 Wash Buffer.
4. Resuspend beads in 900 ul M2 Buffer + 22.5 ul Ribolock (1:40).

Capture I, Washes, and Elution

5. Resuspend beads with 150ul of 4M Urea Buffer. Add 50ul of Biotinylated RNA. Combine for total volume 200ul.
6. Incubate for 60 minutes at room temp, 900rpm.
7. Magnet separate supernatant.
8. Wash 3X with 150 µl of 1X 4M Urea RAP Buffer for 5 minutes per wash at 37C, 750rpm.
9. Wash 3X with 150 µl of M2 Wash Buffer.
10. Resuspend in 50ul of Elution Buffer (5.7M Guanidine plus 1% NLS). Incubate at 65C X 2 minutes, shaking at 1200 rpm.

11. Collect elution into separate tube and elute beads with 50ul of additional Elution Buffer. Incubate at 65min X 2 minutes, shaking at 1200 rpm. Pool elutions.

Capture II, Washes, and Elution

12. Dilute with Post-Capture I with 400ul of 1X RAP Buffer. Incubate with 50ul of pre-washed SA beads.
13. Incubate for 60 minutes at room temp, 900rpm.
14. Magnet separate supernatant.
15. Wash 3X with 150 µl of 1X 4M Urea RAP Buffer for 5 minutes per wash at 37C, 750rpm.
16. Wash 3X with 150 µl of M2 Wash Buffer.
17. Resuspend in 50ul of Elution Buffer (5.7M Guanidine plus 1% NLS). Incubate at 65C X 2 minutes, shaking at 1200 rpm.
18. Collect elution into separate tube and elute beads with 50ul of additional Elution Buffer. Incubate at 65min X 2 minutes, shaking at 1200 rpm. Pool elutions.

Capture III, Washes, and Elution

19. Dilute with Post-Capture I with 400ul of 1X RAP Buffer. Incubate with 50ul of pre-washed beads.
20. Incubate for 60 minutes at room temp, 900rpm.
21. Magnet separate supernatant.
22. Wash 3X with 150 µl of 1X 4M Urea RAP Buffer for 5 minutes per wash at 37C, 750rpm.
23. Wash 3X with 150 µl of M2 Wash Buffer.
24. Resuspend in 50ul of Elution Buffer (5.7M Guanidine plus 1% NLS). Incubate at 65C X 2 minutes, shaking at 1200 rpm.
25. Collect elution into separate tube and elute beads with 50ul of additional Elution Buffer. Incubate at 65min X 2 minutes, shaking at 1200 rpm. Pool elutions.
26. Zymo RNA Clean and Concentrate in an IC column.
27. Elute in 17 ul H2O.

28. Run 2 ul of capture 3 eluate on HS DNA Tapestation to assay sizes and capture.
29. Proceed with Dephoshorylation/End repair and Ril19 ligation.

Step 4: Dephosphorylation/End repair

1. Fragment 10 ng inputs for 3 minutes at 91*C for 3 minutes but not the captures in 1X Fast AP Buffer, then add enzymes on top.
2. Add 16 ul of Click'd RNA to FastAP mix according to the table below.

Component	**1 rxn**
Cleaned "Clicked" RNA	16 ul
10X Fast AP Buffer	2
Fast AP	1
RNase Inhibitor	1
Total	20

3. Incubate for 10 min at 37C.
4. Add PNK mix to each reaction according to the table below.

Component	**1 rxn**
10X PNK Buffer	3
PNK enzyme	3

ddH20	4
Total	10

5. Incubate for an additional 10 min at 37C.
6. Add 20 ul H2O to bring volume up to 50 ul.
7. Clean using RNA clean and concentrate/Zymo column >17 nt.
8. Elute in 12 ul H2O.

STEP 5: FIRST RNA LIGATION

1. Add 11uL of cleaned RNA + 1.5uL DMSO + 1uL Ril19 adapter (20uM).
    a. 5' – /Phosphate/rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrG<> – 3' = RIL19 Needs: 5'-P and 3' ddC (or 3'-C3 spacer)
2. Heat at 65C for 2 min, then hold at 4C.
3. Prepare Ligation Mastermix.

Component	1 rxn
10X T4 RNA Ligase Buffer	2
ATP (100 mM)	0.2
50% PEG 8000	6
RNase Inhibitor	0.3
T4 RNA ligase	Don't add to MM

Total	8.5

4. Add 8.5uL of master mix to each sample
5. Mix mastermix well by flicking and pop-spin repeatedly, use low retention tips for ligation reactions.
6. Add 1.3 uL of T4 RNA Ligase High Concentration to each reaction.
7. Incubate for 1 hr 15 min at 24C, 1600rpm (1 min on, 5 min off).
8. Clean using Silane beads.
9. 12 uL Silane beads in 60 uL RLT per sample.
10. Incubate for 1 minute at RT.
11. Add 76 uL 100% EtOH.
12. Incubate for 2 minutes at RT.
13. Place on Magnet, discard supernatant.
14. Wash 3X with 80% EtOH.
15. Dry at RT and elute in 12 uL ddH20.

STEP 6: REVERSE TRANSCRIPTION

1. Add 1uL of AR17 (5uM) to 11uL of each sample.
    a. 5′ – ACACGACGCTCTTCCGA – 3′ = AR17 (no modifications, standard desalting)
2. Heat at 65C for 2 min, then hold at 4C.
3. Add 8.5uL of RT mastermix to each sample, according to the table below.

Component	1 rxn
5X First Strand (FS) Buffer	4
H2O	1

10 mM dNTP	2
RNAse Inhibitor	0.5
Maxima RT	1
Total	8.5

4. Incubate at 50C for 30 min, 75C for 15 min, then hold at 4C o/n.
5. Add 4uL of Exo-SAP IT.
6. Incubate for 15 min at 37C.
7. Add 1uL of 0.5M EDTA.
8. Place on ice for 3 min.
9. Add 2.5uL of 1M NaOH.
10. Heat at 80C for 6 min, then hold at 4C.
11. Add 2.5uL of 1M HCl to neutralize base.
12. Clean using Silane beads.
13. Add 12 uL Silane beads in 90 uL RLT per sample.
14. Incubate for 1 minute at RT.
15. Add 105 uL 100% EtOH.
16. Incubate for 2 minutes at RT.
17. Place on Magnet, discard supernatant.
18. Wash 3X with 80% EtOH.
19. Dry at RT and elute in 8 uL ddH20.

STEP 7: CDNA LIGATION

1. Add 1uL of DMSO + 0.6uL of 3TR3 linker (80uM) to 7.5uL of cDNA.
   a. 5'/Phosphate/AGATCGGAAGAGCACACGTCTG<> — 3Tr3 Needs: 5'-P and 3' ddC (or 3'-C3 spacer)
2. Heat at 80C for 2 min, then hold at 4C.

3. Add 11.8uL of cDNA ligation master mix to each sample (see table below).

Component	1x Rxn
10X T4 RNA Ligase Buffer	2
ATP (100 mM)	0.2
50% PEG 8000	9
T4 RNA ligase	Don't add to MM
Total	11.2

4. Add 1.3 uL of T4 RNA Ligase High Concentration to each reaction.
5. Incubate overnight at 24C, 1600rpm (1 min on, 5 min off).
6. This ligation can go as short as 4 hours.
7. Clean using Silane beads.
8. Add 12 uL Silane beads in 60 uL RLT per sample.
9. Incubate for 1 minute at RT.
10. Add 76 uL 100% EtOH.
11. Incubate for 2 minutes at RT.
12. Place on Magnet, discard supernatant.
13. Wash 3X with 80% EtOH.
14. Dry at RT and elute in 32 uL ddH20.

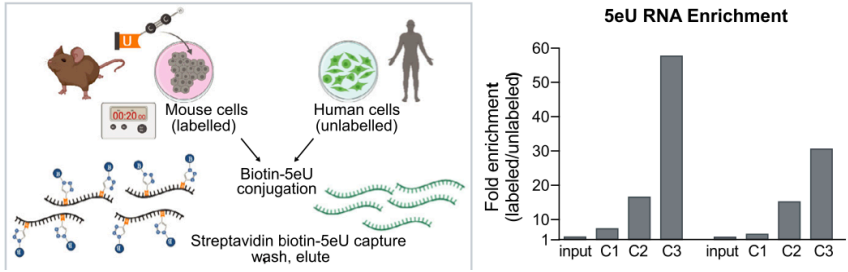


Figure 2: Human-mouse mixing experiment showing that enrichment of 5EU labeled RNA is specifically enriched over three rounds of capture.

STEP 8: PCR AMPLIFICATION OF CDNA

1. To a new strip of PCR tubes, add 23uL sample + 2uL of 25 uM mixed primers (indices) + 25uL of 2X Q5 Master Mix.

PCR enrichment primers	**Sequence**
Primer 1: from IndexedPrimers BARCODES.xls	5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
Primer 2: 2P-Universal	5'CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'

2. Save the remaining ligated cDNA as pre-PCR sample and freeze at -20C (can be used for reamplification).
3. PCR Conditions

	emp (°C)	Seconds
Initial Denaturation	98°C	180 s

4 cycles:	98°C	10 s
	68°C	30 s
	72°C	60 s
2 cycles (input) 6 cycles (captures)	98°C	10 s
	70°C	30 s
	72°C	60 s
Final Extension	72°C	180 s
Hold	4°C	Indefinitely

4. Clean with SPRI (1.4X).
5. Add 1.4X SPRI beads (70ul).
6. Mix well by pipetting and incubate at RT for 10 min (mix by pipetting every 3 min).
7. Place on magnet and discard supernatant.
8. Wash 2x with 100 ul 80% EtOH on bead.
9. Let dry for 4-5 min at RT.
10. Elute in 12 ul H2O.
11. Use 1 ul for DNA Qubit and run samples on HS DNA Bioanalyzer.

STEP 9: LIBRARY POOLING AND GEL CUT (HISEQ 200 CYCLE, PAIRED END KIT)

1. Pool 5eu capture 3 samples as described in excel sheet pooling:
2. Load on 1% E-gel with 50 bp ladder. Ensure there is at least 2 lanes of space between ladder and your sample pool. Run until bands are almost near the end.
3. Image before (left) and after (right) gel cut.
4. Add 3x volume ADB Buffer, shake at 50C for at least 10 min (Actual: 25 min).
5. Transfer to Gel Recovery Column, bind, wash 2x with RNA/DNA Wash buffer, dry spin, elute in 10 ul H2O.
6. Qubit 1 ul and Tapestation 1.5 ul against ladder.
7. Paired end. Read 1 (100 bp) - index (8 bp) - Read 2 (100 bp). Final loading 8 pM.

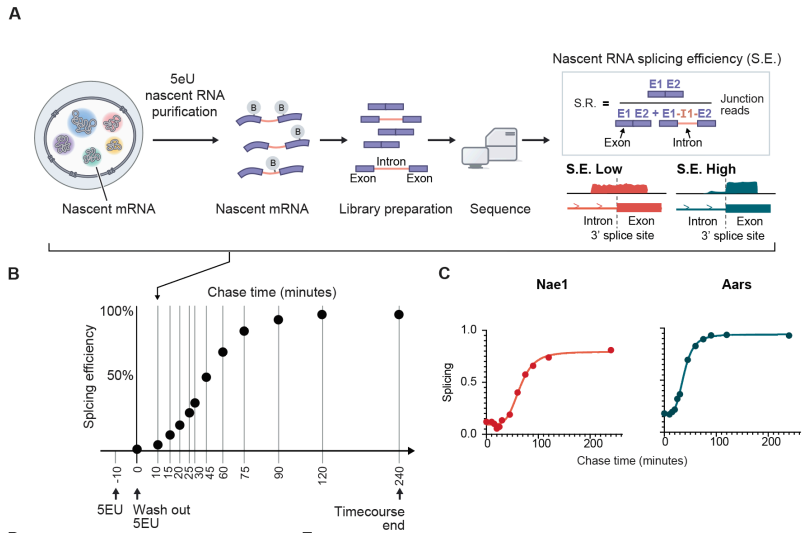


Figure 3: 5EU time course overview with sample data. *(A) Overview of 5EU method and calculation of splicing efficiency. (B) Overview of pulse chase experiment. (C) Splicing ratio over time for two genes Nae1 and Aars.*

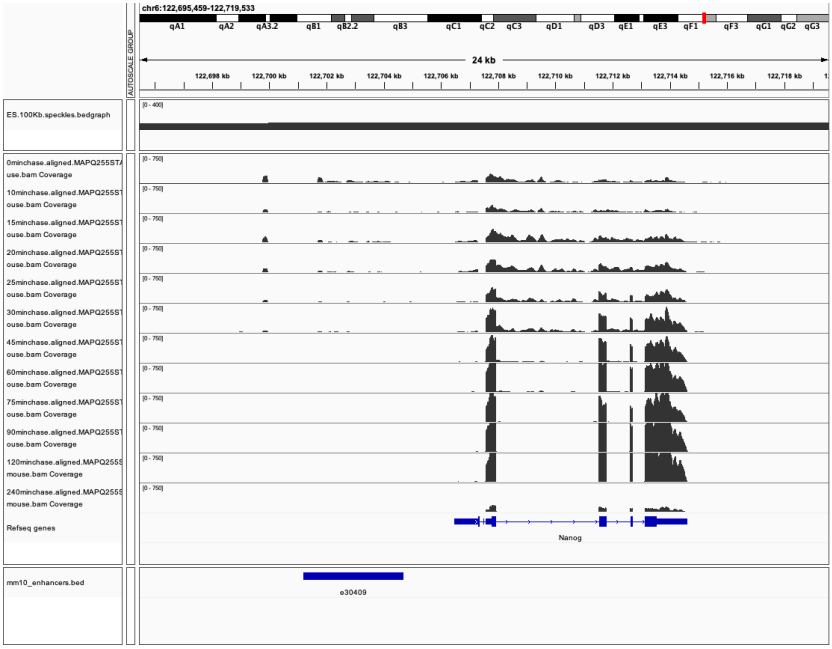


Figure 4: Example of Nanog 5EU pulse chase with annotated enhancer upstream of the promoter.

Chapter 4

3D GENOME ORGANIZATION AROUND NUCLEAR SPECKLES DRIVES MRNA SPLICING EFFICIENCY

Prashant Bhat, Amy Chow, Benjamin Emert, Olivia Ettlin, Sofia A. Quinodoz, Yodai Takei, Wesley Huang, Isabel Goronzy, Mario R. Blanco, Mitchell Guttman

A modified version of this chapter was published as: Bhat, P. *et al.* 3D genome organization around nuclear speckles drives mRNA splicing efficiency. bioRxiv doi: https://doi.org/10.1101/2023.01.04.522632 (In Revision)

"Form follows function - that has been misunderstood. Form and function should be one, joined in a spiritual union." – Frank Lloyd Wright

4.1 ABSTRACT

The nucleus is highly organized such that factors involved in transcription and processing of distinct classes of RNA are organized within specific nuclear bodies. One example is the nuclear speckle, which is defined by high concentrations of protein and non-coding RNA regulators of pre-mRNA splicing. What functional role, if any, speckles might play in the process of mRNA splicing remains unknown. Here we show that genes localized near nuclear speckles display higher spliceosome concentrations, increased spliceosome binding to their pre-mRNAs, and higher co-transcriptional splicing levels relative to genes that are located farther from nuclear speckles. We show that directed recruitment of a pre-mRNA to nuclear speckles is sufficient to increase mRNA splicing levels. Finally, we show that gene organization around nuclear speckles is highly dynamic with differential localization between cell types corresponding to differences in Pol II occupancy. Together, our results integrate the longstanding observations of nuclear speckles with the biochemistry of mRNA splicing and demonstrate a critical role for dynamic 3D spatial organization of genomic DNA in driving spliceosome concentrations and controlling the efficiency of mRNA splicing.

4.2 INTRODUCTION

The nucleus is highly organized such that DNA, RNA and protein molecules involved in transcription and processing of distinct RNA classes (e.g., ribosomal RNA, histone mRNAs, snRNAs, mRNAs) are spatially organized within or near specific nuclear bodies[1–5] (e.g., nucleolus[6,7], histone locus body[8,9], Cajal body[9–11], nuclear speckles[12,13]). While it has been long speculated that nuclear bodies may play a crucial role in RNA biogenesis, such a role has never been directly demonstrated[14–16]. In theory, nuclear bodies could represent structures that are critical for transcription and/or processing of specialized classes of RNA[2] (i.e., structure enables function), or instead they could represent an emergent property whereby regions of shared regulation simply self-assemble in three-dimensional (3D) space (i.e., function results in structure)[17]. Distinguishing between these possibilities has proven challenging because many of the molecular components contained within these nuclear bodies serve dual roles – as catalytic components required for transcription or RNA processing and as structural components required for the integrity of these structures[14–16,18–25].

To explore this question, we focused on the relationship between nuclear structure and mRNA splicing. In higher eukaryotes, most RNA Polymerase II (Pol II) transcribed genes contain intronic sequences that must be removed from precursor messenger RNAs (pre-mRNAs) to generate mature mRNA transcripts[26,27]. mRNA splicing is predominantly co-transcriptional such that nascent pre-mRNAs are spliced as they are transcribed[28–34]. Incomplete splicing yields mRNAs that are degraded by nonsense-mediated decay and results in decreased protein levels[35], and disruption of mRNA splicing is associated with many human diseases[36] including cancer[37–39], neurodegeneration[40–43], and immune dysregulation[44,45]. Due to this central importance, splicing needs to be highly efficient to ensure the fidelity of mRNA and protein production.

Early studies visualizing the localization of mRNA splicing factors – including proteins (e.g., SRRM1, SRSF1, SF3a66) and non-coding RNAs

(e.g., U1, U2)[46,47] – observed that these factors were not uniformly distributed throughout the nucleus but instead were enriched within specific 3D territories referred to as nuclear speckles[48–50]. Because of this preferential localization of splicing regulators, nuclear speckles were initially thought to represent the site of mRNA splicing[12,13,51]. However, this proposal was challenged by the subsequent observations that: (i) DNA and nascent pre-mRNAs are not located within nuclear speckles[52,53], (ii) speckles contain inactive spliceosome components[12,13], (iii) splicing factors diffuse away from speckles to bind nascent pre-mRNAs and catalyze the splicing reaction[54–58], and (iv) transcriptional inhibition does not lead to disruption of the nuclear speckle as would be expected if they were organized around active splicing[49,59,60]. These observations led to the prevailing notion that speckles are not critical for mRNA splicing but instead act as storage assemblies of inactive spliceosomes[61–66]. While numerous additional roles for nuclear speckles have been proposed over the years[1,12,13,16,51,67–75], to date these proposals are primarily based on correlational observations and have yet to be directly tested. Accordingly, despite their initial description over 40 years ago[48–50], what functional role, if any, nuclear speckles might play in the process of mRNA splicing remains unknown[76].

Recently, we developed genome-wide methods to explore the higher-order three-dimensional organization of DNA and RNA in the nucleus[77–79]. Using these and related approaches, we and others identified that nuclear speckles represent major structural hubs that organize interchromosomal contacts corresponding to genomic regions containing highly transcribed Pol II genes[78,80,81] and their associated nascent pre-mRNAs[67,68,79]. Based on these observations, we sought to revisit the role of nuclear speckles in splicing. Specifically, we hypothesized that organization of highly transcribed Pol II genes on the periphery of nuclear speckles would increase the concentration of spliceosomes at these nascent pre-mRNAs, thereby increasing their splicing efficiency. Here we demonstrate an essential role for 3D organization of genomic DNA in controlling the efficiency of mRNA splicing.

4.3 RESULTS

4.3.1 SNRNAS PREFERENTIALLY BIND PRE-MRNAS OF GENES THAT ARE CLOSE TO SPECKLES

To explore DNA localization around the nuclear speckle, we first computed speckle contacts for all genomic regions using both genomic (RNA & DNA SPRITE)[79] and microscopy (seqFISH+)[81] approaches in mouse embryonic stem (ES) cells. We observed that DNA regions that exhibit high SPRITE-based speckle contact frequencies (e.g., Tcf3, Foxj1, and Nrxn2) were preferentially located adjacent to SF3a66, a protein marker of nuclear speckles (**Figure 1A**). Conversely, DNA regions with low SPRITE-based speckle contact frequencies on the same chromosomes (e.g., Grik2, Efemp1, Zfand5) were located farther away from SF3a66 foci **(Figure 1A)**. Comparing 2,460 paired genomic regions, we observed that SPRITE-based speckle contact frequency and DNA distance to SF3a66 were inversely correlated ($r = -0.72$), indicating that SPRITE accurately measures genomic distance to nuclear speckles (**Figure 1B**). We refer to genomic regions with the highest 5% of speckle contact frequencies as speckle close and those with the lowest 5% as speckle far.

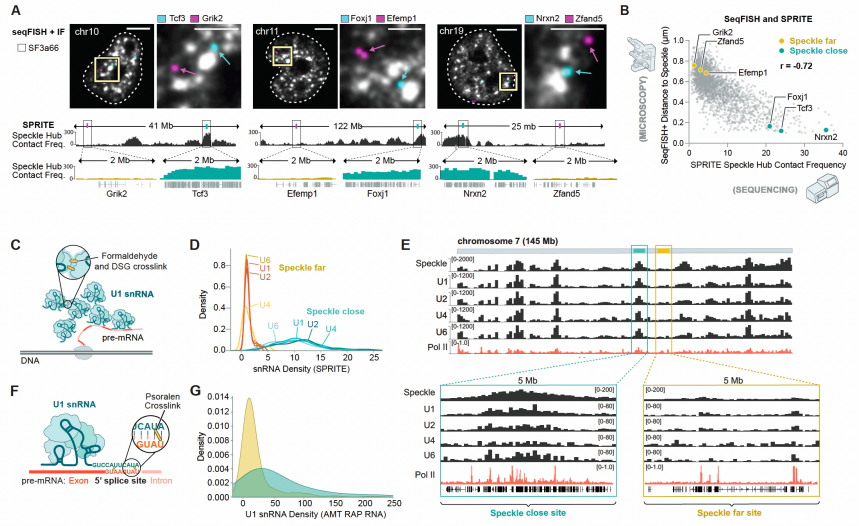
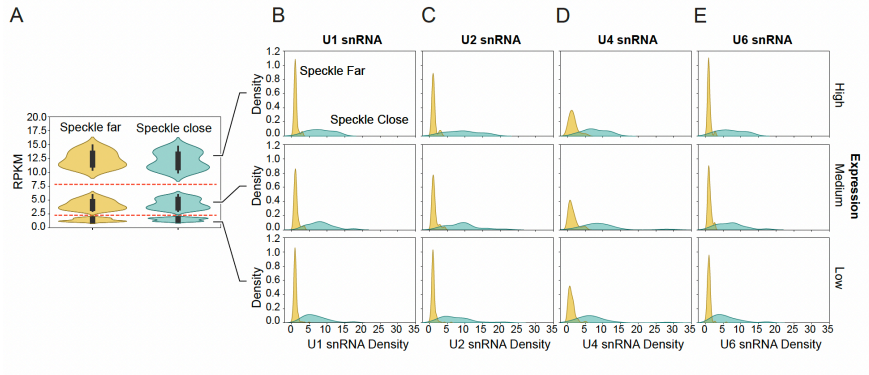


Figure 1 | snRNAs preferentially bind pre-mRNAs of genes that are close to speckles. *(A) Three reconstructed images for DNA seqFISH+ and immunofluorescence (SF3A66) in mouse ES cells comparing speckle close genes (Tcf3, Foxj1, Nrxn2 in blue) and speckle far genes (Grik2, Efemp1, Zfand5 in purple) (top). Images are maximum intensity z-projected for 1 μm section. White lines represent nuclear segmentation. Scale bars in zoom out panels are 5 μm and zoom in panels are 2.5 μm. Speckle contact frequencies from SPRITE for chromosomes 10, 11, and 19 at 100-kb resolution (bottom). Zoom in, speckle contact frequencies from SPRITE for the 2 Mb region around genes shown in top. (B) Genome-wide comparison of DNA seqFISH+ distance to exterior of speckle (μm) and SPRITE speckle hub contact frequency for 2460 paired genomic regions. Pearson r correlation is -0.72. (C) Schematic of types of RNA-DNA interactions captured by SPRITE. Formaldehyde and DSG crosslink nucleic acids and proteins to each other and SPRITE can measure the number, type (DNA or RNA), and sequence of molecules within each crosslinked complex. (D) Normalized density of U1, U2, U4, U6 snRNAs on speckle close versus speckle far genomic regions. Normalization for each snRNA is to the mode of the speckle far distribution to visualize all snRNA densities on the same scale. RPKM for both speckle far and close genes is thresholded between 2.5-7.5. (E) Whole chromosome 7 view of SPRITE contact frequencies at 1-Mb resolution and zoom in views at 100-kb resolution for speckle hub, U1, U2, U4 and U6 snRNAs. Pol II-S2P ChIP-seq density at 100-kb resolution. (F) Schematic of direct RNA-RNA interactions capture by AMT RAP RNA84. Psoralen forms direct crosslinks between RNA-RNA hybrids, affinity purification (not shown) selectively captures U1 snRNA, and all directly hybridized pre-mRNAs. (E) U1 snRNA density from AMT RAP RNA for speckle close versus speckle far regions.*

Having defined genome-wide proximity to nuclear speckles, we explored the localization of the spliceosome – the molecular machinery that carries out splicing and consists of U-rich small nuclear RNAs (snRNAs) and associated proteins[82] – across the genome. We considered two possible models for spliceosome association with pre-mRNA. In the direct-recruitment model, the spliceosome is directly recruited by either Pol II or the nascent pre-mRNA, which would result in the spliceosome associating with transcribed regions proportional to their mRNA levels. Alternatively, in the speckle-recruitment model, the spliceosome would accumulate preferentially at nascent pre-mRNAs that are localized near nuclear speckles.

To test these two models, we mapped the localization of the U1, U2, U4, and U6 snRNAs across the genome using RNA & DNA SPRITE (RD-SPRITE, **Figure 1C**). As expected, these snRNAs are enriched over genomic DNA regions that are actively transcribed into pre-mRNA. However, rather than simply reflecting pre-mRNA levels as is predicted by the direct-recruitment model, we observed that regions that are close to nuclear speckles display ~10-fold higher enrichment of snRNAs, independent of gene expression levels (**Figure 1D, Supplemental Figure 1A-E**). For example, two neighboring genomic regions on mouse chromosome 7 that are transcribed at comparable levels, but are located at different distances relative to speckles, display a ~4-fold difference in snRNA levels (**Figure 1E**). These results indicate that spliceosome concentrations are highest at nascent pre-mRNAs that are in proximity to nuclear speckles.



Supplemental Figure 1 | snRNA density for differently expressed genes. *(A)To ensure that splicing factor difference were not due to expression differences between speckle close and speckle far genes, we divided genes up based on expression ranges: high expression (RPKM > 10), medium expression (RPKM = 2.5-7.5), low expression (RPKM = 0-2.5). The distribution of expression within these ranges were the same for speckle close and speckle far genes. (B) U1 snRNA density is plotted for high (top), medium (middle), and low expression genes (bottom). (C) U2 snRNA density is plotted for high (top), medium (middle), and low expression genes (bottom). (D) U4 snRNA density is plotted for high (top), medium (middle), and low expression genes (bottom). (E) U6 snRNA density is plotted for high (top), medium (middle), and low expression genes (bottom).*

Because RD-SPRITE utilizes protein-protein crosslinking (formaldehyde + DSG) to map RNA-DNA contacts[77], this approach captures associations that are indirect and therefore may not reflect the proportion of pre-mRNAs directly engaged by spliceosomes[78,79] (**Figure 1C**). To measure the number of spliceosomes that directly bind to nascent pre-mRNAs, we used psoralen-mediated crosslinking (which forms covalent crosslinks only between directly hybridized nucleic acids[83]) to map U1 interactions with pre-mRNAs (**Figure 1F**). We previously showed that this approach is highly specific at mapping U1 binding to 5' splice sites at exon-intron junctions[84]. Using this data, we computed the frequency of U1 binding to each pre-mRNA (which accounts for differences in pre-mRNA levels by measuring the number of U1 bound RNAs divided by RNA abundance) and compared U1 binding frequency to the distance between the nascent locus and nuclear speckles. We observed ~3-fold higher levels of U1

binding to pre-mRNAs transcribed from speckle close genes compared to those transcribed from speckle far genes (**Figure 1G**).

Together, these results indicate that proximity of genomic DNA regions to nuclear speckles is associated with increased concentrations of spliceosomes and spliceosome engagement on pre-mRNA.

4.3.2 CO-TRANSCRIPTIONAL SPLICING EFFICIENCY VARIES BASED ON PROXIMITY TO NUCLEAR SPECKLES

We reasoned that higher concentrations of spliceosome components (enzyme) at nascent pre-mRNAs (substrate) located proximal to nuclear speckles would lead to increased co-transcriptional splicing efficiencies (e.g., the proportion of spliced products to total mRNA produced, **Figure 2A**) relative to pre-mRNAs that are located farther from the speckle.

To focus on splicing of pre-mRNAs that occurs near the DNA locus from which they are transcribed (which we refer to as co-transcriptional splicing), we analyzed nascent RNA that is associated with chromatin using a stringent biochemical purification procedure[85,86] (**Figure 2B**). Using these data, we computed the splicing efficiency for each gene by taking the ratio of spliced reads relative to total pre-mRNA reads (spliced counts + unspliced counts) (**Figure 2A**). Overall, we observed that genes located closest to nuclear speckles showed a >2-fold higher splicing ratio compared to genes that are farthest from nuclear speckles (41.0% vs 19.1%) (**Figure 2C-D**). More generally, we observed a strong correlation between speckle contact frequency and splicing efficiency ($r=0.92$, $p<0.0001$, **Figure 2E**). To ensure that this difference in splicing efficiency is not simply due to differences in transcription levels between speckle close and speckle far genes, we measured splicing efficiency only for genes that are transcribed at comparable levels and observed a similar increase in splicing efficiency at genes that are located close to nuclear speckles relative to genes that are farther (**Supplemental Figure 2**).

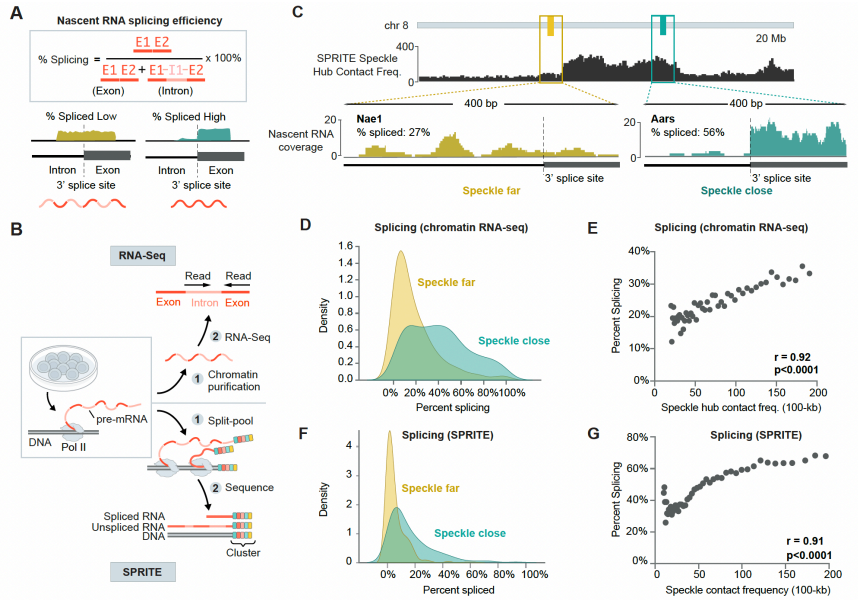
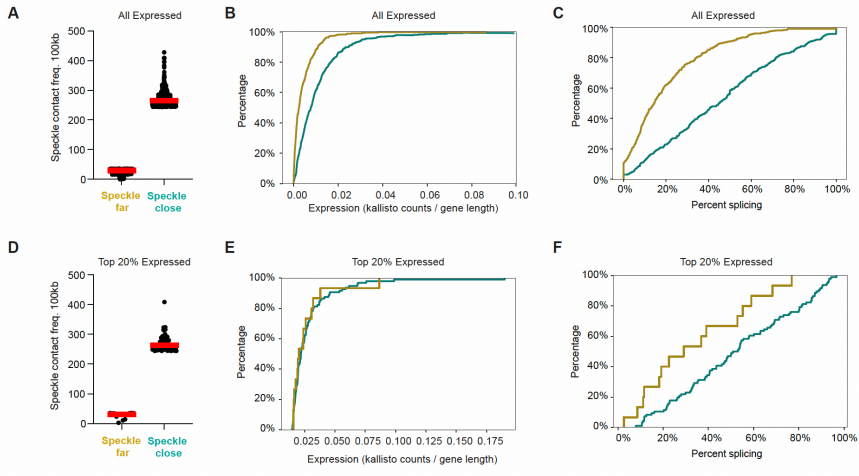


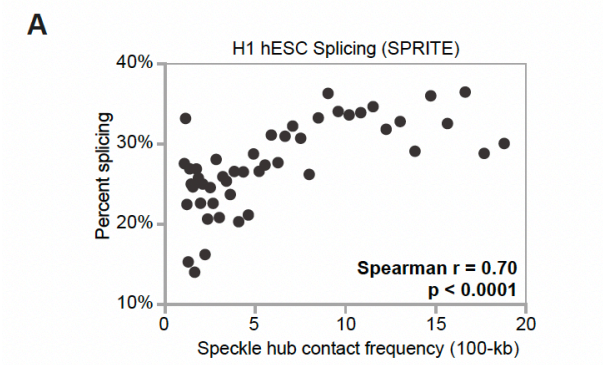
Figure 2 | Co-transcriptional splicing efficiency varies based on proximity to nuclear speckles. *(A) Nascent RNA splicing efficiency calculation. Splicing efficiency of a gene is calculated by taking the ratio of exon to total pre-mRNA counts from RNA sequencing (exons + introns). (B) Schematic of nascent RNA sequencing and SPRITE methods used to measure splicing efficiency. (C) SPRITE speckle hub contact frequency for a 20-Mb region on chromosome 8 (top). Nascent RNA coverage from chromatin RNA sequencing for a speckle far (Nae1) and speckle close (Aars) gene around a single 3'splice site (bottom). Percent spliced across entire gene is 27% (Nae1) and 56% (Aars). (D) Density plot of percent spliced for genes located within speckle close or speckle far 100-kb genomic regions (461 speckle close genes and 460 speckle far genes). (E) SPRITE speckle hub contact frequency (x axis) and percent spliced for genes from nascent RNA sequencing within each bin (y axis) across 50 bins. Each point/bin contains at least 20 genes and reflects the average splicing for that bin. Pearson r correlation = 0.92.(F) Density plot of percent spliced within 100-kb genomic intervals from SPRITE for speckle close and speckle far regions (312 speckle close and 311 speckle far 100-kb regions). (G) SPRITE speckle hub contact frequency (x axis) and percent spliced within genomic bins from SPRITE (y axis) across 50 bins. Each point/bin contains at least 20 regions and reflects the average splicing for that bin. Pearson r correlation = 0.91.*



Supplemental Figure 2 | Higher splicing efficiency in speckle close regions when comparing to genes of similar expression in speckle far regions. *(A) Speckle contact frequencies for speckle close and speckle far genes ≥2 exons. (B) ECDF plot of expression for speckle close and speckle far genes in (A). (C) ECDF plot of splicing efficiency for speckle close and speckle far genes in (A). (D) Speckle contact frequencies for speckle close and speckle far genes corresponding to top 20% of expression in mouse ES cells. (E) ECDF plot of expression for speckle close and speckle far genes in (D). (F) ECDF plot of splicing efficiency for speckle close and speckle far genes in (D).*

To further validate this effect and exclude the possibility that the observed splicing differences might reflect mature mRNA in our biochemical purification, we used an orthogonal method to measure mRNA levels on chromatin. Specifically, we used RD-SPRITE to analyze splicing ratios of RNAs[87] exclusively when they were associated with the DNA of their own nascent locus (**Figure 2B**). We then computed splicing efficiency as the fraction of exons over the total number of exons and introns. Consistent with the chromatin RNA-Seq data, we observed ~3 fold higher splicing in speckle-close (16.1%) to speckle-far (5.5%) regions (**Figure 2F**). Furthermore, we observed a strong correlation between the splicing efficiency per gene and its speckle contact frequency ($r=0.91$, $p<0.0001$; **Figure 2G**). We observed a similar effect when focusing on human H1-hES cells; genes that have higher speckle contact frequencies show higher splicing efficiencies ($r = 0.70$, $p<0.001$; **Supplemental Figure 3**).

Together, these results indicate that the pre-mRNA splicing efficiency is highest for speckle-associated genes and that this splicing efficiency is achieved while the pre-mRNA is bound at its nascent locus.



Supplemental Figure 3 | Splicing efficiency in H1-hESCs correlates with speckle proximity. *(A) SPRITE speckle hub contact frequency at 100-kb resolution (x axis) in H1-hESCs and percent spliced within genomic bins from SPRITE (y axis) across 50 bins. Spearman r correlation = 0.70.*

4.3.3 PRE-MRNA ORGANIZATION AROUND NUCLEAR SPECKLES IS SUFFICIENT TO DRIVE INCREASED MRNA SPLICING

Because genes differ in multiple ways beyond their nuclear speckle proximity (e.g., gene length, alternative splicing patterns, and sequence-specific features), it remains possible that the observed increase in splicing efficiency is due to other gene-specific or genomic DNA features (e.g., chromatin structure[88–93]) that might also correlate with speckle proximity.

To directly test whether speckle proximity drives splicing efficiency, we designed a splicing reporter that can be directly recruited to nuclear speckles, allowing us to measure its splicing efficiency within individual cells. Specifically, we generated a reporter that produces an mRNA that is

translated into GFP when spliced, but not when unspliced (**Figure 3A**). Increased GFP signal reflects increased reporter splicing and can be quantitatively measured within each cell via a fluorescence readout (**Figure 3A**). In the intron of this reporter, we embedded an MS2 bacteriophage RNA hairpin that binds with high affinity to the MS2 bacteriophage coat protein (MCP)[94]. We used this system to localize the pre-mRNA reporter to specific nuclear locations by co-expressing the splicing reporter together with specific MCP-fusion proteins that are known to localize at different locations within the nucleus (**Figure 3B**). Specifically, we expressed SRRM1 and SRSF1, two proteins that localize within nuclear speckles[22,95]. SRRM1 is primarily localized in nuclear speckles (punctate), while SRSF1 exhibits both speckle (punctate) and nucleoplasmic (diffuse) localization. As controls, we expressed several non-speckle proteins, including SRSF3 and SRSF9 (two splicing proteins that are not enriched within nuclear speckles but are localized throughout the nucleoplasm[96,97]) and LBR (a protein that is anchored in the nuclear membrane and associates with the transcriptionally inactive nuclear lamina[98]).

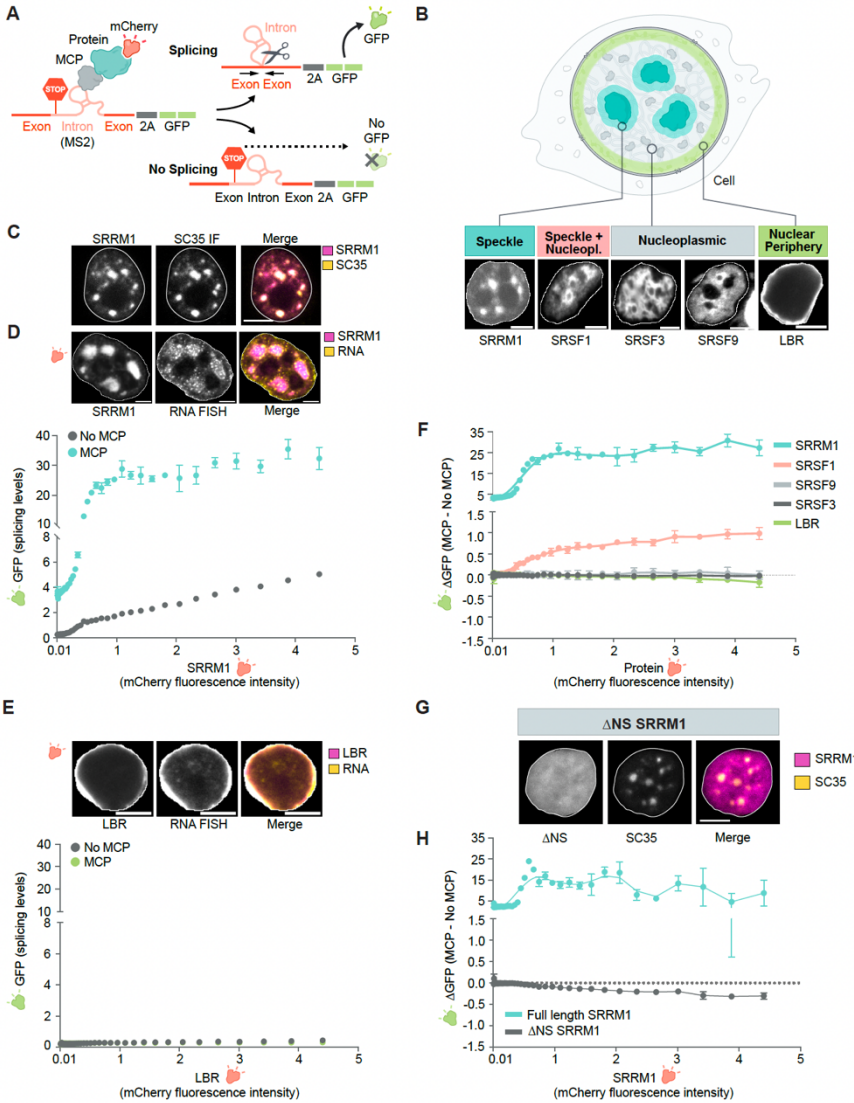


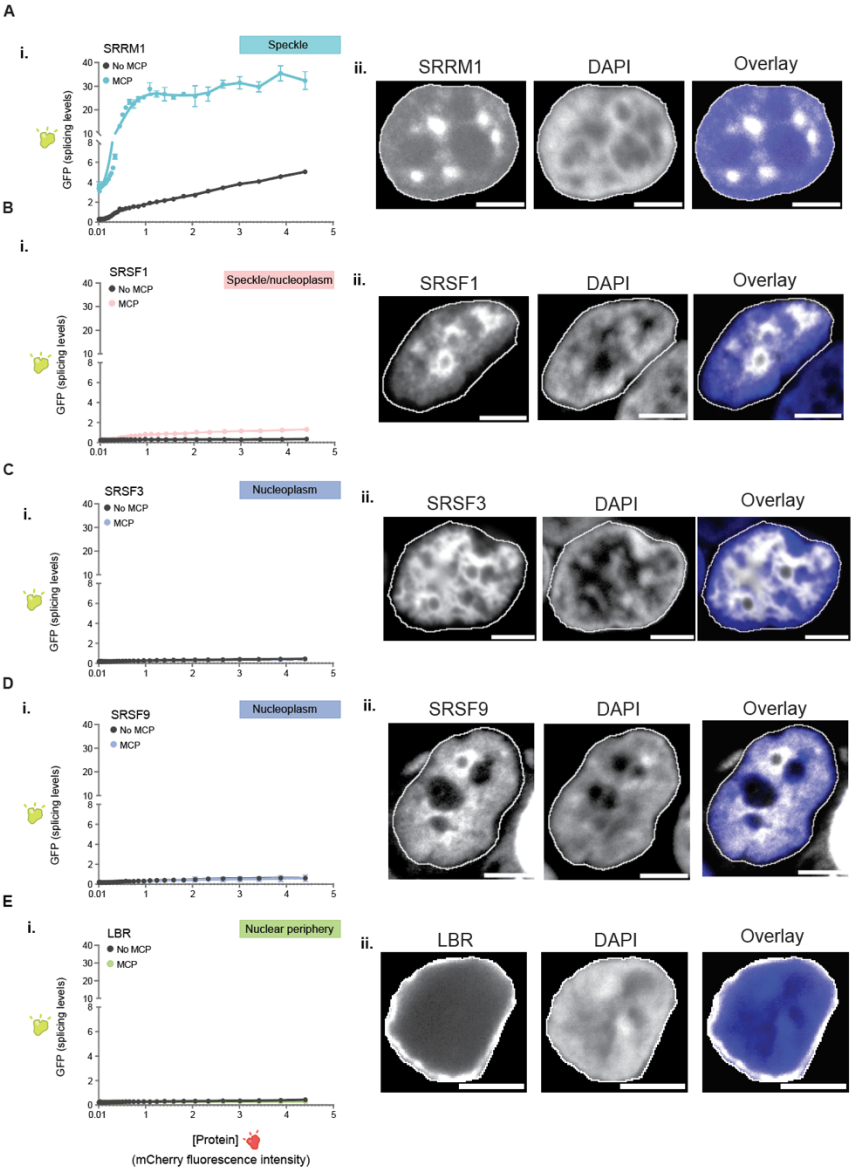
Figure 3 | pre-mRNA organization around nuclear speckles drives splicing efficiency. *(A) Schematic of pre-mRNA splicing assay via a fluorescence based read out. Individual proteins of interest are mCherry-tagged (shown) or without an MCP tag (not shown). MCP protein binds to the complementary MS2 stem loop embedded within the intron of the pre-mRNA reporter. GFP is expressed only when the reporter is spliced and measured via FACS. (B) Schematic of specific nuclear locations (speckle, speckle+nucleoplasm, nucleoplasm, nuclear periphery, top) and mCherry fluorescence of their corresponding*

proteins (SRRM1, SRSF1; SRSF3, SRSF9; LBR, bottom). Nucleus is outlined in white. Scale bar is 5 μm. (C) Fluorescence microscopy for mCherry-SRRM1 (top left). co-immunofluorescence for SC35 (top middle), and merge (top right). Scale bar is 5 μm. (D) Localization of SRRM1+MCP with mCherry reporter and single-molecule RNA FISH. Nucleus is outlined in white. Scale bars, 5 μm (top). GFP levels (y axis) versus fluorescence intensity (levels) of SRRM1 (x axis) (bottom). Error bars are S.E.M for three replicates. (E) Localization of LBR+MCP with mCherry reporter and single-molecule RNA FISH. Nucleus is outlined in white. Scale bars, 5 μm (top). GFP levels (y axis) versus fluorescence intensity (levels) of LBR (x axis) (bottom). Error bars are S.E.M for three replicates. (F) Difference of GFP expression between constructs with MCP and no MCP (y axis) versus mCherry fluorescence intensity (x axis) for all constructs tested. Error bars are S.E.M for three replicates. (G) Fluorescence microscopy for mCherry-SRRM1-ΔNS (bottom left). Immunofluorescence for SC35 (bottom middle), and merge (bottom right). Error bars are S.E.M for three replicates. Scale bar is 5 μm. (H) Difference of GFP expression between SRRM1 full length and SRRM1 ΔNS constructs with MCP and no MCP (y axis) versus mCherry fluorescence intensity (x axis). Error bars are S.E.M.

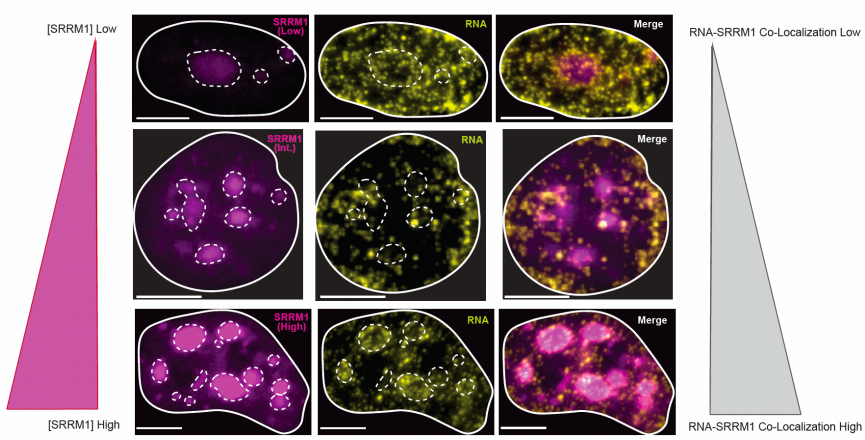
We transfected each of these proteins fused to MCP and mCherry (to directly visualize localization) and, using fluorescence microscopy, confirmed that each protein localized in the nucleus as expected (**Figure 3B**, **Supplemental Figure 4A-E**). We observed that SRRM1-MCP co-localized with endogenous SC35, a well-characterized marker of nuclear speckles (**Figure 3C**), while SRSF3 and SRSF9 localized diffusively throughout the nucleus and LBR localized to the periphery of the nucleus (**Figure 3B, Supplemental Figure 4A-E**). Next, we confirmed that the MS2-containing reporter RNA co-localized along with the MCP fusion protein using RNA FISH coupled with fluorescence microscopy of mCherry (**Figure 3D-4E**). We observed that the MS2-RNA localizes within nuclear speckles when co-expressed with SRRM1-MCP and localizes at the nuclear periphery when co-expressed with LBR-MCP. As expected, cells that express higher concentrations of the MCP-fusion protein exhibit increased co-localization of MS2-RNA (**Supplemental Figure 5**).

Having demonstrated the ability to drive recruitment of an mRNA to a specific nuclear location, we sought to test the impacts of nuclear speckle localization on splicing efficiency. To establish the baseline splicing efficiency and account for non-MCP dependent effects on GFP expression

– including transfection or specific protein-dependent effects – we expressed each protein without MCP. We quantified the relationship between directed recruitment and splicing efficiency by measuring the difference in GFP fluorescence with and without MCP for each protein construct (ΔGFP) relative to expressed protein levels (mCherry).


A
i.
SRRM1
Speckle
No MCP
MCP
GFP (splicing levels)
ii.
SRRM1
DAPI
Overlay
B
i.
SRSF1
Speckle/nucleoplasm
No MCP
MCP
GFP (splicing levels)
ii.
SRSF1
DAPI
Overlay
C
i.
SRSF3
Nucleoplasm
No MCP
MCP
GFP (splicing levels)
ii.
SRSF3
DAPI
Overlay
D
i.
SRSF9
Nucleoplasm
No MCP
MCP
GFP (splicing levels)
ii.
SRSF9
DAPI
Overlay
E
i.
LBR
Nuclear periphery
No MCP
MCP
GFP (splicing levels)
ii.
LBR
DAPI
Overlay
[Protein]
(mCherry fluorescence intensity)

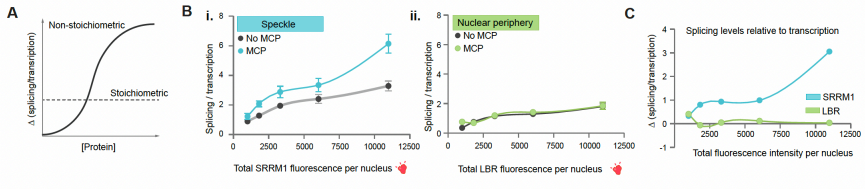
Supplemental Figure 4 | pre-mRNA organization around nuclear speckles drives splicing efficiency. *(i) GFP fluorescence (splicing levels) (y axis) versus mCherry fluorescence intensity for constructs with MCP or without MCP (Left) for: (A) SRRM1, (B) SRSF1, (C) SRSF3, (D) SRSF9, (E) LBR. (ii) Imaging of each protein (A-E) with DAPI and overlay (Right) with nucleus outlined in white. Images are a more complete representation of those displayed in Figure 3B. Scale bars, 5 µm.*



Supplemental Figure 5 | MS2-RNA localization to speckles is concentration dependent. *SRRM1+MCP fluorescence microscopy (left) combined with RNA FISH (middle) and overlay (right). Low (top left), intermediate (middle left), and high (bottom left) SRRM1 expression. Scale bar is 5 µm.*

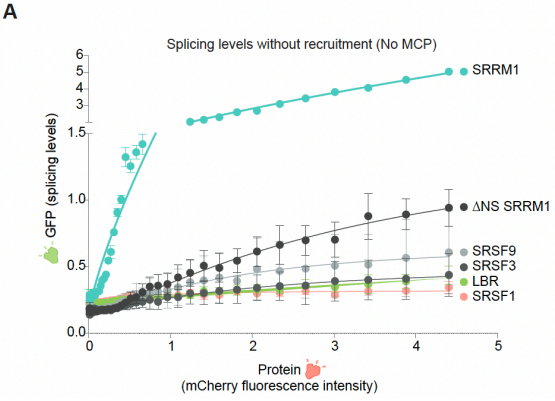
Recruitment of MS2-RNA specifically to speckle proteins SRRM1 or SRSF1 resulted in a non-linear change in GFP levels relative to speckle protein levels (nonlinear four parameter logistic regression; $R^2 = 0.92$ and 0.94, respectively; **Figure 3F; Supplemental Figure 4A and 4B**). To ensure that this observed effect is specifically due to nuclear speckle recruitment, we recruited this MS2-RNA to the diffusely localized splicing proteins SRSF3 and SRSF9 or to the nuclear lamina using LBR. In all cases, we observed that these conditions had no impact on GFP levels (**Figure 3E-3F; Supplemental Figure 4C-4E**). To ensure that the observed increase in GFP levels accurately reflects increased splicing efficiency and not simply higher transcription of the reporter, we compared the change in GFP protein levels to GFP mRNA transcription levels (measured by single molecule FISH). We observed the same non-

linear increase in GFP protein levels relative to transcription levels when the pre-mRNA was recruited to the nuclear speckle, but not when recruited to a control location (**Supplemental Figure 6B – 6C;** See Methods for quantification).



Supplemental Figure 6 | Splicing efficiency relative to transcription around nuclear speckles. *(A) Schematic of two alternative models of how splicing efficiency would change relative to transcription. In the stoichiometric model, the ratio of splicing levels to transcription is directly proportional and constant across all concentrations of speckle protein; in the non-stoichiometric model, the levels of splicing increase relative to transcription across the concentration range of speckles. (B) GFP levels divided by intensity of brightest FISH spot (y axis) versus fluorescence intensity (levels) of SRRM1 (x axis) (bottom). Error bars are S.E.M for at least 3 cells per bin. (C) GFP levels divided by intensity of brightest FISH spot (y axis) versus fluorescence intensity (levels) of LBR (x axis) (bottom). Error bars are S.E.M for at least 3 cells per bin. (D) Difference of normalized GFP expression in B and C between constructs with MCP and no MCP (y axis) versus mCherry fluorescence intensity (x axis) for SRRM1 and LBR constructs.*

These results indicate that direct recruitment of a pre-mRNA to nuclear speckle proteins, but not to other nuclear proteins, is sufficient to increase splicing efficiency. To ensure that this effect is specifically due to the ability of these proteins to localize within the nuclear speckle, we



B

Construct	4PL Nonlinear Fit (R^2)
SRRM1	0.99
ΔNS SRRM1	0.79
SRSF1	0.66
SRSF3	0.43
SRSF9	0.36
LBR	0.58

Supplemental Figure 7 | Splicing activity retained with SRRM1 full length and ΔNS. *(A) GFP levels (y axis) versus fluorescence intensity (levels) of protein constructs with no MCP domain (x axis) (bottom). Error bars are S.E.M for three replicates except ΔNS (four replicates). (B) Table of four-parameter logistic regression fits for all constructs tested.*

expressed a truncated form of SRRM1 that lacks the domain responsible for nuclear speckle localization but retains its previously defined catalytic RNA processing domain[20] (ΔNS-SRRM1; **Figure 3G**). We confirmed that ΔNS-SRRM1 no longer localizes within nuclear speckles (**Figure 3G**). Interestingly, expression of ΔNS-SRRM1 leads to a loss of the MCP-dependent increase in splicing efficiency (ΔGFP) and instead shows a response similar to that observed for other non-speckle-associated proteins (**Figure 3H**). This is true even though ΔNS-SRRM1 retains its splicing activity (**Supplemental Figure 7**).

Together, these results demonstrate that directed recruitment of a pre-mRNA to nuclear speckles leads to a non-linear increase in mRNA splicing efficiency.

4.3.4 DIFFERENTIAL GENE POSITIONING AROUND NUCLEAR SPECKLES CORRESPONDS TO DIFFERENTIAL POL II OCCUPANCY

Previous studies have shown that mRNA splicing is functionally coupled with Pol II transcription, yet how this coupling is achieved is still not fully resolved[99–107]. Because proximity to speckles is associated with increased splicing efficiency and genes that are proximal to speckles tend to be highly transcribed, we reasoned that dynamically organizing actively transcribed genes around nuclear speckles may act to couple Pol II transcription and mRNA splicing.

To address this, we explored whether differential Pol II activity is associated with differential organization of genomic DNA around nuclear speckles. Specifically, we compared genomic DNA organization around nuclear speckles in two distinct mouse cell types with different gene expression programs: mouse ES cells and mouse myoblasts. We generated SPRITE maps in mouse myocytes derived from differentiated MM14 mouse myoblast cells and computed genome-wide nuclear speckle

distances from >14 million SPRITE clusters (**Supplemental Figure 8A-8D**). We observed that DNA regions located close to speckles correspond to genomic regions containing high-density of RNA Pol II in differentiated myocytes (Spearman correlation = 0.69, $p<0.0001$, **Figure 4A-4B**). Importantly, not all highly transcribed Pol II genes organize around the speckle; for example, the Chd2 gene on mouse chromosome 7 contains high levels of Pol II – comparable to that of the nearby Btbd1 gene – yet is located farther from the speckle, likely because Chd2 is transcribed from an otherwise Pol II sparse location (**Figure 4A**).

Next, we compared myocyte speckle distances to those measured in mouse ES cells. Overall, we observe that ~25% of the genome is speckle-proximal in either mouse ES or myocytes. Of these, ~40% are speckle proximal in both cell types whereas ~30% are speckle-proximal only in ESCs and the other ~30% are close to speckles only in myocytes (**Figure 4C)**. Because speckle proximity is correlated with Pol II density, we explored whether the changes in speckle proximity between myocytes and ES cells corresponded to changes in Pol II localization. Indeed, these unique speckle-proximal regions correspond to genomic regions that contain the largest differences in RNA Pol II between myocytes and ES cells (Spearman correlation = 0.52, $p<0.001$, **Figure 4D**). Similarly, genomic regions that are speckle-proximal in ES cells but not in myocytes correspond to regions that contain higher amounts of Pol II in ES cells relative to myocytes. For example, the genomic neighborhood containing the pluripotency marker Nanog is highly expressed and displays high speckle contact frequency in ES cells (**Figure 4E-4F)**. In contrast, Nanog is not expressed in myocytes and the same genomic region displays a low speckle contact frequency (**Figure 4F)**. Conversely, myogenic differentiation leads to widespread transcriptional upregulation of skeletal muscle specific genes, such as Titin (Ttn) and MyoD1 in myocytes. We observed that these regions were highly expressed and located proximal to nuclear speckle hubs in myocytes, whereas these same regions in mES cells were not expressed and were localized away from nuclear speckles (**Figure 4E-5F; Supplemental Figure 8E)**.

To explore these changes more generally, we performed SPRITE on four distinct human cell types: H1 human embryonic stem cells (H1 hESC), H1 ES-derived endoderm cells, human foreskin fibroblasts (HFF-c6), and human lymphoblastoid cells (GM12878) (**Figure 4G; Supplemental Figure 9A-9C**). Similar to the differential speckle contacts observed in mouse, we observed differential speckle localization for genes that are cell type specific. For example, the genomic region containing vimentin (also known as fibroblast intermediate filament) was most speckle proximal in HFF relative to the other three cell types. In contrast, the retinol binding protein 4 (RBP4) was most speckle proximal in ES-derived endoderm cells, consistent with the fact that RPB4 is primarily expressed by endoderm-derived liver[108].

Together, these results demonstrate that differential RNA Pol II activity corresponds to differential gene organization relative to nuclear speckle in distinct cell states (**Figure 4H**).

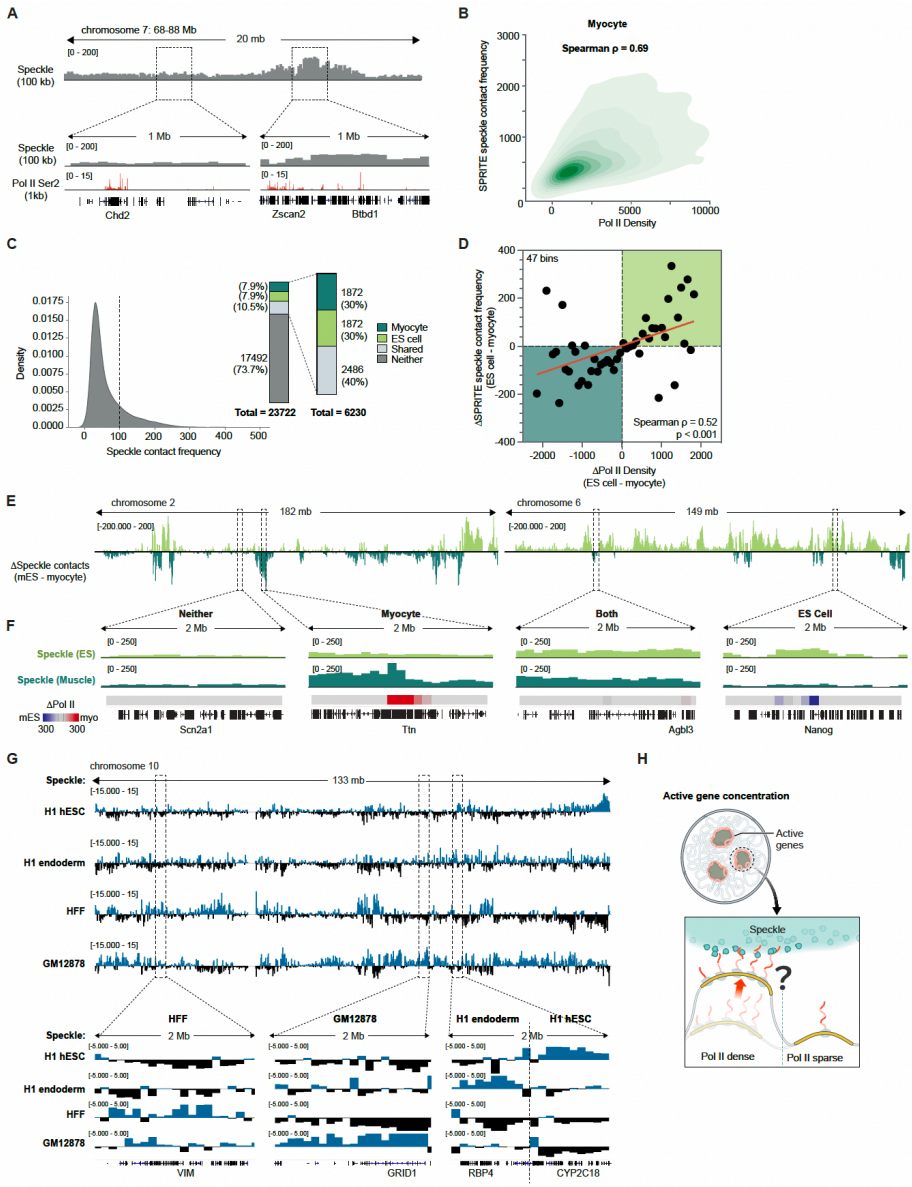
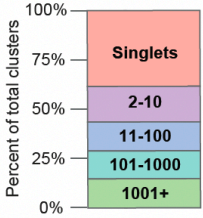


Figure 4 | Differential gene positioning around nuclear speckles corresponds to differential Pol II occupancy. *(A) SPRITE speckle hub contact frequency at 100-kb*

resolution for a 20-Mb region on chromosome 7 in mouse myocytes. Pol II-S2P ChIP-seq density at 1-kb resolution. (B) Ser2-P Pol II density (x axis) and normalized SPRITE speckle contact frequency (100-kb resolution) for myocytes. Spearman correlation = 0.69; p<0.0001. (C) Distribution of SPRITE speckle contact frequencies (100-kb resolution) for normalized mES and myocyte cell SPRITE (left). Distribution of number of genomic regions categorized as speckle hubs in myocyte, ES cells, both, or neither (right). (D) Difference in Ser2-P Pol II density (x axis) versus difference in SPRITE speckle hub contact frequency (y axis) between ES cells and myocytes at 1 Mb resolution. 47 bins. Spearman correlation = 0.52; p<0.001. (E) Difference in speckle hub contact frequency between mESCs (bottom) and myocytes for chromosomes 2 and 6. (F) 2-Mb zoom in regions of speckle contact frequencies and Ser2P Pol II densities for Scn2a1 (speckle in neither), Ttn (myocyte specific), Agbl3 (speckle in both) and Nanog (mES cell specific). (G) Difference in SPRITE speckle hub contact frequency for chromosome 10 between sample and average of the other three samples for each of H1 hESC, H1 endoderm, HFF, and GM12878 human cell lines. Zoom ins are 2-Mb regions of speckle contact frequencies for VIM (HFF specific), GRID1 (GM12878 specific), RBP4 (endoderm specific), and CYP2C18 (H1 hESC specific). (H) Model explaining how Pol II density may act to reposition genomic DNA into proximity with the nuclear speckle.

A



B

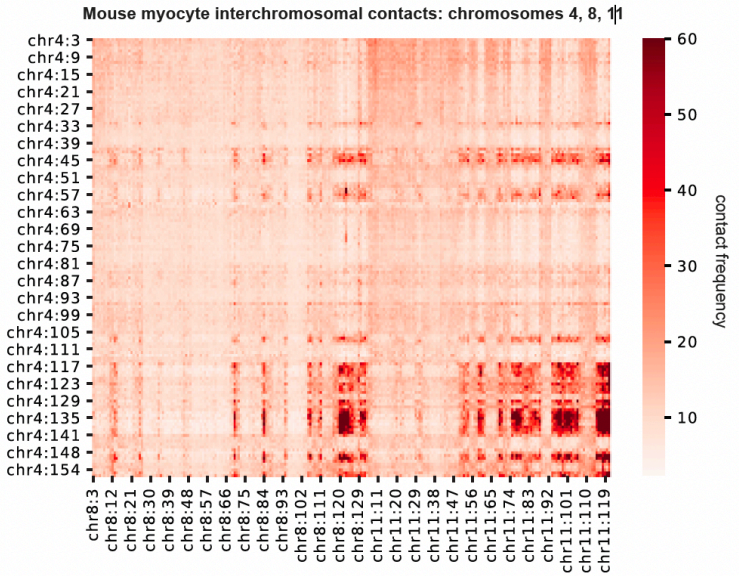
Alignment statistics	60hr
Number of clusters > 2 reads	14,364,072
% reads aligned	73.76%
% unique alignment	56.44%
% multimappers	17.32%

C

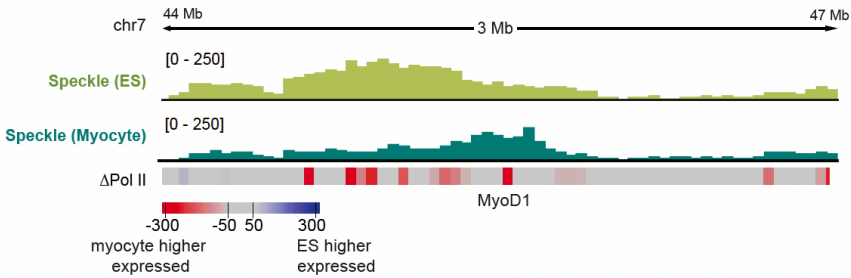
Num. tags	60hr
0 tags	0.20%
1 tags	1.10%
2 tags	7.30%
3 tags	8.90%
4 tags	11.80%
5 tags	70.60%

Ligation efficiency each round = 95%
= (0.71)^(1/5)"

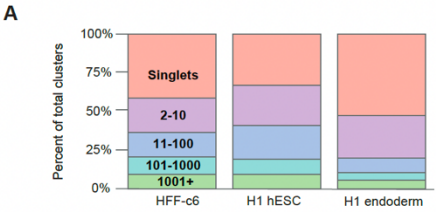
D



E

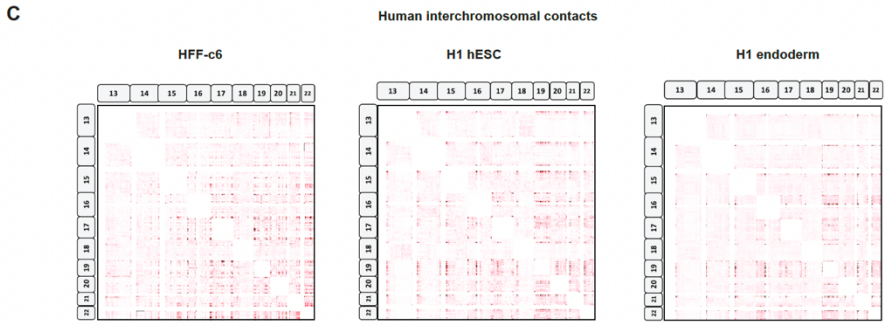


Supplemental Figure 8 | SPRITE analysis of mm14 myocyte cells. *(A) Distribution of SPRITE cluster sizes for myocyte SPRITE. The percentage of reads was calculated for different SPRITE cluster sizes (1, 2-10, 11-100, 101-1000, and over 1001 reads) and reported as the percentage of total reads. Cluster size is defined as the number of reads with the same barcode. (B) Alignment statistics. (C) A summary of ligation efficiency statistics to confirm tags have successfully ligated to each DNA molecule. (D) Mouse myocyte interchromosomal contacts on chromosomes 4, 8, 11. (E) ES cell speckle contact frequency (light green) and skeletal muscle speckle contact frequency (dark green) for genomic locus near MyoD1 (expressed in myocyte). ΔPol II refers to difference in Ser2P-Pol II ChIP seq signal between mES cells and myocytes at 100-kb resolution, red is high in myocyte and blue high in ES.*



B

SPRITE statistics	Number of clusters > 2 reads
HFF-c6	6,626,446
H1 hESC	3,090,232
H1 endoderm	2,632,186



Supplemental Figure 9 | SPRITE analysis of human cells. *(A) Distribution of SPRITE cluster sizes for HFF-c6, H1-hESC, and H1 endoderm SPRITE. The percentage of reads was calculated for different SPRITE cluster sizes (1, 2-10, 11-100, 101-1000, and over 1001 reads) and reported as the percentage of total reads. Cluster size is defined as the number of reads with the same barcode. (B) SPRITE statistics. (C) Human interchromosomal contacts on chromosomes 13 - 22.*

4.4 DISCUSSION

Our results suggest a model that integrates the longstanding observations of nuclear speckles with the biochemistry of mRNA splicing. In this model, nuclear speckles consist of high concentrations of inactive spliceosomes which when activated diffuse away to engage pre-mRNAs[12,13,54,60,61]. When a nascent pre-mRNA is located closer to a speckle, there is a reduced volume through which the active spliceosomes need to diffuse to interact with the pre-mRNA. This decrease in diffusion volume creates a higher concentration of spliceosomes in the vicinity of speckle-close genes and results in increased spliceosome binding to these pre-mRNAs and conversion into spliced mRNA (**Figure 5**).

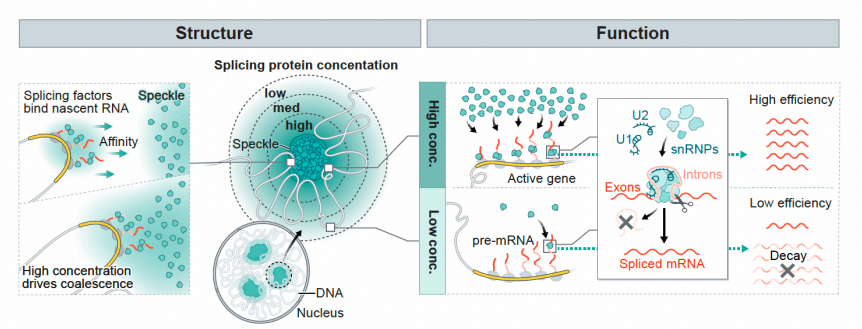


Figure 5 | Integrated model for how gene organization around nuclear speckles impact splicing. *Model of how 3D genome organization drives mRNA splicing. Because nascent pre-mRNAs have high affinity for splicing factors and Pol II dense regions contain the highest concentrations of nascent pre-mRNAs, these genomic regions can achieve multivalent contacts with splicing factors that are enriched within nuclear speckles. Because nuclear speckles contain the highest concentration of these factors within the nucleus, these multivalent contacts may drive coalescence (self-assembly) of these genomic DNA sites with the nuclear speckle. Genomic regions and pre-mRNAs close to nuclear speckles have higher levels of spliceosomes than regions farther away. Locally concentrating pre-mRNAs, genomic DNA, and spliceosomes at speckle-proximal regions leads to increased splicing efficiency whereas a speckle far gene transcribed at the same level is not spliced as efficiently.*

Because speckle proximity is correlated with Pol II density and genes are differentially organized relative to speckles based on transcriptional activity, high levels of transcription may act to reposition genomic DNA closer to nuclear speckles. Because nascent pre-mRNAs have high affinity for splicing factors (including SR proteins and other RNA binding proteins) and Pol II dense regions contain the highest concentrations of nascent pre-mRNAs, these genomic regions may achieve multivalent contacts with splicing factors that are enriched within nuclear speckles. These multivalent contacts may in turn drive coalescence (self-assembly) of these genomic DNA sites with the nuclear speckle[2] (**Figure 5**). Indeed, this self-assembly concept explains how newly transcribed ribosomal DNA genes and snRNA gene loci coalesce into the nucleolus[2,7] and Cajal bodies[17,109,110], respectively. Although RNA Pol II density is associated with speckle proximity[78], not all highly transcribed genes in a cell type are organized around the speckle. Because differential splicing efficiency would impact mRNA and protein levels in a cell, varying genome organization relative to speckles may drive differences in splicing efficiencies and therefore create another dimension of gene expression control.

mRNA splicing and Pol II transcription are known to be kinetically coupled[56,99,106,111] such that increasing the transcription of a gene leads to a non-linear increase in its splicing efficiency (referred to as 'economy of scale' splicing[112]). While individual splicing proteins have been shown to associate with the C-terminal domain of Pol II[99,100,103,113–116] direct binding of splicing factors to Pol II would predict a linear relationship between transcription and splicing and therefore cannot fully explain this coupling. Moreover, Pol II is not sufficient to stimulate splicing efficiency in cellular extracts[117]. This suggests that there must be some additional cellular mechanism required to functionally couple transcription and splicing in cells; our results suggest that this mechanism may be differential gene organization relative to nuclear speckles. Specifically, high levels of Pol II transcription would act to reposition genomic DNA into proximity with the nuclear speckle and increase splicing efficiency at these genes.

Consistent with this notion, it was previously observed that increasing transcription of an individual reporter gene leads to non-linear increases in its splicing efficiency and this coincides with an increased proximity between the gene locus and nuclear speckles[112]. Because the increase in spliceosome concentration achieved at DNA regions positioned at the nuclear speckle would exceed the proportional concentration of the pre-mRNAs transcribed at that locus, this model would explain the observed non-linear increase in splicing efficiency that is achieved when a gene is recruited to the nuclear speckle. In this way, spatial organization around nuclear speckles may act to couple Pol II transcription and mRNA splicing efficiency.

More generally, our results suggest a novel mechanism by which nuclear organization can coordinate regulatory processes in the nucleus and ensure robust non-linear control. Beyond speckles, there are many other bodies that similarly organize RNA processing enzymes with their co-transcriptional DNA and RNA targets[1,2,79]. These compartments include nascent ribosomal RNA loci and rRNA processing factors (e.g., snoRNAs, nucleolin) within the nucleolus[7,118], histone mRNAs and histone processing factors (e.g., U7 snRNA) in histone locus bodies[8,9], and snRNAs and their processing factors (e.g., scaRNAs) within Cajal bodies[10,11,119]. In each of these examples, these nuclear bodies organize around active transcription of the genes that they process[79]. Our results suggest that this structural arrangement may be an important and shared role for coordinating the co-transcriptional efficiency of RNA processing. Specifically, assembling genomic DNA encoding nascent pre-RNAs and their associated regulatory factors within the nucleus could act to increase the local concentration of these factors and therefore couple the efficiency of RNA processing to transcription of these specialized RNAs. This organization would enable localization of these RNA processing enzymes at their targets as they are being produced. The importance of ensuring robust and efficient co-transcriptional processing and coordinating these processes in space and time may explain why all known classes of RNA processing are associated with specialized nuclear bodies and why

disruption of nuclear bodies is a common hallmark in various human diseases[76,120–126].

4.5 METHODS

4.5.1 VISUALIZATION AND QUANTIFICATION OF DNA SEQFISH+ DATA

The DNA seqFISH+ and immunofluorescence data in mouse ES-E14 cells[81] were downloaded from Zenodo (https://zenodo.org/record/3735329#.Y1t7Xuxuf0o). The pseudo-color images of DNA seqFISH+ spots were reconstructed from rounded voxel location of the decoded DNA seqFISH+ spots (seed values of 4 or 5), and then applied with a multidimensional Gaussian filter (sigma = 1) with scipy.ndimage.gaussian_filter package in python 3.7.13. The raw immunofluorescence images of nuclear speckles were reconstructed from csv files that contain intensity values of the SF3A66 antibody in each nucleus. The DNA seqFISH+ and immunofluorescence images were overlaid and contrasted by using ImageJ. The distance between the DNA locus and SF3A66 region was computed as previously described[81].

4.5.2 COMPUTING GENOME-WIDE SPECKLE CONTACT FREQUENCIES FROM SPRITE DATA

We computed genomic DNA distance to the speckle hub using the approach previously described[78]. Briefly, speckle hub regions were defined by clustering all significant inter-chromosomal contacts. We then computed a continuous distance metric for each bin of the genome (we state in figure legends and figures the resolution used: 1kb, 100kb, 1Mb) by identifying all SPRITE clusters containing both the genomic DNA bin and an inter-chromosomal DNA region contained within the speckle hub. We excluded all clusters where the genomic DNA bin and the speckle hub

bin were only contained on the same chromosome in order to ensure that these distances were not driven by local (speckle-independent) contacts. Using these clusters, we then computed a contact score by divided by its respective cluster size (2/total number of reads within the cluster, as previously described[78]) and summed all contact scores for each genomic bin. This produces a continuous contact score for each genomic bin where low scores are farther from the speckle and higher scores are closer to the speckle.

4.5.3 COMPARISON OF SPRITE AND SEQFISH+

To compare SPRITE and SeqFISH+Immunofluorescence measurements, we used SPRITE contact frequencies from contact maps binned at 1 Mb resolution focusing only on SPRITE clusters containing 2 to 1000 reads and down-weighting for cluster size (described above). The distance for SeqFISH+ represented the distance between the DNA spot and the periphery of the SF3A66 domain. When a DNA region and speckle are close, the SeqFISH+ distance is expected to be low and the SPRITE contact frequency is expected to be high.

4.5.4 SNRNA ENRICHMENT CALCULATION FROM RNA & DNA SPRITE

We computed RNA-DNA contacts frequencies for U1, U2, U4, and U6 snRNAs in 1-Mb or 100-kb bins across the genome, weighted by cluster size. For the same 1-Mb bins, we computed each genomic bins' speckle hub contact frequency. To calculate transcription rate, we labeled mES cells for 10 minutes with 5-ethynyl uridine and purified the resulting RNA as previously described[127]. We aligned reads to mm10 and calculated reads per kilobase of gene per million reads (RPKM) mapped for each gene. After this, we computed the median RPKM for the region and filtered bins by expression level (RPKM 1-2 (low), 3-6 (medium), and 10-15 (high).

To compute snRNA enrichment genome-wide, we first computed speckle contact frequencies of each genomic bin at 1 Mb. Next, we computed the contact frequency of U1, U2, U4, and U6 snRNAs for each of these same 1 Mb bins. We then performed a rank normalization of speckle contacts and defined speckle far as the regions corresponding to the lowest 5% of speckle contact frequencies and speckle close as the top 5% of regions. To normalize all values to the same range, the contact score of each snRNA bin value was divided by the median of the speckle far contacts. To compare only regions of equivalent expression, we thresholded regions corresponding to low, medium or high expression. To do this, we computed the median RPKM for the region and filtered bins by expression level (RPKM 1-2 (low), 3-6 (medium), and 10-15 (high). Density plots for speckle close and speckle far regions, for each snRNA, and for each expression level were plotted using the seaborn kde function.

4.6.5 U1 SNRNA ENRICHMENT CALCULATION FROM PSORALEN CROSSLINKING (RAP-RNA AMT)

To compute direct U1 snRNA-pre-mRNA interactions, data from RAP-RNA from AMT crosslinking[84] (GEO IDs: GSM1348350 (input RNA AMT) and GSM1348348 (U1 AMT RAP RNA)) was re-analyzed. In this procedure, cells are treated with a psoralen crosslinker to form direct crosslinks between directly base pair hybridized RNA-RNA sequences. Affinity capture for U1 snRNA and sequencing of associated RNAs identifies the RNAs that were directly bound to U1. To normalize for transcript abundance, input RNA libraries were sequenced in parallel.

To control for U1 occupancy on pre-mRNAs of varying expression, the enrichment of U1 snRNA on pre-mRNAs was computed by dividing the contact frequency for each 100-kb genomic bin in the capture by the input, where the input sample reflects transcription levels. Speckle contact frequencies for the same 100-kb bins were computed as above. Also as above, we performed a rank normalization of speckle contacts and defined the top and bottom 5% of speckle contact frequencies as speckle close and

speckle far. We plotted the density for all speckle close and far regions for the U1 snRNA using the seaborn kde function.

4.5.6 NASCENT SPLICING EFFICIENCY CALCULATION FROM CHROMATIN RNA SEQUENCING

Total chromatin RNA sequencing[128] was re-analyzed from GEO ID: GSM2123095 and re-aligned using the kallisto-bustools workflow[129] to two references separately: a cDNA reference (for exon reads and exon-exon junction reads) and a genomic DNA reference genome (for exon-intron and intron reads). Splicing ratio was computed as the fraction of normalized exon counts over normalized intron + exon (total) counts. We filtered for speckle close and far regions as above and plotted the distribution of percent splicing using the seaborn kde function. For the continuous distribution plot, we plotted all speckle contact frequencies (x-axis) versus the average splicing ratio in each of 50 bins, where each bin contains at least 20 genes.

To calculate splicing efficiency for genes of similar expression, we first computed the normalized expression of genes (≥ 2 exons per gene) by dividing the total counts by the length of the gene. This normalized expression was rank normalized from 0 to 1 and the top 20% of expressed genes were compared. This corresponded to 15 speckle far genes and 96 speckle close genes. For all genes ≥2 exons, this corresponded to 392 speckle far genes and 394 speckle close genes. The empirical cumulative distribution function for expression and splicing efficiency were plotted using the seaborn ecdfplot function.

4.5.7 SPLICING EFFICIENCY CALCULATION FROM RNA & DNA SPRITE (RD-SPRITE)

Because RNA & DNA SPRITE captures interactions occurring between DNA and RNA, we reasoned that any mRNA that was in a SPRITE cluster with its own DNA locus corresponded to nascent chromatin associated RNA. Indeed, we previously showed that this approach accurately captures and quantifies nascent pre-mRNA levels[87]. Using these clusters, we computed splicing efficiency based on the total number of exon reads

in a nascent genomic bin divided by the total number of exons and introns (total pre-mRNA reads) within that same bin. To ensure that we had robust coverage to estimate this frequency, we filtered for genomic regions that contained at least 50 RNA reads (exons + introns). We filtered for speckle close and far regions as above and plotted the distribution of percent splicing using the seaborn kde function. For the continuous distribution plot, we plotted all speckle contact frequencies (x-axis) versus the average splicing ratio in each of 50 bins, where each bin contains at least 20 genes.

4.5.8 PLASMID GENERATION FOR MS2-MCP ASSAY

mCherry-fused, MCP-tagged expression plasmid

The Gateway destination plasmid pCAG-NSTF-DEST-V5 (gift from P. McDonel) was modified by digestion/ligation methods to add mCherry in frame following the V5 tag. This was the -MCP destination vector. To generate the +MCP version, digestion/ligations methods were used to remove the NSTF cassette and to replace it with 2xMCP. These destination vectors were used in Gateway LR recombination reactions with entry clones for each protein of interest. Entry clones were obtained from DNASU.

deltaIDR-SRRM1 entry clone

The SRRM1 entry clone from DNASU was modified using the Q5 site directed mutagenesis kit (New England Biolabs) to delete the predicted disordered region as annotated by Uniprot (https://www.uniprot.org/uniprotkb/Q8IYB3/entry). The resulting clone lacked one additional amino acid at the C-terminus as determined by sanger sequencing of the clone and alignment with the predicted sequence.

4.5.9 IMAGING ANALYSIS FOR MS2/MCP REPORTER ASSAY

4.5.9.1 RNA FISH

To visualize RNA localization in MCP-MS2 recruitment assays, we co-transfected HEK293 cells with splicing reporter and domain recruitment constructs then performed single-molecule RNA FISH as previously described[130]. 24-hours after transfection, we rinsed samples once with 1X PBS then fixed in 4% buffered formaldehyde for 10 minutes at room temperature. Following fixation, we rinsed the samples twice with 1X PBS then permeabilized in 70% ethanol overnight at 4ºC. For hybridization, we rinsed the samples once with wash buffer (10% formamide 2X SSC) then added hybridization buffer (10% formamide, 10% dextran sulfate, 2X SSC) containing RNA FISH probes targeting GFP RNA. These probes were kindly provided by Arjun Raj (University of Pennsylvania). After adding the hybridization solution, we covered samples with glass coverslips and hybridized overnight at 37ºC in a humidified container. Following hybridization, we rinsed the samples once with wash buffer to remove coverslips and then washed twice for 30 minutes at 37ºC. We added 50 µg/mL 4′,6-diamidino-2-phenylindole (DAPI) to the second wash to stain nuclei. Following washes, we rinsed the samples twice with 2X SSC, added SlowFade™ Diamond Antifade solution and proceeded with imaging on a Nikon spinning-disk confocal equipped with Andor Zyla 4.2P sCMOS camera, Nikon LUNF-XL laser unit, and Yokogawa CSU-W1 with 50 µm disk patterns. For each sample, we selected at least ten positions on the basis of DAPI signal and acquired z-stacks at 0.5 µm intervals using a x60 oil objective.

4.5.9.2 IMMUNOFLUORESCENCE

Cells were fixed on coverslips with 4% formaldehyde in PBS for 15 min at room temperature and permeabilized with 0.5% Triton X-100 in PBS for 10 min at room temperature. After washing twice with PBS containing 0.05% Tween (PBSt) and blocking with 2% BSA in PBSt for 30 min, cells were incubated with primary antibodies for anti-SC35 antibody at 1:200 dilution (Abcam, ab11826) overnight at 4 °C in 1% BSA in PBSt. After

overnight incubation at 4 °C, cells were washed three times in 1× PBSt and incubated for 1 h at room temperature with secondary antibodies labeled with Alexa fluorophores (Invitrogen) diluted in 1× PBSt (1:500). Next, coverslips were washed three times in PBSt, rinsed in PBS, rinsed in double-distilled H_2O, mounted with ProLong Gold with DAPI (Invitrogen, P36935) and stored at 4 °C until acquisition.

4.5.9.3 IMAGE ANALYSIS

To quantify RNA recruitment to nuclear lamina or speckles, we used Cellpose (https://github.com/mouseland/cellpose) to segment nuclear boundaries based on DAPI signal and used the Raj Lab smFISH pipeline (https://github.com/arjunrajlaboratory/rajlabimagetools) to localize intranuclear reporter RNA[130,131]. We then quantified mCherry fluorescence intensity at the position of each reporter RNA molecule. To account for heterogeneity in mCherry expression across cells, we calculated the rank pixel intensity to measure relative RNA-mCherry colocalization across conditions. We note that expression heterogeneity precluded us from segmenting speckle domains consistently across cells. In addition, to account for heterogeneity in co-transfection efficiency, we had a blinded author manually select non-mitotic cells co-transfected with both the splicing reporter and the domain recruitment construct.

Due to the sequence and length (GUACAUCUGGUCCAUCCUUCCUAGCUGCGUCCUGGUGGCGCAGGUGUGGGGGAUCGGCAGGUGCCUACCACUAUGCUGUCUAUUACAG; 88 nucleotides;) the intron in our splicing reporter our splicing reporter, we were unable to design smFISH probes selectively targeting nascent RNA. Instead, we used a probeset targeting exons present in both nascent and mature RNA. Since only nascent (unspliced) RNA contain the MS2 hairpin, our results likely underestimate the extent of reporter RNA recruitment.

4.5.10 OVEREXPRESSION OF MS2/MCP CONSTRUCTS IN HEK293T

For MS2/MCP experiments that required a wide range of protein expression, human HEK293T cells were used instead of mESCs because they allow for a wide range of expression levels and enabled investigation of the effect of varying concentrations of proteins (with and without recruitment) on splicing efficiency.

HEK293T cells were cultured in complete media consisting of DMEM (GIBCO, Thermo Fisher Scientific) supplemented with 10% FBS (Seradigm Premium Grade HI FBS, VWR), 1X penicillin-streptomycin (GIBCO, Thermo Fisher Scientific), 1X MEM non-essential amino acids (GIBCO, Thermo Fisher Scientific), 1 mM sodium pyruvate (GIBCO, Thermo Fisher Scientific) and maintained at 37°C under 5% CO2. For maintenance, 800,000 cells were seeded into 10 mL of complete media every 3-4 days in 10 cm dishes.

To assess splicing efficiency of the MS2 splicing reporter, exons 5-6 of mouse IRF7 (ENMUST00000026571.10) containing its endogenous intron were fused upstream of 2A self-cleaving peptide and eGFP and cloned into an MSCV vector (PIG, Addgene)[132]. This splicing reporter has a stop codon embedded within the intron, thereby only when the reporter is spliced will eGFP be translated. An MS2 stem loop was introduced into the intron to enable recruitment of the nascent pre-mRNA splicing reporter specifically to MCP-tagged proteins. The MS2 and tagged protein constructs were co-transfected into HEK293Ts. Splicing, as measured by GFP fluorescence, was assayed 24 and 48 hours after transfection by flow cytometry (Macsquant) and analyzed using FloJo analysis software. Transfections were performed using BioT transfection reagent (Bioland) according to the manufacturer's recommendations. Transfected constructs included SRRM1, SRSF1, SRSF3, SRSF9, and LBR; all constructs were fused to a C terminal mCherry tag. Constructs harboring the MCP tag were fused to two tandem repeats of the MCP peptide at the N terminus.

4.5.11 GFP EXPRESSION AS A FUNCTION OF VARIOUS PROTEINS FUSED TO MCHERRY

For each construct (+/- MCP), we sorted on GFP and mCherry (doubly transfected cells). Because each protein expressed is fused to an mCherry, we assumed that the increase in mCherry fluorescence is proportional to the concentration of the protein of interest within the cell. As a control, we also sorted cells that contained constructs expressing GFP only or mCherry only to ensure there was no spillover of the fluorescence detection between constructs. Additionally, we sorted untransfected cells to set a baseline threshold to filter out cells with background autofluorescence. To that end, because the range of expression is variable due to differences in transfection efficiency/etc, we thresholded cells that contained the same range of mCherry fluorescence intensity (between 0 and 5) and contained non-zero GFP values. The upper threshold of 5 for mCherry fluorescence was chosen because that represents the upper bound of mCherry expression for the protein construct with the overall lowest levels of expression (SRRM1 + MCP). Next, because most points for all constructs were in the lower range of mCherry fluorescence (0-1), mCherry fluorescence (x axis) for each construct was logarithmically binned to 50 bins between 0-5 and the average GFP value for each bin was plotted. Each construct had at least three replicates. Data were merged after binning and the S.E.M. is plotted.

4.5.12 DIFFERENCE IN GFP EXPRESSION CALCULATIONS

For each x-value (50 mCherry bins), the difference in average GFP fluorescence was computed between MCP and no MCP constructs. The average difference of at least three replicates were plotted for all constructs with S.E.M.

4.5.13 NON-LINEAR REGRESSION STATISTICS

Data from each construct (ΔGFP for SRRM1, SRSF1, SRSF3, SRSF9, and LBR or GFP levels for constructs without the MCP domain) were fitted using a four-parameter logistic curve and goodness of fit was calculated using GraphPad Prism 9 software.

4.5.14 QUANTIFICATION OF GFP SPLICING LEVELS NORMALIZED BY TRANSCRIPTION LEVELS

We assessed how GFP splicing levels changed relative to transcription of the MS2 reported plasmid. To do this, we analyzed our SRRM1 and LBR microscopy data which is composed of mCherry to label the protein constructs, RNA FISH to label the mRNA reporter, and GFP protein as a marker of splicing levels. We quantified the intensity of GFP per cell as a proxy for splicing and the intensity of the brightest RNA FISH spot corresponding to transcription levels (Ding et al NSMB). We normalized the GFP protein levels by the intensity of the brightest transcription site and quantified this ratio as a function of mCherry protein levels (i.e., SRRM1+MCP fused to mCherry versus SRRM1-MCP fused to mCherry) for five logarithmically spaced bins. Additionally, as above, we set a threshold for each construct for the same range of mCherry expression to ensure that the change of GFP relative to transcription corresponded to the same mCherry bins across samples.

4.5.15 MYOBLAST CELL CULTURE AND DIFFERENTIATION

C2C12 mouse skeletal myoblasts were passaged at 50-60% confluency every 1-2 days using the Wold lab protocol: (https://www.encodeproject.org/documents/a5f5c35a-cdda-4a45-9742-22e69ff50c9c/@@download/attachment/C2C12_Wold_protocol.pdf). Undifferentiated myoblasts grow in growth medium (20% fetal bovine serum). Myogenic differentiation was initiated upon reaching confluence by switching the cells to medium containing 2% horse serum supplemented with insulin. Differentiation was performed for 60 hours by rinsing fully confluent cells once with PBS and adding
25mL of low-serum differentiation medium. Fresh differentiation medium was changed every 24
hours up to the 48h timepoint and 12 hours afterward were crosslinked using SPRITE crosslinking procedures[77].

4.5.16 HUMAN CELL CULTURE

HFFc6 cells were cultured in DMEM medium supplemented with 20% heat-inactivated FBS. Cells were crosslinked according to our previous SPRITE crosslinking procedure[77]. Details of culture conditions are available on the 4DN portal https://data.4dnucleome.org/biosources/4DNSRC6ZVYVP/.

H1 hESC cells were maintained on matrigel matrix (Corning, 354277) in feeder free media using mTeSR1 (Stemcell Tech, 85850). Every 4-5 days cells were passaged using ReLeSR reagent (Stemcell tech, 05872).

4.5.17 H1 HESC DIFFERENTIATION TO DEFINITIVE ENDODERM

A detailed protocol from Maéhr lab is available on the 4DN portal (https://data.4dnucleome.org/protocols/680ed3dd-04aa-49bc-aac0-8c88da6fddb6/).

Briefly, H1 hESC cells were grown to 80-90% confluency, dissociated into single cells, pelleted and resuspended in mTeSR1 supplemented with 1uM Y27632 (Tocris, 1254). Cells were seeded onto a 6 well coated plate with Growth Factor Reduced Matrigel. On day 1, cells were fed with mTeSR1 and incubated for 24hrs. On day 2, cells were changed with fresh media containing RMPI1640 (Thermo, 21870) supplemented with 0.2% Hyclone FBS (GE Healthcare, SH30070.03) 100 ng/mL Activin A (R&D Systems, 338-AC-01M), 3 μM CHIR 99021 (Tocris, 4423), and 50 nM PI 103 (Tocris, 2930) and incubated for 24hrs. On day 3, cells were changed with fresh media containing RPMI1640 supplemented with 0.2% Hyclone FBS, 100 ng/mL Activin A, and 250 nM LDN-193189 (Tocris, 6053) and incubated for 24hrs. On day 4, Cells were changed again with fresh media containing RPMI1640 supplemented with 0.2% Hyclone FBS, 100 ng/mL Activin A, and 250 nM LDN-193189 (Tocris, 6053). On day 5, cells were crosslinked with SPRITE crosslinking procedures as previously described.

4.5.18 SPRITE CLUSTER SIZE CALCULATIONS

DNA SPRITE and RNA & DNA-SPRITE were performed as previously described[79]. 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. GM12878 SPRITE data was generated previously[78].

4.5.19 SPECKLE HUB DEFINITION

We computed speckle hub contacts from the myocyte data and human cells using the same approaches as previously described[78]. Briefly, the speckle hubs were defined by computing all inter-chromosomal contacts from DNA SPRITE at 1Mb resolution. Using these contacts, we computed p-values to identify all significant inter-chromosomal contacts and clustered these regions. As observed in the mouse ES data, we identified two, mutually exclusive sets of DNA regions, one of these two sets corresponds to the speckle hub and the other being the nucleolar hub. We used this speckle hub region to compute the speckle distance for each region of the genome by computing the number of SPRITE clusters containing the genomic DNA region and at least one of the regions contained within the nuclear speckle hub. To exclude this calculation being dominated by linearly proximal contacts on the same chromosome, we only counted clusters if they contained the genomic region of interest and a speckle hub region that was not contained on the same chromosome. Overall, the distribution of speckle hub scores across 1 megabase genomic regions are similar between mouse ES cells and myocytes, although the precise regions differ (see “SPRITE speckle hubs contact frequency” section).

4.5.20 COMPARING SPRITE DATASETS

To map and compare speckle hub contact frequencies (mouse ES vs myocyte; human SPRITE datasets) in each cell type, we performed a quantile normalization of the speckle hub contacts for each cell line to account for differences in coverage for each SPRITE.

4.5.21 DATA VISUALIZATION

Scatter plots were generated using GraphPad Prism (v9.5.1) and kernel density plots were generated using the Seaborn package. Sequencing data was visualized using IGV (v2.9.4).

4.5.22 STATISTICS AND REPRODUCIBILITY

Data are presented as mean ± S.E.M, or as indicated in the figure legends. Statistical analyses were performed using two-sided z-tests. Methods and details on individual statistical analyses and tests can be found in the respective figure legends. The number of times individual experiments were replicated is noted in their respective figure legends. For SPRITE experiments, one replicate mouse myocyte, HFF-c6, and H1 endoderm were performed and two replicates for H1-hESCs.

4.6 REFERENCES

1. Bhat, P., Honson, D. & Guttman, M. Nuclear compartmentalization as a mechanism of quantitative control of gene expression. *Nat Rev Mol Cell Biol* **22**, 653–670 (2021).
2. Quinodoz, S. A. & Guttman, M. Essential Roles for RNA in Shaping Nuclear Organization. *Cold Spring Harb Perspect Biol* **14**, (2022).
3. Dundr, M. & Misteli, T. Biogenesis of nuclear bodies. *Cold Spring Harbor perspectives in biology* vol. 2 Preprint at https://doi.org/10.1101/cshperspect.a000711 (2010).
4. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nat Rev Mol Cell Biol* **16**, 245–257 (2015).
5. Strom, A. R. & Brangwynne, C. P. *The liquid nucleome - phase transitions in the nucleus at a glance*. *Journal of cell science* (2019). doi:10.1242/jcs.235093.
6. Pederson, T. The nucleolus. *Cold Spring Harb Perspect Biol* **3**, (2011).
7. Lafontaine, D. L. J., Riback, J. A., Bascetin, R. & Brangwynne, C. P. The nucleolus as a multiphase liquid condensate. *Nature Reviews Molecular Cell Biology* vol. 22 165–182 Preprint at https://doi.org/10.1038/s41580-020-0272-6.
8. Marzluff, W. F. & Koreski, K. P. Birth and Death of Histone mRNAs. *Trends in Genetics* vol. 33 745–759 Preprint at https://doi.org/10.1016/j.tig.2017.07.014 (2017).
9. Nizami, Z., Deryusheva, S. & Gall, J. G. The Cajal body and histone locus body. *Cold Spring Harbor perspectives in biology* vol. 2 Preprint at https://doi.org/10.1101/cshperspect.a000653 (2010).
10. Ogg, S. C. & Lamond, A. I. Cajal bodies and coilin - Moving towards function. *Journal of Cell Biology* vol. 159 17–21 Preprint at https://doi.org/10.1083/jcb.200206111 (2002).
11. Machyna, M., Neugebauer, K. M. & Staněk, D. Coilin: The first 25 years. *RNA Biol* (2015) doi:10.1080/15476286.2015.1034923.

12. Spector, D. L. & Lamond, A. I. Nuclear speckles. *Cold Spring Harb Perspect Biol* **3**, 1–12 (2011).
13. Lamond, A. I. & Spector, D. L. Nuclear speckles: a model for nuclear organelles. *Nat Rev Mol Cell Biol* **4**, 605–612 (2003).
14. Adriaens, C. *et al.* Blank spots on the map: some current questions on nuclear organization and genome architecture. *Histochem Cell Biol* **150**, 579–592 (2018).
15. Shachar, S. & Misteli, T. Causes and consequences of nuclear gene positioning. *J Cell Sci* **130**, 1501–1508 (2017).
16. Belmont, A. S. Nuclear Compartments: An Incomplete Primer to Nuclear Compartments, Bodies, and Genome Organization Relative to Nuclear Architecture. *Cold Spring Harb Perspect Biol* **14**, (2022).
17. Shevtsov, S. P. & Dundr, M. Nucleation of nuclear bodies by RNA. *Nat Cell Biol* **13**, 167–173 (2011).
18. Wagner, S., Chiosea, S. & Nickerson, J. A. The spatial targeting and nuclear matrix binding domains of SRm160. *Proc Natl Acad Sci U S A* **100**, 3269–3274 (2003).
19. Pereira, T. A., Ramos, D. N. & Lopez, R. F. V. Hydrogel increases localized transport regions and skin permeability during low frequency ultrasound treatment. *Sci Rep* **7**, 44236 (2017).
20. Szymczyna, B. R. *et al.* Structure and function of the PWI motif: a novel nucleic acid-binding domain that facilitates pre-mRNA processing. *Genes Dev* **17**, 461–475 (2003).
21. Sharma, A., Takata, H., Shibahara, K., Bubulya, A. & Bubulya, P. A. Son is essential for nuclear speckle organization and cell cycle progression. *Mol Biol Cell* **21**, 650–663 (2010).
22. Ilık, İ. A. *et al.* Son and srrm2 are essential for nuclear speckle formation. *Elife* **9**, 1–48 (2020).
23. Blencowe, B. J., Issner, R., Nickerson, J. A. & Sharp, P. A. A coactivator of pre-mRNA splicing. *Genes Dev* **12**, 996–1009 (1998).

24. Blencowe, B. J., Nickerson, J. A., Issner, R., Penman, S. & Sharp, P. A. Association of nuclear matrix antigens with exon-containing splicing complexes. *J Cell Biol* **127**, 593–607 (1994).
25. Blencowe, B. J. *et al.* The SRm160/300 splicing coactivator subunits. *RNA* **6**, 111–120 (2000).
26. Deutsch, M. & Long, M. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res* **27**, 3219–3228 (1999).
27. Gehring, N. H. & Roignant, J.-Y. Anything but Ordinary – Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. *Trends in Genetics* **37**, 355–372 (2021).
28. Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**, 1128–1133 (2009).
29. Reimer, K. A., Mimoso, C. A., Adelman, K. & Neugebauer, K. M. Co-transcriptional splicing regulates 3′ end cleavage during mammalian erythropoiesis. *Mol Cell* (2021) doi:10.1016/j.molcel.2020.12.018.
30. L., W. *et al.* Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res* **22**, 2031–2042 (2012).
31. Khodor, Y. L. *et al.* Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes Dev* **25**, 2502–2512 (2011).
32. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18**, 1435–1440 (2011).
33. Khodor, Y. L., Menet, J. S., Tolan, M. & Rosbash, M. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *RNA* **18**, 2174–2186 (2012).
34. Herzel, L., Ottoz, D. S. M., Alpert, T. & Neugebauer, K. M. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. *Nature Reviews Molecular Cell Biology* Preprint at https://doi.org/10.1038/nrm.2017.63 (2017).

35. Lewis, B. P., Green, R. E. & Brenner, S. E. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc Natl Acad Sci U S A* **100**, 189–192 (2003).
36. Xiong, H. Y. *et al.* RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
37. Wang, E. & Aifantis, I. RNA Splicing and Cancer. *Trends Cancer* **6**, 631–644 (2020).
38. Bonnal, S. C., López-Oreja, I. & Valcárcel, J. Roles and mechanisms of alternative splicing in cancer — implications for care. *Nat Rev Clin Oncol* **17**, 457–474 (2020).
39. Frankiw, L., Baltimore, D. & Li, G. Alternative mRNA splicing in cancer immunotherapy. *Nat Rev Immunol* **19**, 675–687 (2019).
40. Nik, S. & Bowman, T. V. Splicing and neurodegeneration: Insights and mechanisms. *Wiley Interdiscip Rev RNA* **10**, e1532 (2019).
41. Li, D., McIntosh, C. S., Mastaglia, F. L., Wilton, S. D. & Aung-Htut, M. T. Neurodegenerative diseases: a hotbed for splicing defects and the potential therapies. *Transl Neurodegener* **10**, 16 (2021).
42. Ling, S.-C., Polymenidou, M. & Cleveland, D. W. Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. *Neuron* **79**, 416–438 (2013).
43. Pascual, M., Vicente, M., Monferrer, L. & Artero, R. The Muscleblind family of proteins: an emerging class of regulators of developmentally programmed alternative splicing. *Differentiation* **74**, 65–80 (2006).
44. Anna, A. & Monika, G. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet* **59**, 253–268 (2018).
45. Yabas, M., Elliott, H. & Hoyne, G. F. The Role of Alternative Splicing in the Control of Immune Homeostasis and Cellular Differentiation. *Int J Mol Sci* **17**, 3 (2015).

46. Huang, S. & Spector, D. L. U1 and U2 small nuclear RNAs are present in nuclear speckles. *Proc Natl Acad Sci U S A* **89**, 305–8 (1992).
47. Matera, A. G. & Ward, D. C. Localization of the Human Sp1 Transcription Factor Gene to 12q13 by Fluorescence in Situ Hybridization. *Genomics* **17**, 793–794 (1993).
48. Perraud, M., Gioud, M. & Monier, J. C. [Intranuclear structures of monkey kidney cells recognised by immunofluorescence and immuno-electron microscopy using anti-ribonucleoprotein antibodies (author's transl)]. *Ann Immunol (Paris)* **130C**, 635–647 (1979).
49. Spector, D. L., Schrier, W. H. & Busch, H. Immunoelectron microscopic localization of snRNPs. *Biol Cell* **49**, 1–10 (1983).
50. Lerner, E. A., Lerner, M. R., Janeway, C. A. J. & Steitz, J. A. Monoclonal antibodies to nucleic acid-containing cellular constituents: probes for molecular biology and autoimmune disease. *Proc Natl Acad Sci U S A* **78**, 2737–2741 (1981).
51. Hall, L. L., Smith, K. P., Byron, M. & Lawrence, J. B. Molecular anatomy of a speckle. in *Anatomical Record - Part A Discoveries in Molecular, Cellular, and Evolutionary Biology* (2006). doi:10.1002/ar.a.20336.
52. Cmarko, D. *et al.* Ultrastructural analysis of transcription and splicing in the cell nucleus after bromo-UTP microinjection. *Mol Biol Cell* **10**, 211–223 (1999).
53. Thiry, M. The interchromatin granules. *Histol Histopathol* **10**, 1035–1045 (1995).
54. Huang, S. & Spector, D. L. Intron-dependent recruitment of pre-mRNA splicing factors to sites of transcription. *J Cell Biol* **133**, 719–732 (1996).
55. Misteli, T., Cáceres, J. F. & Spector, D. L. The dynamics of a pre-mRNA splicing factor in living cells. *Nature* **387**, 523–527 (1997).

56. Misteli, T. & Spector, D. L. RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo. *Mol Cell* **3**, 697–705 (1999).
57. Jiménez-García, L. F. & Spector, D. L. In vivo evidence that transcription and splicing are coordinated by a recruiting mechanism. *Cell* **73**, 47–59 (1993).
58. Xie, S. Q., Martin, S., Guillot, P. V., Bentley, D. L. & Pombo, A. Splicing speckles are not reservoirs of RNA polymerase II, but contain an inactive form, phosphorylated on serine2 residues of the C-terminal domain. *Mol Biol Cell* (2006) doi:10.1091/mbc.E05-08-0726.
59. Melcák, I. *et al.* Nuclear pre-mRNA compartmentalization: trafficking of released transcripts to splicing factor reservoirs. *Mol Biol Cell* **11**, 497–510 (2000).
60. Spector, D. L., Fu, X. D. & Maniatis, T. Associations between distinct pre-mRNA splicing components and the cell nucleus. *EMBO J* **10**, 3467–3481 (1991).
61. Mattaj, I. W. RNA processing. Splicing in space. *Nature* vol. 372 727–728 Preprint at https://doi.org/10.1038/372727a0 (1994).
62. Wansink, D. G. *et al.* Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. *J Cell Biol* **122**, 283–293 (1993).
63. Hendzel, M. J. & Bazett-Jones, D. P. RNA polymerase II transcription and the functional organization of the mammalian cell nucleus. *Chromosoma* **103**, 509–516 (1995).
64. Zhang, G., Taneja, K. L., Singer, R. H. & Green, M. R. Localization of pre-mRNA splicing in mammalian nuclei. *Nature* **372**, 809–812 (1994).
65. Spector, D. L. Macromolecular domains within the cell nucleus. *Annu Rev Cell Biol* **9**, 265–315 (1993).
66. Lewis, J. D. & Tollervey, D. Like attracts like: getting RNA processing together in the nucleus. *Science* **288**, 1385–1389 (2000).

67. Tammer, L. *et al.* Gene architecture directs splicing outcome in separate nuclear spatial regions. *Mol Cell* **82**, 1021-1034.e8 (2022).
68. Barutcu, A. R. *et al.* Systematic mapping of nuclear domain-associated transcripts reveals speckles and lamina as hubs of functionally distinct retained introns. *Mol Cell* **82**, 1035-1052.e9 (2022).
69. Alexander, K. A. *et al.* p53 mediates target gene association with nuclear speckles for amplified RNA expression. *Mol Cell* **81**, 1666-1681.e6 (2021).
70. Khanna, N., Hu, Y. & Belmont, A. S. HSP70 transgene directed motion to nuclear speckles facilitates heat shock activation. *Current Biology* **24**, 1138–1144 (2014).
71. Lawrence, J. B. & Clemson, C. M. Gene associations: True romance or chance meeting in a nuclear neighborhood? *Journal of Cell Biology* **182**, 1035–1038 (2008).
72. Kim, J., Venkata, N. C., Hernandez Gonzalez, G. A., Khanna, N. & Belmont, A. S. Gene expression amplification by nuclear speckle association. *J Cell Biol* (2020) doi:10.1083/jcb.201904046.
73. Wang, K. *et al.* Intronless mRNAs transit through nuclear speckles to gain export competence. *Journal of Cell Biology* (2018) doi:10.1083/jcb.201801184.
74. Dias, A. P., Dufu, K., Lei, H. & Reed, R. A role for TREX components in the release of spliced mRNA from nuclear speckle domains. *Nat Commun* **1**, 97 (2010).
75. Ilık, İ. A. & Aktaş, T. Nuclear speckles: dynamic hubs of gene expression regulation. *FEBS J* **289**, 7234–7245 (2022).
76. Hirose, T., Ninomiya, K., Nakagawa, S. & Yamazaki, T. A guide to membraneless organelles and their various roles in gene regulation. *Nat Rev Mol Cell Biol* (2022) doi:10.1038/s41580-022-00558-8.
77. Quinodoz, S. A. *et al.* 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 (2022).
78. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).
79. Quinodoz, S. A. *et al.* RNA promotes the formation of spatial compartments in the nucleus. *Cell* **184**, 5775-5790.e30 (2021).
80. Chen, Y. *et al.* Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *Journal of Cell Biology* (2018) doi:10.1083/jcb.201807108.
81. Takei, Y. *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021).
82. Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu Rev Biochem* **89**, 359–388 (2020).
83. Calvet, J. P. & Pederson, T. Base-pairing interactions between small nuclear RNAs and nuclear RNA precursors as revealed by psoralen cross-linking in vivo. *Cell* (1981) doi:10.1016/0092-8674(81)90205-1.
84. Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159**, 188–199 (2014).
85. Pandya-Jones, A. & Black, D. L. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15**, 1896–1908 (2009).
86. Wuarin, J. & Schibler, U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* **14**, 7219–7225 (1994).
87. Goronzy, I. N. *et al.* Simultaneous mapping of 3D structure and nascent RNAs argues against nuclear compartments that preclude transcription. *Cell Rep* **41**, (2022).
88. Luco, R. F., Allo, M., Schor, I. E., Kornblihtt, A. R. & Misteli, T. Epigenetics in alternative pre-mRNA splicing. *Cell* **144**, 16–26 (2011).
89. Kornblihtt, A. R. Promoter usage and alternative splicing. *Curr Opin Cell Biol* **17**, 262–268 (2005).

90. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res* **23**, 1256–1269 (2013).
91. Curado, J., Iannone, C., Tilgner, H., Valcárcel, J. & Guigó, R. Promoter-like epigenetic signatures in exons displaying cell type-specific splicing. *Genome Biol* **16**, 236 (2015).
92. Agirre, E., Oldfield, A. J., Bellora, N., Segelle, A. & Luco, R. F. Splicing-associated chromatin signatures: a combinatorial and position-dependent role for histone marks in splicing definition. *Nat Commun* **12**, 682 (2021).
93. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).
94. Pichon, X., Robert, M.-C., Bertrand, E., Singer, R. H. & Tutucci, E. New Generations of MS2 Variants and MCP Fusions to Detect Single mRNAs in Living Eukaryotic Cells. *Methods Mol Biol* **2166**, 121–144 (2020).
95. Tripathi, V. *et al.* SRSF1 regulates the assembly of pre-mRNA processing factors in nuclear speckles. *Mol Biol Cell* (2012) doi:10.1091/mbc.E12-03-0206.
96. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Sci Adv* **7**, (2021).
97. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, (2017).
98. Olins, A. L., Rhodes, G., Welch, D. B. M., Zwerger, M. & Olins, D. E. Lamin B receptor: multi-tasking at the nuclear envelope. *Nucleus* **1**, 53–70 (2010).
99. Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes and Development* vol. 26 2119–2137 Preprint at https://doi.org/10.1101/gad.200303.112 (2012).
100. David, C. J., Boyne, A. R., Millhouse, S. R. & Manley, J. L. The RNA polymerase II C-terminal domain promotes splicing

activation through recruitment of a U2AF65-Prp19 complex. *Genes Dev* **25**, 972–983 (2011).
101. Brodsky, A. S. *et al.* Genomic mapping of RNA polymerase II reveals sites of co-transcriptional regulation in human cells. *Genome Biol* **6**, R64 (2005).
102. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics* vol. 15 163–175 Preprint at https://doi.org/10.1038/nrg3662 (2014).
103. Fong, N. & Bentley, D. L. Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. *Genes Dev* **15**, 1783–1795 (2001).
104. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *Journal of Molecular Biology* vol. 428 2623–2635 Preprint at https://doi.org/10.1016/j.jmb.2016.04.017 (2016).
105. de la Mata, M. *et al.* A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**, 525–532 (2003).
106. McCracken, S. *et al.* The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**, 357–360 (1997).
107. Ip, J. Y. *et al.* Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res* **21**, 390–401 (2011).
108. Tsutsumi, C. *et al.* Retinoids and retinoid-binding protein expression in rat adipocytes. *J Biol Chem* **267**, 1805–1810 (1992).
109. Dundr, M. Seed and grow: a two-step model for nuclear body biogenesis. *Journal of Cell Biology* **193**, 605–606 (2011).
110. Kaiser, T. E., Intine, R. V. & Dundr, M. De novo formation of a subnuclear body. *Science (1979)* (2008) doi:10.1126/science.1165216.
111. Das, R. *et al.* Functional coupling of RNAP II transcription to spliceosome assembly. *Genes Dev* **20**, 1100–1109 (2006).

112. Ding, F. & Elowitz, M. B. Constitutive splicing and economies of scale in gene expression. *Nat Struct Mol Biol* **26**, 424–432 (2019).
113. Emili, A. *et al.* Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. *RNA* **8**, 1102–1111 (2002).
114. Rosonina, E. *et al.* Role for PSF in mediating transcriptional activator-dependent stimulation of pre-mRNA processing in vivo. *Mol Cell Biol* **25**, 6734–6746 (2005).
115. Morris, D. P. & Greenleaf, A. L. The Splicing Factor, Prp40, Binds the Phosphorylated Carboxyl-terminal Domain of RNA Polymerase II *. *Journal of Biological Chemistry* **275**, 39935–39943 (2000).
116. Yuryev, A. *et al.* The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. *Proceedings of the National Academy of Sciences* **93**, 6975–6980 (1996).
117. Lazarev, D. & Manley, J. L. Concurrent splicing and transcription are not sufficient to enhance splicing efficiency. *RNA* **13**, 1546–1557 (2007).
118. Pederson, T. The nucleolus. *Cold Spring Harb Perspect Biol* **3**, 1–15 (2011).
119. Nizami, Z., Deryusheva, S. & Gall, J. G. The Cajal body and histone locus body. *Cold Spring Harbor perspectives in biology* vol. 2 Preprint at https://doi.org/10.1101/cshperspect.a000653 (2010).
120. Shin, Y. & Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **357**, 6357 (2017).
121. Galganski, L., Urbanek, M. O. & Krzyzosiak, W. J. Nuclear speckles: Molecular organization, biological function and role in disease. *Nucleic Acids Research* Preprint at https://doi.org/10.1093/nar/gkx759 (2017).
122. Morimoto, M. & Boerkoel, C. F. The role of nuclear bodies in gene expression and disease. *Biology (Basel)* **2**, 976–1033 (2013).

123. Mensah, M. A. *et al.* Aberrant phase separation and nucleolar dysfunction in rare genetic diseases. *Nature* **614**, 564–571 (2023).
124. Maehama, T., Nishio, M., Otani, J., Mak, T. W. & Suzuki, A. Nucleolar stress: Molecular mechanisms and related human diseases. *Cancer Sci* (2023) doi:10.1111/cas.15755.
125. Alberti, S. & Hyman, A. A. Biomolecular condensates at the nexus of cellular stress, protein aggregation disease and ageing. *Nat Rev Mol Cell Biol* **22**, 196–213 (2021).
126. Adriaens, C. *et al.* p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. *Nat Med* **22**, 861–868 (2016).
127. Banerjee, A. K. *et al.* SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses. *Cell* **183**, 1325-1339.e21 (2020).
128. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* (2016) doi:10.1038/nature20149.
129. Melsted, P. *et al.* Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat Biotechnol* **39**, 813–818 (2021).
130. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* **5**, 877–879 (2008).
131. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat Methods* **18**, 100–106 (2021).
132. Mayr, C. & Bartel, D. P. Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**, 673–684 (2009).

Chapter 5

CONCLUSION AND FUTURE DIRECTIONS

"Somewhere, something incredible is waiting to be known." – Carl Sagan

5.1 CONCLUSIONS AND FUTURE DIRECTIONS

In this thesis, we aimed to highlight the functional roles of specific nuclear compartments in various aspects of gene regulation, including transcriptional initiation, higher-order chromatin regulation, epigenetics, and post-transcriptional RNA processing. We used these examples to describe several emerging principles by which compartmentalization can have important quantitative implications for gene regulation. These included: (i) increasing affinity and kinetic rates, (ii) increasing specificity of target recognition, (iii) memory and persistence of transcriptional states, (iv) time-dependent and concentration-dependent transitions in responses, and (v) the ability to achieve non-stoichiometric responses. We anticipate that this emerging perspective will enable a more general understanding of how changes in spatial concentration of molecules in the nucleus drive emergent quantitative and kinetic properties in gene regulation.

Yet, nuclear compartmentalization remains a new and active area of research and there are still many open questions and challenges in this nascent field.

5.1.2 FUNCTIONS OF NUCLEAR COMPARTMENTS. While nuclear compartments can play key functional roles in specific cases, not all nuclear compartments may be functionally important. For example, although HP1a has been shown to form condensates[163], it was recently shown that heterochromatin maintenance can occur independently of this property[163]. Moreover, repressed DNA is known to be organized around

the nuclear lamina – a nuclear compartment that is enriched for various repressive chromatin regulators and heterochromatin factors[164]. Yet, several studies showed that synthetically anchoring DNA to the nuclear lamina had little to no effect of transcriptional regulation[165–167]. As such, in some cases, nuclear compartments may simply represent emergent properties of functional states (e.g., repressed DNA tends to associate with the nuclear lamina) rather than critical mechanisms for controlling function.

To date, what functional roles, if any, most nuclear compartments might play remains largely unknown. One of the main challenges in addressing this functional question is that studying nuclear compartments, their formation, and functional effects are often confounded. For example, it is challenging to disrupt the formation of transcription factor condensates without also impacting their activation domains[95] or to disrupt the ability for PRC1 proteins to form a condensate without impacting its ability to condense chromatin[41]. A key example of this challenge is illustrated by the nucleolus, which is one of the most well-characterized nuclear compartments. While it is now well-defined that the molecular components associated with ribosome biogenesis are contained within this compartment, whether this spatial organization is essential for ribosome biogenesis or whether this process can occur even when these components are not spatially assembled remains untested because there is no simple way to disrupt nucleolar assembly without also impacting ribosome biogenesis (e.g., 45S pre-rRNA transcription) directly. As such, it is often difficult to disentangle the cause-and-effect relationship between compartmentalization of molecules and their functional roles.

Recent technological innovations in engineering condensates and DNA organization around nuclear bodies have begun to enable direct manipulation and exploration of these questions[168–171] (see **Chapter 1, Box 1**). Yet, while these methods represent a major advance towards addressing these questions, most of these tools still rely on creating compartments and condensates through overexpression of molecular components and as such do not perfectly represent the normal cellular

context. Defining the functional roles of different nuclear compartments will require creative new genetic approaches to test the causal role of compartmentalization in gene regulation.

5.1.2 MECHANISMS OF COMPARTMENT FORMATION. In addition, the mechanisms by which specific compartments form is still unclear in most cases. Although specific compartments have been well mapped out, there are still many open questions about the molecular mechanisms by which most nuclear compartments are formed. For example, what molecules act as seeds that nucleate formation of a nuclear compartment and whether spatial enrichment of molecules is mediated by stoichiometric interactions with a seed, concentration-dependent condensate formation, or a combination of multiple events is largely unknown. For example, the inactive X chromosome (Barr body) is well characterized to form a nuclear compartment, yet the precise mechanism is not fully resolved. While there is a clear indication that individual Xist binding proteins can form condensates[47,56], whether or not the entire Barr body itself forms a condensate remains unknown.

One of the key challenges is that it remains difficult to measure phase separation and molecular dynamics and assembly in living cells. In the case of phase separation, to date, most studies have characterized these properties *in vitro* using purified proteins or by imaging tagged and overexpressed proteins *in vivo.* Both of these approaches present limitations since they do not represent the endogenous concentrations that occur within the cell. Moreover, nuclear compartments are often too small to accurately measure their internal diffusion rates or spherical shapes or to observe individual molecules and their stoichiometries. Improvements in super-resolution microscopy have begun to address some of these challenges and we anticipated that further advances in these microscopy-based approaches will be needed to resolve many of these questions.

5.1.3 MOLECULAR COMPONENTS CONTAINED WITHIN NUCLEAR COMPARTMENTS. It remains largely unknown what molecular components are contained within individual nuclear compartments. For

example, while we know that RNA Pol II can form spatially enriched compartments within the nucleus, we still do not know the full assembly of protein components that are contained within these condensates or what DNA regions (if any) these condensates localize at.

The main challenge is that nuclear compartments can: (i) contain dozens to thousands of distinct molecular components, including DNA, RNA, and protein, (ii) occupy small (nanometer) to large (micrometer) distances in the nucleus, and (iii) undergo precise concentration-dependent transitions. Current methods for measuring molecular organization – including microscopy and genomic methods – are limited in their ability to measure these features because they cannot currently map the combinatorial and spatial organization of these complex assemblies. Recent advances in highly multiplexed super-resolution microscopy methods (e.g., seqFISH+[172,173]) and genomic methods for mapping multiway spatial interactions (e.g., SPRITE[10,56]) have begun to shed light on some of these questions. We expect that further technological developments to increase scale, resolution, and detection modalities will be critical for fully defining the molecular and spatial composition of nuclear compartments.

5.1.4 RELATIONSHIP BETWEEN BIOPHYSICAL PROPERTIES AND FUNCTIONAL ROLE OF COMPARTMENTS. Finally, it remains largely unclear what the functional implications might be for nuclear compartments that have different biophysical properties. For example, are there particular functional implications for compartments that are liquid-like and are they functionally distinct from compartments that may display gel or solid properties? One proposal is that the ability to form a liquid-like state would lead to a barrier that would limit molecular diffusion across its boundaries[32,174]. Another proposal is that liquid-like condensates may enable persistence of molecular assembly even upon the loss of the initiating molecule[47]. While attractive, these proposals have yet to be experimentally demonstrated for biological compartments. Testing the relationship between the biophysical properties of a compartment and its functional roles will require the abilities to accurately define their physical

properties *in vivo*, specifically disrupt or transform these biophysical states, and measure their functional roles.

We expect that advances in experimental tools will enable further advances in studying the quantitative relationships between compartmentalization and gene regulation and will likely uncover additional functional roles that have yet to be defined.